ON THE "HUMAN" IN HUMAN-ARTIFICIAL INTELLIGENCE INTERACTION

EDITED BY: Stefano Triberti, Davide La Torre, Jianyi Lin, Ilaria Durosini and Manuel Ruiz Galan
PUBLISHED IN: Frontiers in Psychology and Frontiers in Computer Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-322-3
DOI 10.3389/978-2-88974-322-3

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

ON THE "HUMAN" IN HUMAN-ARTIFICIAL INTELLIGENCE INTERACTION

Topic Editors:
Stefano Triberti, University of Milan, Italy
Davide La Torre, SKEMA Business School, Sophia Antipolis Campus, France
Jianyi Lin, Khalifa University, United Arab Emirates
Ilaria Durosini, European Institute of Oncology (IEO), Italy
Manuel Ruiz Galan, University of Granada, Spain

Citation: Triberti, S., La Torre, D., Lin, J., Durosini, I., Galan, M. R., eds. (2022). On the "Human" in Human-Artificial Intelligence Interaction.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-322-3

Table of Contents

frontiers in Psychology

EDITORIAL
published: 24 December 2021
doi: 10.3389/fpsyg.2021.808995

Editorial: On the "Human" in Human-Artificial Intelligence Interaction

Stefano Triberti[1,2]*, *Ilaria Durosini*[2], *Jianyi Lin*[3,4], *Davide La Torre*[1,5] and *Manuel Ruiz Galán*[6]

[1] Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy, [2] Applied Research Division for Cognitive and Psychological Science, IEO, European Institute of Oncology IRCCS, Milan, Italy, [3] Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Milan, Italy, [4] Department of Mathematics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, [5] SKEMA Business School and Université Cote d'Azur, Sophia Antipolis Campus, Sophia Antipolis, France, [6] Department of Applied Mathematics, University of Granada, Granada, Spain

Keywords: artificial intelligence, eXplainable Artificial Intelligence (XAI), human centered AI, human technology interaction, cyberpsychology, attitudes toward technology

OPEN ACCESS

Edited by:
Hamidreza Namazi, Monash University Malaysia, Malaysia

Reviewed by:
Edwin Lughofer, Johannes Kepler University of Linz, Austria
Yenchun Jim Wu, National Taiwan Normal University, Taiwan

***Correspondence:**
Stefano Triberti
stefano.triberti@unimi.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *04 November 2021*
Accepted: *06 December 2021*
Published: *24 December 2021*

Citation:
Triberti S, Durosini I, Lin J, La Torre D and Ruiz Galán M (2021) Editorial: On the "Human" in Human-Artificial Intelligence Interaction. Front. Psychol. 12:808995. doi: 10.3389/fpsyg.2021.808995

Editorial on the Research Topic

On the "Human" in Human-Artificial Intelligence Interaction

Artificial Intelligence or technologies able to perform tasks normally requiring human cognitive processes (e.g., reasoning, perception) are revolutionizing many fields such as healthcare and business. For example, medical doctors use artificial intelligence to analyze pathological data and patients' genomic profiles to identify personalized treatment according to a precision medicine approach. In general, artificial intelligence represents an invaluable resource for any professional dealing with the need to understand data and make decisions.

However, desirable utilization of technology largely depends on the interface that allows users to form a representation of software's structure and functions. Research is still needed to provide information on how humans represent artificial intelligence. This is important especially when the future users are not experts in algorithms but they still need to make decisions based on deep learning outcomes. Last but not least, we still have to understand and master the multiple ways artificial intelligence could be used to address *human* issues: how can artificial intelligence contribute to improving people's health, well-being and flourishing?

Psycho-social research shows that technologies are not accepted by users and implemented in real-life on the sole basis of effectiveness. People form attitudes toward technologies that shape their future behavior (Venkatesh and Davis, 2000; Marangunć and Granić, 2015; Gorini et al., 2018; Nunes et al., 2019); or, they evaluate technologies according to pre-existing intentions, needs and misconceptions that may lead to improper usage, errors, and ultimately abandonment (Triberti et al., 2016; Sebri et al., 2020). Without an understanding of the human barriers and motivations for adoption and acceptance of AI, AI is simply just an invention in search of a market.

To understand human responses to AI, we identify five categories of potential scientific areas requiring further investigation for this special issue:

- The study of attitudes and behaviors toward artificial intelligence (Dos Santos et al., 2019; Schepman and Rodway, 2020; Sebri et al., 2020) (area A);

- The study, development, and validation of artificial intelligence-human interfaces; this includes eXplainable Artificial Intelligence (XAI), or the sub-discipline devoted to make "black-box" algorithms understandable to human users (Miller, 2019), and Human Factors research on systems involving artificial intelligence (Knijnenburg et al., 2012; Lau et al., 2020) (area B);
- The research on human characteristics that could hinder or promote effective interaction with artificial intelligence (Oksanen et al., 2020; Sharan and Romano, 2020; Matthews et al., 2021); this includes models and criteria to select personnel expected to work with artificial intelligence (La Torre et al., 2021) (area C);
- The identification of issues in artificial intelligence implementation and/or possible solutions to existing issues, including social science, political science, and philosophy/ethics contributions (Pravettoni et al., 2015; Triberti et al., 2020a,b) (area D);
- Research on the implementation or testing of specific artificial intelligence solutions that require interaction with human users, and provides information relevant to better understand risks and opportunities (Adamo et al., 2015; Bodini et al., 2018) (area E).

The present special issue aimed at collecting innovative and interdisciplinary contributions on the topic of artificial intelligence-human interaction, that emphasize the "human" part and provide insights to improve the development of artificial intelligence that could be really useful and effectively used in society. All the contributions to this special issue indeed touch on one or more of the research areas highlighted above, as it is evidenced below by reference to the designated areas' letters.

Specifically, the contribution by Biancardi et al. (areas A, B, C) deals with the topic of interface, specifically in terms of embodied conversational agents: it elaborates on the topic of adaptation, testing three different models that allow embodied conversational agents to modify their behavior based on the user's response. They show that the way we conceptualize adaptive interfaces affects users' engagement with artificial intelligence.

In this line, the theoretical contribution by Hildt (areas A, B, D) reflects on how humans would like to interact with robots and how the interaction influences both parts. It is suggested that a broader perspective on Human-Robot Interaction is needed that takes the social and ethical implications into account. Although humans tend to react to robots in similar ways as they react to human beings even if they are not, aspects needing more attention include how to deal with simulated human-like behavior that is not grounded in human-like capabilities. Moreover, questions of what social roles to ascribe to robots deserve a central importance in designing them.

Interface and its ethical and practical aspects are elaborated further in the contribution by Holohan and Fiske, dealing mostly with area D, focused on artificial intelligence in psychotherapy and the concept of transference: indeed both these studies show that we may need to update conceptions, theoretical constructs, and terminology to support desirable implementation of artificial intelligence solutions within sensitive contexts, such as healthcare. Design thinking and the associated research methods may be an important resource to conceptualize artificial intelligence solutions that address real-world issues, as suggested by the perspective article by Talamo et al. (area B) focused on systems to support venture capitalists' decision-making. Indeed, one possible way to improve artificial intelligence is to consider users' needs and context since the first steps of the design of both algorithms and interface, consistently with a user-centered approach (Weller, 2019). From a broader point of view, the two reviews by Tariq, Poulin et al. (areas A, C, D) and Abonamah et al. (area D) also help to identify relevant factors involved both in operational excellence and commoditization of artificial intelligence. In particular, the former sheds novel light on how artificial intelligence can provide driving forces for achieving operational excellence in a business company (Gólcher-Barguil et al., 2019) as soon as certain barriers consisting of lack of skills, technologies and strategy can be overcome, while the latter well-interprets and outlines the role of artificial intelligence technologies as commodities within an organization in a comprehensive and systematic way comparing to existing literature (Carr, 2003).

Furthermore, it is important to take into account all psychological, medico-legal, and ethical issues which need to be addressed to artificial intelligence be considered fully capable of patient management in real life. Coppola et al. (areas C and D) provide an overview of the state of the art of artificial intelligence systems regarding medical imaging, with special focus on how artificial intelligence can be implemented in a human-centered field such as contemporary medicine. This approach contributes in addressing important issues associated with artificial intelligence in sensitive contexts (e.g., ethical and organizational) (Keskinbora, 2019; Triberti et al., 2020a), as it encourages health professionals to actively engage in iterative discourse to preserve humanitarian sensitivity in the future models of care.

Tariq, Babar et al. related to category E, propose and test a framework based on Apache Spark for efficiently processing the big datasets resulting from user comment activities triggered by videos on social media. The article shows the potential effectiveness of the devised implementation, which was able to perform the planned analytics operations on social media dataset in a time that well-scales with the data size. Specifically, they provide a new concrete demonstration of processing big data coming from an extended social hub named Dailymotion within a time frame of few minutes using Apache Spark.

Certainly future research needs innovative tools and approaches to address human behavior through the lenses of artificial intelligence. An example of integration between artificial intelligence and social psychology methods is the work by Catellani et al. (area E) who, moving from the psychological concept of framing, test persuasive messages to do home-based physical activities and use the results to inform the development of a Dynamic Bayesian Network predictor. This points toward

the development of artificial intelligence-based tools that autonomously interact with human users to support positive behavioral change. Similarly, Peifer et al. (area E, possibly with interesting hints for future research in areas B and C too) focus on team flow (i.e., a shared experience characterized by the pleasant feeling of absorption in challenging activities and of optimal team-interaction during an interdependent task), a well-known concept in group and work psychology. They identify psychophysiological and behavioral correlates which can be used as input data for a machine learning system to assess team flow in real time. Such approaches constitute notable examples of how artificial intelligence could provide new avenues for research and intervention on human behavior, consistently with the prediction that artificial intelligence will play a more and more important role in psychological research (Lisetti and Schiano, 2000; Daróczy, 2010; Tuena et al., 2020).

In conclusion, this Research Topic provides an overview on artificial intelligence-human interaction, focusing on relevant psychological, technical, and methodological aspects of real-life implementation. Emphasizing the "human" in the human-artificial intelligence interaction provides insights to design the future technologies that could contribute to advance society.

AUTHOR CONTRIBUTIONS

ST and ID drafted the editorial. JL, DL, and MR participated in the discussion on the ideas presented and edited the editorial. All authors approved the submitted version.

FUNDING

ST was supported by MIUR—Italian Ministry of University and Research (Department of Excellence Italian Law n.232, 11th December 2016) for University of Milan. ID was supported by Fondazione Umberto Veronesi.

REFERENCES

Adamo, A., Grossi, G., Lanzarotti, R., and Lin, J. (2015). Robust face recognition using sparse representation in LDA space. *Mach. Vis. Applic.* 26, 837–847. doi: 10.1007/s00138-015-0694-x

Bodini, M., D'Amelio, A., Grossi, G., Lanzarotti, R., and Lin, J. (2018). "Single sample face recognition by sparse recovery of deep-learned lda features," in *International Conference on Advanced Concepts for Intelligent Vision Systems, LNCS, Vol. 11182* (Cham: Springer), 297–308. doi: 10.1007/978-3-030-01449-0_25

Carr, N. G. (2003). IT doesn't matter. *Educau. Rev.* 38, 24–38. doi: 10.1080/0957404032000081692

Daróczy, G. (2010). "Artificial intelligence and cognitive psychology," in *Proceedings of the 8th International Conference on Applied Informatics* (Eger), 61–69.

Dos Santos, D. P., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., et al. (2019). Medical students' attitude towards artificial intelligence: a multicentre survey. *Euro. Radiol.* 29, 1640–1646. doi: 10.1007/s00330-018-5601-1

Gólcher-Barguil, L. A., Nadeem, S. P., and Garza-Reyes, J. A. (2019). Measuring operational excellence: an operational excellence profitability (OEP) approach. *Product. Plan. Cont.* 30, 682–698. doi: 10.1080/09537287.2019.1580784

Gorini, A., Mazzocco, K., Triberti, S., Sebri, V., Savioni, L., and Pravettoni, G. (2018). A P5 Approach to m-Health: design suggestions for advanced mobile health technology. *Front. Psychol.* 9:2066. doi: 10.3389/fpsyg.2018.02066

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *J. Clin. Neurosci.* 64, 277–282. doi: 10.1016/j.jocn.2019.03.001

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Model. User Adapt. Interact.* 22, 441–504. doi: 10.1007/s11257-011-9118-4

La Torre, D., Colapinto, C., Durosini, I., and Triberti, S. (2021). Team formation for human-artificial intelligence collaboration in the workplace: a goal programming model to foster organizational change. *IEEE Trans. Eng. Manage.* doi: 10.1109/TEM.2021.3077195

Lau, N., Hildebrandt, M., and Jeon, M. (2020). Ergonomics in AI: designing and interacting with machine learning and AI. *Ergonom. Des.* 28:3. doi: 10.1177/1064804620915238

Lisetti, C. L., and Schiano, D. J. (2000). Automatic facial expression interpretation: where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmat. Cogn.* 8, 185–235. doi: 10.1075/pc.8.1.09lis

Marangunć, N., and Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Univ. Access Informat. Soc.* 14, 81–95. doi: 10.1007/s10209-014-0348-1

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., et al. (2021). Evolution and revolution: personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Pers. Indiv. Diff.* 169:109969. doi: 10.1016/j.paid.2020.109969

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Art. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Nunes, A., Limpo, T., and Castro, S. L. (2019). Acceptance of mobile health applications: examining key determinants and moderators. *Front. Psychol.* 10:2791. doi: 10.3389/fpsyg.2019.02791

Oksanen, A., Savela, N., Latikka, R., and Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human–technology interactions online. *Front. Psychol.* 11:568256. doi: 10.3389/fpsyg.2020.568256

Pravettoni, G., Folgieri, R., and Lucchiari, C. (2015). "Cognitive science in telemedicine: from psychology to artificial intelligence," in *Tele-oncology TELe-Health,* eds G. Gatti, G. Pravettoni, F. Capello (Cham: Springer). doi: 10.1007/978-3-319-16378-9_2

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

Sebri, V., Pizzoli, S. F. M., Savioni, L., and Triberti, S. (2020). Artificial Intelligence in mental health: professionals' attitudes towards AI as a psychotherapist. *Ann. Rev. Cyberther. Telemed.* 18, 229–233. Available online at: https://www.arctt.info/volume-18-summer-2020

Sharan, N. N., and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6:e04572. doi: 10.1016/j.heliyon.2020.e04572

Triberti, S., Durosini, I., Curigliano, G., and Pravettoni, G. (2020b). Is explanation a marketing problem? The quest for trust in artificial intelligence and two conflicting solutions. *Public Health Genom.* 23, 2–5. doi: 10.1159/000506014

Triberti, S., Durosini, I., and Pravettoni, G. (2020a). A "third wheel" effect in health decision making involving artificial entities: a psychological perspective. *Front. Public Health* 8:117. doi: 10.3389/fpubh.2020.00117

Triberti, S., Villani, D., and Riva, G. (2016). Unconscious goal pursuit primes attitudes towards technology usage: a virtual reality experiment. *Comput. Hum. Behav.* 64, 163–172. doi: 10.1016/j.chb.2016.06.044

Tuena, C., Chiappini, M., Repetto, C., and Riva, G. (2020). "Artificial intelligence in clinical psychology," in *Reference Module in Neuroscience and Biobehavioral Psychology* (Elsevier). doi: 10.1016/B978-0-12-818697-8.00001-7

Venkatesh, V., and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manage. Sci.* 46, 186–204. doi: 10.1287/mnsc.46.2.186.11926

Weller, A. J. (2019). Design thinking for a user-centered approach to artificial intelligence. *J Des. Econ Innovat.* 5, 394–396. doi: 10.1016/j.sheji.2019.11.015

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.

frontiers in Computer Science

OPINION
published: 05 July 2021
doi: 10.3389/fcomp.2021.671012

What Sort of Robots Do We Want to Interact With? Reflecting on the Human Side of Human-Artificial Intelligence Interaction

*Elisabeth Hildt**

Illinois Institute of Technology, Chicago, IL, United States

Keywords: ethics, social robots, human-technology interaction, social cognition, terminology, capabilities, human-robot interaction, social implications

INTRODUCTION

During the past decades, the interplay between humans and robots has been investigated in the field of human-robot interaction (HRI). This research has provided fascinating results on the spectrum and forms of engagement between humans and robots and on the various behaviors displayed by robots aimed at interacting with and influencing humans (Tsarouchi et al., 2016; Ahmad et al., 2017; Saunderson and Nejat, 2019a). Yet, crucial questions regarding how humans want to interact with and be influenced by robots are sidestepped in this research, falling for what could be called a robotistic fallacy. This article outlines some of the current findings on HRI to then critically assess the broader implications of HRI and key questions that must be asked in this context.

Social robots, i.e., robots that engage on a social level with humans, are expected to increasingly assist and support humans in workplace environments, healthcare, entertainment, training and education, and other fields (Ahmad et al., 2017; Richert et al., 2018; Pepito et al., 2020).

By using an interdisciplinary approach that involves behavioral studies and cognitive and social neuroscience, recent research on social cognition and HRI investigates how humans perceive, interact with and react to robots in social contexts (Cross et al., 2019; Henschel et al., 2020). Especially in the context of possible future uses in healthcare or geriatric care, the importance of developing robots with which humans can easily and naturally interact has been stressed (Pepito et al., 2020; Wykowska 2020).

Henschel et al. (2020) argue that research into and knowledge of the neurocognitive mechanisms involved in human-robot interaction will supply critical insights for optimizing social interaction between humans and robots which will in turn help to develop socially sophisticated robots. They write (Henschel et al., 2020, p. 373): "Robots that respond to and trigger human emotions not only enable closer human-machine collaboration, but can also spur human users to develop long-term social bonds with these agents." This approach suggests using cognitive neuroscience to build robots that humans are likely to emotionally interact with, an approach that can be seen as an extension of affective computing (Scheutz, 2011; McDuff and Czerwinski, 2018). In this, the focus is on building social robots so that HRI resembles human-human interaction (HHI). A question rarely asked though, is how humans would like to interact with robots and what sort of robots humans would like to interact with.

For example, Wiese et al. (2017) argue that in order for humans to interact intuitively and socially with robots, robots need to be designed in a way that humans perceive them as intentional agents, i.e., as agents with mental states. They elaborate that this is achieved when robots evoke mechanisms of social cognition in the human brain that are typically evoked in HHI. Consequently, they advocate for integrating behavioral and physiological neuroscience methods in the design and evaluation of

OPEN ACCESS

Edited by:
Manuel Ruiz Galan,
University of Granada, Spain

Reviewed by:
Gabriel Skantze,
Royal Institute of Technology, Sweden

***Correspondence:**
Elisabeth Hildt
ehildt@iit.edu

Received: *22 February 2021*
Accepted: *21 June 2021*
Published: *05 July 2021*

Citation:
Hildt E (2021) What Sort of Robots Do We Want to Interact With? Reflecting on the Human Side of Human-Artificial Intelligence Interaction. Front. Comput. Sci. 3:671012. doi: 10.3389/fcomp.2021.671012

social robots to build robots that are perceived as social companions. A questionnaire-based study by Marchesi et al. (2019) found that at least sometimes, humans adopt the intentional stance to humanoid robots by explaining and predicting robot behavior through mental states and mentalistic terms. Ciardo et al. (2020) investigated participants' sense of agency, i.e., the perceived control felt on the outcome of an action, when participants engaged with robots. Their results indicate that in HRI involving a shared control setting, participants perceived a lower sense of agency, similar to what can be observed in comparable HHI.

BROADER IMPLICATIONS OF HUMAN-ROBOT INTERACTION

These are but a few examples of recent studies in an upcoming research field. Overall, research indicates that humans tend to interact with social robots in similar ways as they interact with humans, with anthropomorphic features facilitating this effect. From the perspective of HRI research, these similarities are considered an advantage, the goal being to build robots with which humans can easily and intuitively interact. In this, the guiding assumption is that cognitive and behavioral studies serve as the basis for building robots with which humans engage in such a way that HRI resembles HHI.

However, this way of reasoning jumps too easily from the empirical observation that HRI resembles HHI to the normative conclusion that HRI should resemble HHI. It could thus be called a robotistic fallacy. A possible reply to this criticism is that what makes HRI that resembles HHI attractive, is that it is user-friendly, allows for effective interaction with robots, and provides the basis for social acceptance of robots. However, so far, these claims are essentially unproven. Further, they rely on very narrow conceptions of user-friendliness and social interaction, as will be outlined below.

While the focus of most HRI studies has been on how HRI is similar to HHI (Irfan et al., 2018; Henschel et al., 2020), there is also the uncanny valley hypothesis, according to which robots with a too anthropomorphic design are considered disturbing (Mori et al., 2012; Richert et al., 2018). Furthermore, current research is primarily confined to laboratory situations and investigates the immediate situation of HRI in experimental settings. When considering real life situations, a multitude of additional factors will come in that have not been researched yet (Jung and Hinds, 2018). Interaction with robots "in the wild" will probably turn out to be much messier and more complex than research studies that highlight and welcome similarities between HRI and HHI currently assume. Reflections on the broader implications of HRI on humans go beyond the immediate experimental setting and include the broader social context. In this context, three aspects are worth considering:

Robot Capabilities

While an immediate HRI can resemble HHI, robots differ in crucial ways from humans. Current robots do not have capabilities that are in any way comparable to human sentience, human consciousness, or a human mind. Robots only *simulate* human behavior. Humans tend to react to this simulated behavior in similar ways as they react to human behavior. This is in line with research according to which humans interact with computers and new media in fundamentally social and natural ways (Reeves and Nass, 2002; Guzman, 2020).

However, capabilities matter. While taking the intentional stance toward robots may help to explain robot behavior and facilitate an interaction with robots, it does not say much about robot capabilities or the quality of the interaction. Superficially, in certain situations, the reactions of a robot simulating human behavior and human emotions and a person having emotions and showing a certain behavior may be similar. But there is a substantial difference in that there is no interpersonal interaction or interpersonal communication with a robot. While this may not play a huge role in confined experimental settings, the situation will change with wider applications of social robots. Questions to be addressed include: What are the consequences of inadequate ascription of emotions, agency, accountability and responsibility to robots? How should one deal with unilateral emotional involvement, lack of human touch and absence of equal level interaction? And how may HRI that simulates HHI influence interpersonal interactions and relationships?

How to Talk About Robot Behavior?

While it may seem tempting to describe robot behavior that simulates human behavior with the same terms as human behavior, the terminology used when talking about robots and HRI clearly needs some scrutiny (see also Salles et al., 2020). For example, in HRI studies in which participants were asked to damage or destroy robots, robots were characterized as being "abused," "mistreated" or "killed" (Bartneck and Hu, 2008; Ackerman, 2020; Bartneck and Keijsers, 2020; Connolly et al., 2020). It is questionable, however, whether concepts like "abuse" or "death" can meaningfully be used for robots. The same holds for ascribing emotions to robots and talking of robots as "being sad" or "liking" something. Claims like "Poor Cozmo. Simulated feelings are still feelings!" (Ackerman, 2020) are clearly misleading, even if meant to express some irony.

In part, this problematic language use results from a strong tendency of anthropomorphizing that is directly implied by an approach that focuses on HRI resembling HHI. In part, it may be considered a use of metaphors, comparable to metaphorical descriptions in other contexts (Lakoff and Johnson, 2003). In part, issues around terminology may be a matter of definition. Depending on the definition given, for example for "mind" or "consciousness", claims that current robots do have minds or consciousness may be perfectly adequate, implying that robot mind or robot consciousness are significantly different from human-like mind or consciousness (Bryson, 2018; Hildt, 2019; Nyholm, 2020). It may be argued that as long as clear definitions are provided, and different conceptions are used for humans and robots, this type of language use is not problematic. However, it will be important to establish a terminology for robots that allows the use of concepts in such a way that there is

no interference with the way these concepts are used for humans. Otherwise, the same term is used for two different things. As a result, humans may expect from robots that are characterized as "being conscious" or "having a mind" much more than is exhibited by the technology.

While these are primarily theoretical considerations for now, empirical studies will certainly find out more about the language used to talk about social robots and the implications.

Robots Influencing Humans

Reflections on the broader implication of HRI beyond laboratory settings include the question of what roles humans, individually and as a society, want to ascribe to robots in their lives, and how much they want to be influenced by robots. For example, in a recent exploratory HRI study (Saunderson and Nejat, 2019b), social robots used different behavior strategies in an attempt to persuade humans in a game, in which the human participants were asked to guess the number of jelly beans. The robots exhibited various verbal and nonverbal behaviors for different persuasive strategies, including verbal cues such as "It would make me happy if you used my guess (. . .)" to express affect, and "You would be an idiot if you didn't take my guess (. . .)" to criticize.

While this certainly is an interesting study, a number of questions come to mind: Is it realistic to assume that one can make a robot happy for taking its guess? In how far can a person be meaningfully blamed by a robot? What would it mean, upon reflection, to be persuaded by a robot? In how far is there deception involved? For sure, it is not the robot itself but the people who design, build and deploy the technology who attempt to elicit a certain human behavior. And there clearly are similarities to commercials and various forms of nudging. In settings like these, aspects to consider include the type of influence, its initiator, the intentions behind and the consequences of the interaction.

CONCLUSION

HRI research has shown that in various regards, humans tend to react to robots in similar ways as they react to human beings. While these are fascinating results, not much consideration has been given to the broader consequences of humans interacting with robots in real-life settings and the social acceptance of social robots. When it comes to potential future applications beyond experimental settings, a broader perspective on HRI is needed that better takes the social and ethical implications of the technology into account. As outlined above, aspects to be considered include how to deal with simulated human-like behavior that is not grounded in human-like capabilities and how to develop an adequate terminology for robot behavior. Most crucially, the questions of what social roles to ascribe to robots and to what extent influence exerted by robots would be considered acceptable need to be addressed. Instead of planning to build robots with which humans cannot but interact in certain ways, it is crucial to think about how humans would like to interact with robots. At the center of all of this is the question "What sort of robots do we want to engage with?" For it is humans who design, build and deploy robots and who by designing, building and deploying robots shape the ways humans interact with robots.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Ackerman, E. (2020). *Can Robots Keep Humans from Abusing Other Robots?* IEEE Spectrum. Available at: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/can-robots-keep-humans-from-abusing-other-robots (Accessed August 19, 2020).

Ahmad, M., Mubin, O., and Orlando, J. (2017). A Systematic Review of Adaptivity in Human-Robot Interaction. *Mti* 1 (3), 14. doi:10.3390/mti1030014

Bartneck, C., and Hu, J. (2008). Exploring the Abuse of Robots. *Is* 9 (3), 415–433. doi:10.1075/is.9.3.04bar

Bartneck, C., and Keijsers, M. (2020). The Morality of Abusing a Robot. *J. Behav. Robotics* 11, 271–283. doi:10.1515/pjbr-2020-0017

Bryson, J. J. (2018). Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Technol.* 20, 15–26. doi:10.1007/s10676-018-9448-6

Ciardo, F., Beyer, F., De Tommaso, D., and Wykowska, A. (2020). Attribution of Intentional agency towards Robots Reduces One's Own Sense of agency. *Cognition* 194, 104109. doi:10.1016/j.cognition.2019.104109

Connolly, J., Mocz, V., Salomons, N., Valdez, J., Tsoi, N., Scassellati, B., et al. (2020). "Prompting Prosocial Human Interventions in Response to Robot Mistreatment," in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20); March 23–26, 2020, Cambridge, United Kingdom. Editors T. Belpaeme and J. Young (New York, NY, USA: ACM), 10. doi:10.1145/3319502.3374781

Cross, E. S., Hortensius, R., and Wykowska, A. (2019). From Social Brains to Social Robots: Applying Neurocognitive Insights to Human-Robot Interaction. *Phil. Trans. R. Soc. B* 374, 20180024. doi:10.1098/rstb.2018.0024

Guzman, A. (2020). Ontological Boundaries between Humans and Computers and the Implications for Human-Machine Communication. *Hmc* 1, 37–54. doi:10.30658/hmc.1.3

Henschel, A., Hortensius, R., and Cross, E. S. (2020). Social Cognition in the Age of Human-Robot Interaction. *Trends Neurosci.* 43 (6), 373–384. doi:10.1016/j.tins.2020.03.013

Hildt, E. (2019). Artificial Intelligence: Does Consciousness Matter? *Front. Psychol.* 10, 1535. doi:10.3389/fpsyg.2019.01535

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., and Belpaeme, T. (2018). "Social Psychology and Human-Robot Interaction: An Uneasy Marriage," in HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion; Chicago, IL, USA, March. 5–8, 2018. New York, NY, USA: ACM, 8. doi:10.1145/3173386.3173389

Jung, M., and Hinds, P. (2018). Robots in the Wild: A Time for More Robust Theories of Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* 7 (1), 5. doi:10.1145/3208975

Lakoff, G., and Johnson, M. (2003). *Metaphors We Live by*. Chicago and London: University of Chicago Press. doi:10.7208/chicago/9780226470993.001.0001

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A. (2019). Do We Adopt the Intentional Stance toward Humanoid Robots? *Front. Psychol.* 10, 450. doi:10.3389/fpsyg.2019.00450

McDuff, D., and Czerwinski, M. (2018). Designing Emotionally Sentient Agents. *Commun. ACM* 61 (12), 74–83. doi:10.1145/3186591

Mori, M., MacDorman, K., and Kageki, N. (2012). The Uncanny valley [from the Field]. *IEEE Robot. Automat. Mag.* 19 (2), 98–100. doi:10.1109/mra.2012.2192811

Nyholm, S. (2020). *Humans and Robots. Ethics, Agency, and Anthropomorphism*. London, New York: Rowman & Littlefield.

Pepito, J. A., Ito, H., Betriana, F., Tanioka, T., and Locsin, R. C. (2020). Intelligent Humanoid Robots Expressing Artificial Humanlike Empathy in Nursing Situations. *Nurs. Philos.* 21 (4), e12318. doi:10.1111/nup.12318

Reeves, B., and Nass, C. (2002). *The Media Equation.How People Treat Computers, Television, and New Media like Real People and Places*. Stanford: CSLI Publications.

Richert, A., Müller, S., Schröder, S., and Jeschke, S. (2018). Anthropomorphism in Social Robotics: Empirical Results on Human-Robot Interaction in Hybrid Production Workplaces. *AI Soc.* 33, 413–424. doi:10.1007/s00146-017-0756-x

Salles, A., Evers, K., and Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neurosci.* 11 (2), 88–95. doi:10.1080/21507740.2020.1740350

Saunderson, S., and Nejat, G. (2019a). How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human-Robot Interaction. *Int. J. Soc. Rob.* 11, 575–608. doi:10.1007/s12369-019-00523-0

Saunderson, S., and Nejat, G. (2019b). It Would Make Me Happy if You Used My Guess: Comparing Robot Persuasive Strategies in Social Human-Robot Interaction. *IEEE Robot. Autom. Lett.* 4 (2), 1707–1714. doi:10.1109/lra.2019.2897143

Scheutz, M. (2011). "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," in *Robot Ethics. The Ethical and Social Implications of Robotics*. Editors P. Lin, K. Abney, and G. A. Bekey (USA: MIT Press), 205–221.

Tsarouchi, P., Makris, S., and Chryssolouris, G. (2016). Human-robot Interaction Review and Challenges on Task Planning and Programming. *Int. J. Computer Integrated Manufacturing* 29 (8), 916–931. doi:10.1080/0951192x.2015.1130251

Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Front. Psychol.* 8, 1663. doi:10.3389/fpsyg.2017.01663

Wykowska, A. (2020). Where I Work. *Nature* 583, 652. https://media.nature.com/original/magazine-assets/d41586-020-02155-1/d41586-020-02155-1.pdf

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fpsyg.2021.686610

Human Behavior Analysis Using Intelligent Big Data Analytics

Muhammad Usman Tariq[1], Muhammad Babar[2], Marc Poulin[1], Akmal Saeed Khattak[3], Mohammad Dahman Alshehri[4] and Sarah Kaleem[5]*

[1] Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates, [2] Department of Computer Science, Allama Iqbal Open University, Islamabad, Pakistan, [3] Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan, [4] College of Computers and Information Technology, Taif University, Taif, Saudi Arabia, [5] Department of Computing and Technology, Iqra University, Karachi, Pakistan

Intelligent big data analysis is an evolving pattern in the age of big data science and artificial intelligence (AI). Analysis of organized data has been very successful, but analyzing human behavior using social media data becomes challenging. The social media data comprises a vast and unstructured format of data sources that can include likes, comments, tweets, shares, and views. Data analytics of social media data became a challenging task for companies, such as Dailymotion, that have billions of daily users and vast numbers of comments, likes, and views. Social media data is created in a significant amount and at a tremendous pace. There is a very high volume to store, sort, process, and carefully study the data for making possible decisions. This article proposes an architecture using a big data analytics mechanism to efficiently and logically process the huge social media datasets. The proposed architecture is composed of three layers. The main objective of the project is to demonstrate Apache Spark parallel processing and distributed framework technologies with other storage and processing mechanisms. The social media data generated from Dailymotion is used in this article to demonstrate the benefits of this architecture. The project utilized the application programming interface (API) of Dailymotion, allowing it to incorporate functions suitable to fetch and view information. The API key is generated to fetch information of public channel data in the form of text files. Hive storage machinist is utilized with Apache Spark for efficient data processing. The effectiveness of the proposed architecture is also highlighted.

Keywords: human behavior, big data, artificial intelligence, Apache Spark, analytics

OPEN ACCESS

Edited by:
Davide La Torre, SKEMA Business School, Sophia Antipolis Campus, France

Reviewed by:
Andrea Seveso, University of Milano-Bicocca, Italy
Jane Heather, Eastern New Mexico University, United States

****Correspondence:***
Muhammad Babar
muhammad.babar@aiuo.edu.pk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *27 March 2021*
Accepted: *09 June 2021*
Published: *06 July 2021*

Citation:
Tariq MU, Babar M, Poulin M, Khattak AS, Alshehri MD and Kaleem S (2021) Human Behavior Analysis Using Intelligent Big Data Analytics. Front. Psychol. 12:686610. doi: 10.3389/fpsyg.2021.686610

INTRODUCTION

Intelligent big data analysis is an evolving pattern in the age of data science, big data, and artificial intelligence (AI). Data has been the backbone of any enterprise and will do so moving forward. Storing, extracting, and utilizing data has been key to any operations of a company (Little and Rubin, 2019). When there were no interconnected systems, data would stay and be consumed in one place. With the onset of Internet technology, the ability and requirement to share and transform data have been exploited (Maceli, 2020). With the spread of social media, the nature of data has changed. Social media can consist of billions of users who continuously provide their digital traces with incredible velocity (Kumar et al., 2018). As the data comes from many sources

and in an unstructured format, it is not easy to handle in traditional relational databases. The need for handling unstructured data gives birth to another type of data called big data, which is unstructured, semi-structured, and unpredictable (Iqbal et al., 2020). This data is created real-time, and the amount of data is increasing daily. The data generated from these social media sites can take the form of text, images, videos, and documents. Only structured data can be processed and stored using an RDBMS. Big data is used to process data with a huge volume that is not possible to process using old database techniques and traditional relational databases, within an acceptable processing time.

Big data is characterized by a large volume of data with a large variety and higher velocity (Wang et al., 2020). Data generated moves through cables, either TV or internet, and data on local TV cables broadcast with large volume, variety, and velocity. The amount of data generated every day in the world is increasing exponentially. The rate of data growth is surprising, and this data comes at a speed, with variety (not necessarily structured), and contains a wealth of information that can be key for gaining an edge in competing businesses. The ability to analyze this massive amount of data brings a new era of innovation, productivity growth, and consumer surplus. "Big data is the term for a collection of data sets so large and complex that it becomes difficult to process it using traditional database management tools or data processing applications" (Cui et al., 2020). The challenges include capturing, curating, storing, searching, sharing, transferring, analyzing, and visualizing this data. This section discusses the related literature.

Big data is described with 5V's instead of 3V (volume, velocity, and variety) and included veracity and value (Grover et al., 2020). The widely known big data examples are social networking sites, such as Facebook, YouTube, Dailymotion, Google, and Twitter (Drosos et al., 2015). These sites receive a tremendous amount of data regularly with different variety, velocity, and veracity. The data include value as well. As the number of users increases, the amount of data also increases day by day. Users and data both keep growing on these sites, and this amount of data is a big challenge for owners and companies. This data contains all useful information that needs to be processed in a concise period. To generate more revenue and increase sales, the companies need the processed and analyzed data. The analysis of this data is not possible through relational or traditional database systems within a given time frame as the resources of this traditional system are not sufficient to accomplish processing and storing this huge amount of data; hence, Hadoop comes into the existence for fulfilling this need. In recent years, a large amount of unstructured data is generated from social media sites, such as Facebook, Twitter, Google, and some Dailymotion forums in the form of images, text, videos, and documents, to access and analyze this type of data, this work is best for practicing in the entire field (Xia et al., 2018). Twitter and Facebook are some of the most famous social media platforms, and the companies find that it is very crucial for obtaining customer feedback and maintaining goodwill.

Dailymotion is one of the best video-sharing social media websites. It is a viral platform that publishes community feedback through its videos and comments, likes, dislikes, published videos, and subscriber information for a particular channel (Stieglitz et al., 2018). The analysis of this type of data is important for acquiring knowledge about users, categories, and interests of users. Most of the production companies have their channels to share daily their movie trailers for getting user feedback before releasing them to the general public. Furthermore, individual users upload their videos to get more subscribers and views. These data points are critical for owners to analyze data to understand the views and feelings of customers about their video and service. Dailymotion has billions of users, who watch hours of videos on their site and generate a massive amount of views (Carlinet et al., 2012). It is estimated that more than a hundred hours of videos are watched per minute, and this amount is increasing day by day. To analyze such a huge amount of data, relational databases are not applicable. Users can use this data to understand how much their marketing program is effective. They can check their view counts and subscribers based on the date range that will show them the peak and downtime of views in a particular time. This will also help to check social trends and behavior of people over time (Lee and Kotler, 2011). For example, users can check how many views their videos have received and how much people have liked their video or product. They can also analyze likes and dislikes from the diverse nature of people around the world.

In this research, we utilized Apache Spark to process datasets of social media. Apache Spark is a parallel and distributed platform that overcomes the challenges faced by the traditional processing mechanisms. The main objective of the project is to demonstrate the use of Apache Spark parallel and distributed framework technologies with other storage and processing mechanisms. The social media data generated from Dailymotion is taken under consideration in this article.

LITERATURE REVIEW

A framework is proposed for computing fast and reliable data analysis and mining feedbacks (Rodrigues and Chiplunkar, 2018). They give the real-time Twitter data input in the framework for getting the results of the analysis to generate fast feedback through sentiment analysis. As per Rodrigues and Chiplunkar (2018), the accuracy of data analysis results is essential, and the Hadoop framework provides more than 84% of results when data is produced from social media. Twitter data is one of the largest social media networks where data is increasing daily (Rodrigues et al., 2017). The researcher used data analysis using the "InfoSphere Big Insights" tool, which is very suitable for enterprise companies to use the power of Hadoop in real-time data analysis. The data analytics in Blomberg's work are beneficial for companies to collect customer feedback and details of current trends (Blomberg, 2012). Many big companies, such as Airlines and some other related companies, use these analytics to reach their customers based on their feedback. For crime investigation, cyber-crime people search individuals who have committed the crime.

An architecture is proposed for the sentiment analysis of Twitter by using Hadoop components simply called the ecosystem of Hadoop (Mahalakshmi and Suseela, 2015). It provides the mechanism of Tweets analysis on clusters of Hadoop. It also provided a complete pictorial form of data from various users and their tweets. Recently, newspapers are not read as often and people use television and the internet for most sources of information. Furthermore, many tasks are now done online, such as trading stocks. Buying and selling of shares can be done through the internet from a single laptop or even through mobile (Khan and Khan, 2018). Customers watch every second trend of the stock exchange through their mobile. In this way, they are aware of market fluctuations. To predict the market, Hadoop is used for the analysis of real-time data. The industries and academics deal with a considerable amount of data and perform analysis on terabytes and even petabytes of data. To access their desired result, they use the Hadoop ecosystem and MapReduce to distribute work around various clusters (Dubey et al., 2015). This project is based on a stock market prediction based on Hadoop. They use Hive commands to create Hive tables to load data.

This is the era of technology, only few people use newspapers and other old media for trading on the stock exchange. Because of mobile technologies, users can directly buy or sell their shares from the online stock market. Also, users get every second update through their mobile (Jose et al., 2019). Hence, investors also used these technologies to discuss trade, market status, and dealing with security issues. This type of data is collected in the form of big data. Similarly, when planes fly, they keep transmitting data to headquarters or airbases. The air traffic control uses this data to track and monitor the current position and status of the flight. All this information is processed on a real-time basis. Since multiple air crafts transmit data regularly, the amount of transmitted data received by the flight controller is enormous, and it is accumulated in a vast volume within a concise time (Barros and Couto, 2013). It is a very challenging task to manage and process this massive amount of data called big data. In this study, the researcher demonstrates the methods to process this type of data.

Hundreds or even thousands of airline flights are canceled every year, which costs more money to passengers and owners. Many airlines are canceled due to bad weather conditions. Using Hadoop and MapReduce, the historical prediction can be maintained, predicting the delay and cancelation of a flight from historical data of weather and airlines (Patgiri et al., 2020). The historical dataset was taken to perform operations using pig and MapReduce, which produce output predictions based on temperature, snowfall, lousy weather, and many other factors. It also predicts the influence of cost due to delays and the cancelation of a flight. A model is proposed that determines the total number of flights canceled during 2012–2014, and their analysis is broken into months of each year. Researchers also analyzed the results of all flights diverted during each month of the year between 2012 and 2014.

The trend analysis is also analyzed for e-commerce websites. Using this project, we can easily find the trend of fashions, technologies, and music that varies from one geographical location to another (Satish and Kavya, 2017). Through trend analysis, companies can think of new products based on the needs of the customer, and they can do good strategic planning based on these trends. Amazon is one of the big e-commerce websites where people worldwide visit and see newly added products (Kaushik et al., 2018). The trend analysis is used to check the upcoming events all over the world. New trends come in fashion, living standards, traveling through cars, and many more. Hadoop is used to analyze this trend in this project and depending on these trends and upcoming events, new products were added. The search keywords from Google were taken and analyzed using Hadoop for finding occasional and even periodic events. Through these analyses, it is important to increase sales and attract an audience. This project will focus on data generated from Dailymotion for data mining and processing to make decisions to check their product market value. To accomplish this target, Hadoop, the distributed file system, is used.

The Hive is utilized to analyze temperature data and apply processing on 800,000 records (Lydia and Swarup, 2016). This analysis is done through the Hive query language, shortly, called HQL commands. This project supports in applying HQL commands for analyzing the data. Some of the common commands which are used are given below. Apache has implemented MapReduce, which is very time-consuming because of needed skills in programming languages,

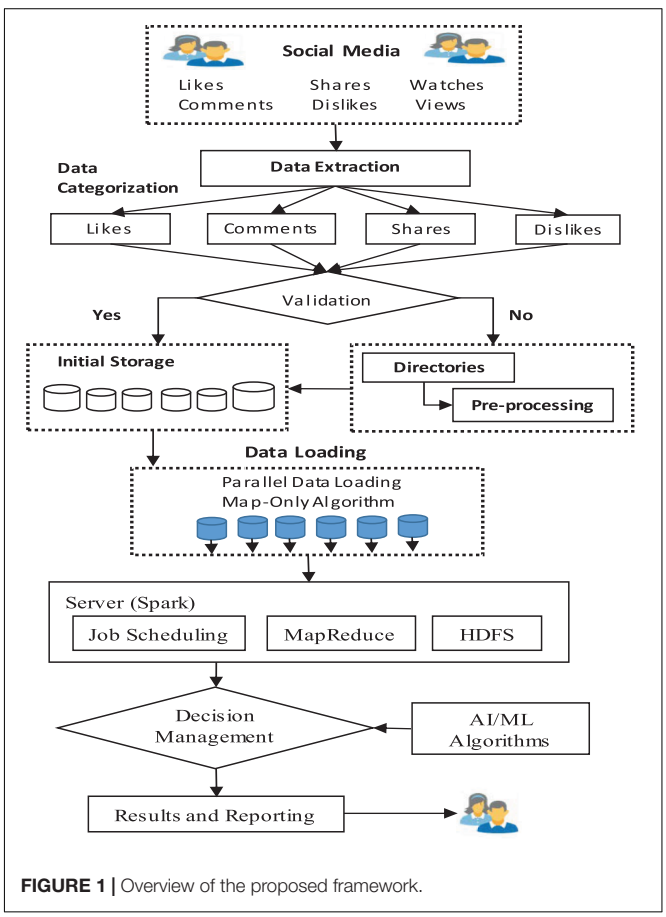

FIGURE 1 | Overview of the proposed framework.

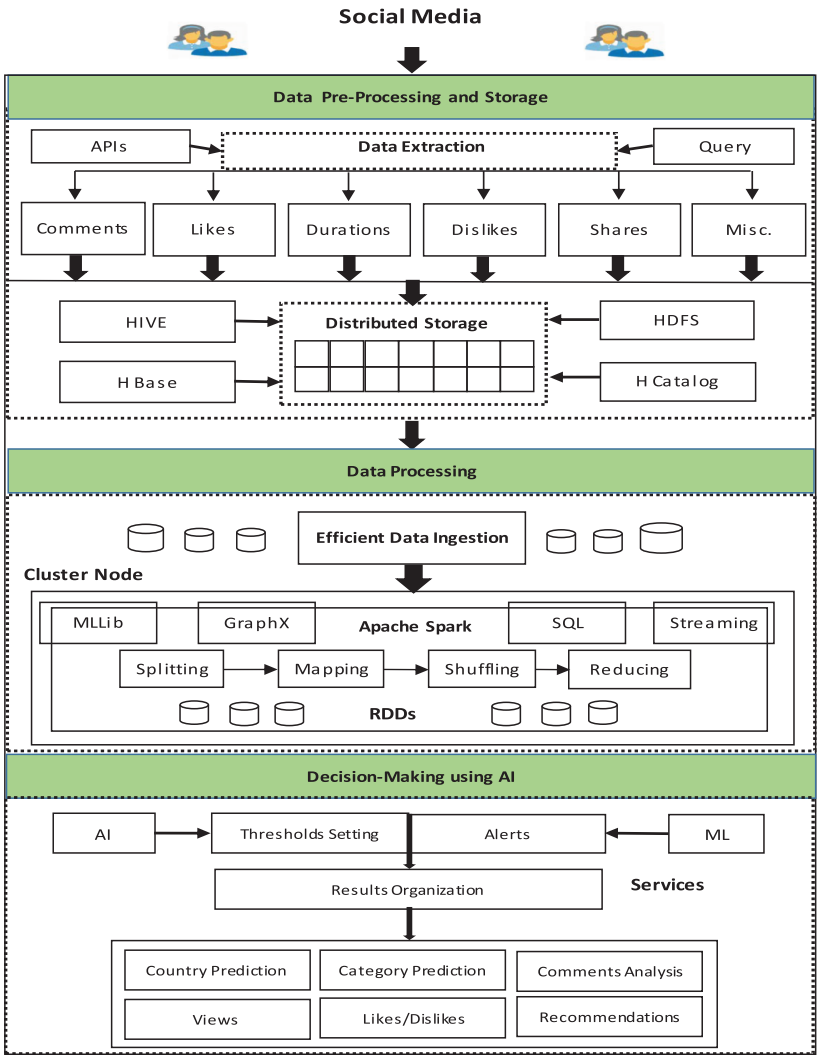

FIGURE 2 | Proposed architecture.

such as Jave (Lydia et al., 2016). The social media platform implemented Hive for its query-based features and similarity with SQL commands. For Warehousing projects, Hive is highly recommended (Capriolo et al., 2012; Salehi and Bernstein, 2018). With an increase in working remotely today, people worldwide can more easily work in one team, allowing multiple experts from different fields and domains, where they can input different types of data. MapReduce has no built-in support of the iterative type of programs; whereas Hadoop allows for processing iterative type of programs and applications from the Hadoop Program without any modification (Paul et al., 2016).

PROPOSED FRAMEWORK

An overview of the proposed framework is given in **Figure 1**. The framework is a parallel and distributed framework. Initially, the data related to a particular video is extracted and the video is scraped. The extracted data is recorded and aggregated in a specific format. Initially, the dataset is checked for anomalies and perm pre-processing. Afterward, the data is loaded into the proposed system using the parallel mechanism to speed up the data ingestion process. The processing of data is carried out by using the Apache Spark framework. The processed data is further utilized for decision-making using machine learning and AI approaches. Finally, the report is provided for decision-making. The detailed architecture of the proposed framework is depicted in **Figure 2**. The proposed architecture is composed of three layers: data pre-processing and storage, data processing, and decision management. A detailed description of the different layers is provided in the upcoming section.

Pre-processing and Data Storage

The application programming interface (API) of Dailymotion is utilized to extract data from a particular channel through

a specific set of queries. This project focuses on fetching data of a particular channel of Dailymotion using its API. We use the Dailymotion developer console to get a unique access key for fetching Dailymotion public channel data. The data is extracted in the form of a CSV file. The CSV file contains all the information about the channel and videos on that channel. The data available in the CSV file contain several anomalies including noise, corrupt data, denormalize data, duplicate values, and null values. Therefore, there is a need for preprocessing techniques to remove the anomalies. The proposed framework

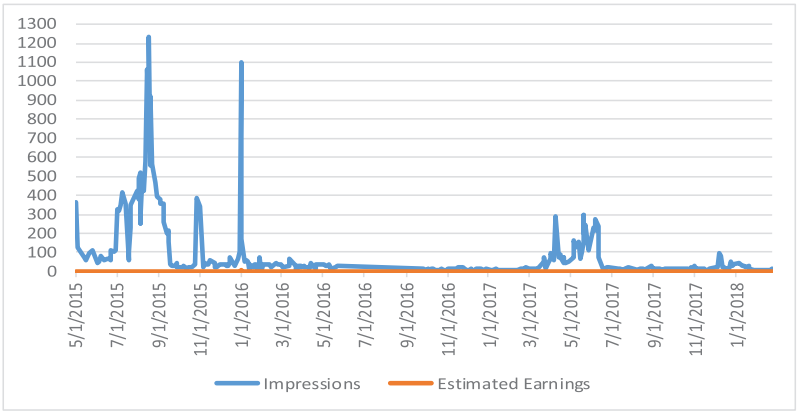

FIGURE 3 | Impressions vs. estimated earnings.

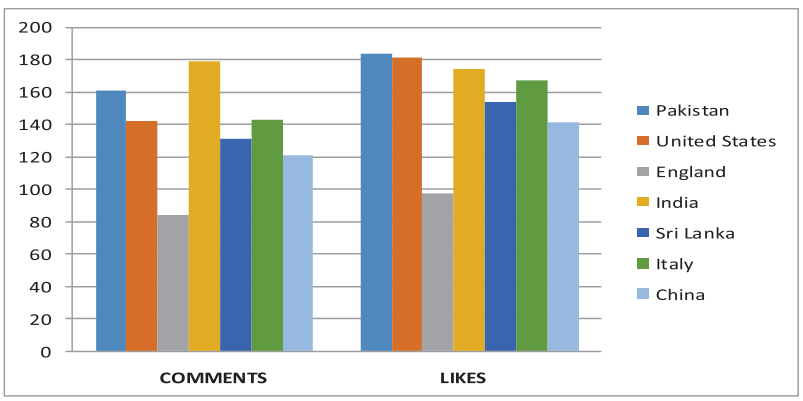

FIGURE 4 | Country-wise comments and likes.

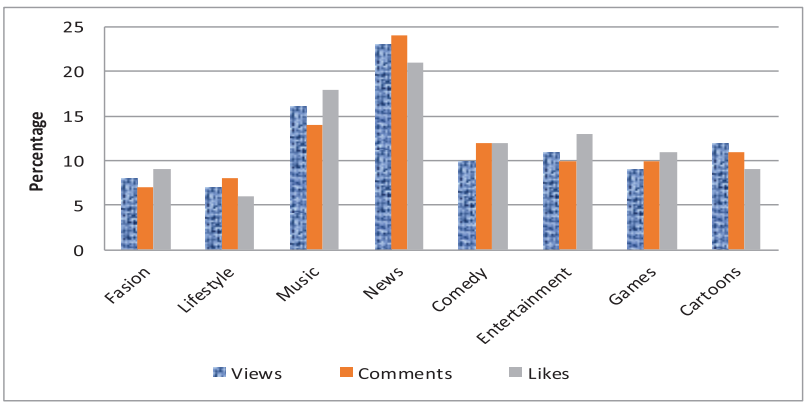

FIGURE 5 | Category-wise detail.

utilized data cleaning, data transformation, data normalization, and data integration. The accuracy of information relies on the recognition and removal of meaningless data.

The noise identification is done before noise removal. The data cleaning includes the detection and rectification of the imprecise data. The normalization is used to transform variables in data into specific series. The transformation is performed by converting the format of available data into a suitable format of processing. The big data must be stored in a specific systematic mechanism to process it efficiently. The proposed architecture utilizes the Hadoop Distributed File System (HDFS) distributed storage mechanism to store huge and gigantic datasets. HDFS grips a huge quantity of data and offers access at ease. The big datasets are stored across many nodes to be processed in parallel.

The Hive storage mechanism is also utilized and integrated with HDFS. The reason for the utilization of the Hive storage is the compatibility of CSV files with Hive that makes the loading process easy. The data is initially extracted in the text file that is in the form of unstructured data. To process analysis techniques, specific delimiters on CSV files are defined to load into Hive. It also works as an interface for data warehousing of Apache Hadoop-based data. It is a data warehousing infrastructure developed on top of Hadoop that allows querying data for data analysis. The CSV data is converted to Optimized Row Columnar (ORC) data and then loaded into the Hive table. A Dailymotion data table is created with a specific set of required fields. The H-catalog is used as a table storage management tool that processes the Hive tabular data into the Hadoop application for processing. The H-catalog is built on top of Hive that incorporates Hive data definition. Hive enables users to treat a file as an Structured Query Language (SQL) table with rows and columns. It provides read and write interfaces for Hadoop technologies.

Data Processing Using Parallel Framework

The data processing of huge datasets is the key module of the proposed model. An integrated approach is used to process the

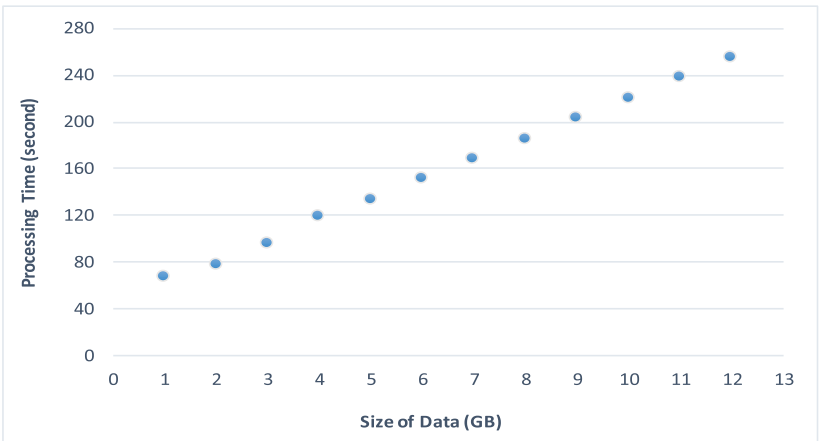

FIGURE 6 | Processing time.

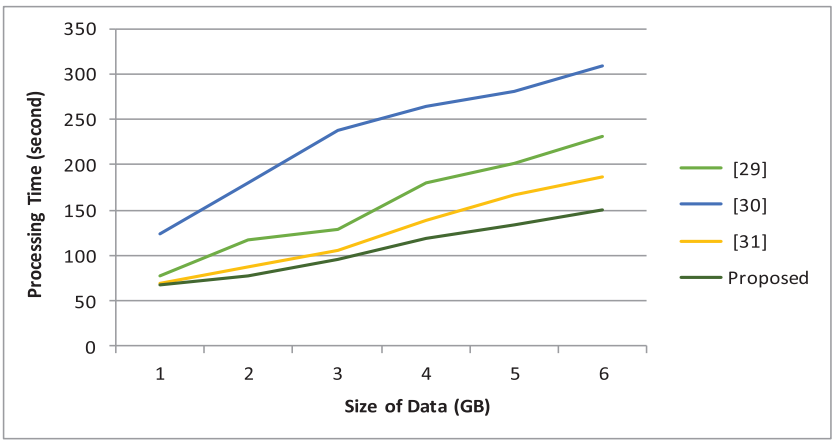

FIGURE 7 | Comparative analysis.

big data. Special storage techniques are taken into consideration for efficient processing. HDFS and Hive storage techniques are integrated to achieve optimal distributed storage. The Apache Spark parallel and distributed framework are applied for fast and real-time stream processing of big data. The programming paradigm utilized by Apache Spark is the MapReduce paradigm. The MapReduce is the rationale for parallel functional processing. The data is loaded into the Spark framework using a parallel mechanism (e.g., map-only algorithm). Apache Spark maps the complex queries with MapReduce jobs for simplifying the complex process. The queries of Spark can be mapped into the phases of the MapReduce framework. Spark SQL handles the selection operations. Spark is a master-slave architecture, and the overall cluster is managed by the Spark master node. The proposed Spark architecture processes the data based on Resilient Distributed Dataset (RDD). An RDD is a distributed collection and immutable that can be wrought on in parallel. The RDD includes an object and is produced by ingesting an external dataset. The data collected from billions of customers is utilized as an actionable metric to perform better decision-making and get more customer satisfaction. The input is categorized into region, likes, duration, etc. The regional data is then analyzed to check the views from different regions and countries. The detail of the viewer and watch time are noted for future decisions. The likes of each video are analyzed to check the interest of the viewer. The daily view and comments analytics are created by running the queries on imported data.

Decision Management

The decision management layer is a bridge between the proposed architecture and the outer world. It utilizes AI and ML algorithms. The thresholds are set using AI to analyze the specific dataset. The users are alerted using the AI mechanism. Based on the output, companies decide the enhancement of their investment decision-making using AI. The decisions can be utilized to market the projects. The proposed system utilizes the Dailymotion data to market the products based on region, country, and even based on a particular interest of users. Companies can find the peak and slow time of their viewership through a share, view count, subscriber, and audience retention. The companies can also find the trending product at a particular time. The changing behavior of people can be an important insight of companies.

RESULTS AND DISCUSSION

This section describes the implementation detail and results. This project focuses on fetching data of a particular channel of Dailymotion using its API. We use the Dailymotion developer console to get a unique access key for fetching Daily motion public channel data. The data is extracted in the form of a CSV file. The CSV file contains all the information about the channel and videos. After getting the API key, the.Net (C#) console application can be developed for fetching information based on search criteria. A text file will be generated by using this program, which will then be loaded from HDFS into the Hive database. In this project, we fetch YouTube data of a specific channel using API. We used Google Developers Console and generated a unique access key required to fetch YouTube public channel data. Once the API key is generated, a.Net (C#) based console application is designed to use the Dailymotion API for fetching video information based on search criteria. The text file output generated from the console application is then loaded from the HDFS file into the Hive database. The user can directly interact with HDFS using various commands. The queries will be run on big data through Hive to get the required data. This data will then be used by management for analysis. Besides, Apache Spark 3.0 is utilized for real-time stream processing of big data. The pyspark library is used for the implementation of spark workers. The MLLib library is utilized for applying the Machine Learning (ML) algorithm in the spark context. The graphX library is utilized for graph implementation.

We analyze the data and perform various operations to find the number of comments on the particular video and also the person who has uploaded the video. The dataset utilized contains the channel ID, category, duration, view count, comment count, like count, and country code. Dailymotion also provides video monetization options for its users, and most Dailymotion users have their channels with a monetized video that generates revenue for them through video ads. We extracted a CSV file from Dailymotion, and then uploaded it on Hadoop HDFS storage to analyze. The extracted file contains some meaningless information. The final file contains three columns: date, number of impressions, and earnings. We have generated the report of earnings within the particular time frame, and the detailed sum of an impression on a video is shown in **Figure 3**. The country-wise comments and like counts are shown in **Figure 4**. The category-wise detail of views, comments, and likes are illustrated in **Figure 5**.

Figure 6 demonstrates the processing time of the proposed architecture. Besides, the comparative analysis of the proposed architecture with state-of-the-art is demonstrated in **Figure 7**.

CONCLUSION

The use of big data in the field of social media is essential. The organizations that use big data have a huge advantage over the one which is still practicing relational database techniques. These organizations better know the importance of big data than the one which has no big data implementation. This product is intended to show the data analysis of Dailymotion and some key results. This article proposed a model using Apache Spark. The proposed architecture is three-layered architecture. The main objective of this project is to demonstrate the use of Apache Spark parallel and distributed framework technologies with other storage and processing mechanisms. The effectiveness of the proposed

architecture is also highlighted. In this way, many other features can be determined, and the company could know the details of its competitor and clients. If a company uploads its marketing video on Dailymotion, its video becomes more prominent than the base of views and likes.

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

MT: idea and logic. MB: writer, logic, and implementation. MP: supervision and review. AK: review and drafting. MA: review and implementation. SK: drafting and review. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

Taif University Researchers Supporting Project number (TURSP-2020/126), Taif University, Taif, Saudi Arabia.

REFERENCES

Barros, C. P., and Couto, E. (2013). Productivity analysis of European airlines, 2000–2011. *J. Air Transp. Manag.* 31, 11–13. doi: 10.1016/j.jairtraman.2012.10.006

Blomberg, J. (2012). *Twitter and Facebook Analysis: It's Not Just for Marketing Anymore*, Vol. 309. Denver, CO: SAS Global Forum.

Capriolo, E., Wampler, D., and Rutherglen, J. (2012). *Programming Hive: Data Warehouse and Query Language for Hadoop*. Sebastopol, CA: O'Reilly Media, Inc.

Carlinet, Y., Huynh, T. D., Kauffmann, B., Mathieu, F., Noirie, L., and Tixeuil, S. (2012). "Four months in daily motion: dissecting user video requests," in *Proceedings of the 2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC)*, (Piscataway, NJ: IEEE), 613–618.

Cui, Y., Kara, S., and Chan, K. C. (2020). Manufacturing big data ecosystem: a systematic literature review. *Robotics Comput. Integr. Manuf.* 62:101861. doi: 10.1016/j.rcim.2019.101861

Drosos, D., Tsotsolas, N., Chalikias, M., Skordoulis, M., and Koniordos, M. (2015). "A survey on the use of social networking sites in Greece," in *Creativity in Intelligent, Technologies and Data Science*, eds A.G. Kravets, P. Groumpos, M. Shcherbakov, M. Kultsova (New York, NY: Springer International Publishing), 556–570.

Dubey, A. K., Jain, V., and Mittal, A. P. (2015). "Stock market prediction using hadoop map-reduce ecosystem," in *Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (Piscataway, NJ: IEEE), 616–621.

Grover, V., Lindberg, A., Benbasat, I., and Lyytinen, K. (2020). The perils and promises of big data research in information systems. *J. Assoc. Inf. Syst.* 21:9.

Iqbal, R., Doctor, F., More, B., Mahmud, S., and Yousuf, U. (2020). Big data analytics: computational intelligence techniques and application areas. *Technol. Forecast. Soc. Change* 153:119253. doi:10.1016/j.techfore.2018.03.024

Jose, J., Mana, S. C., and Samhitha, B. K. (2019). An efficient system to predict and analyze stock data using hadoop techniques. *Int. J. Recent Technol. Eng.* 8, 2277–3878.

Kaushik, K., Mishra, R., Rana, N. P., and Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: a study of helpful reviews on Amazon.in. *J. Retail. Consum. Serv.* 45, 21–32. doi: 10.1016/j.jretconser.2018.08.002

Khan, J., and Khan, I. (2018). The impact of macroeconomic variables on stock prices: a case study Of Karachi Stock Exchange. *J. Econ. Sustain. Dev.* 9, 15–25.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). "Benchmarking aggression identification in social media," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (Santa Fe, NM: Association for Computational Linguistics), 1–11.

Lee, N. R., and Kotler, P. (2011). *Social Marketing: Influencing Behaviors for Good*. Thousand Oaks, CA: Sage Publications.

Little, R. J. A., and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, Vol. 793. Hoboken, NJ: John Wiley & Sons.

Lydia, E. L., and Swarup, M. B. (2016). Analysis of big data through hadoop ecosystem components like flume, mapreduce, pig and hive. *Int. J. Comput. Sci. Eng.* 5, 21–29.

Lydia, E. L., Swarup, M. B., and Laxmi, M. V. (2016). A literature inspection on big data analytics. *Int. J. Innov. Res. Eng. Manag.* 3.

Maceli, M. (2020). Internet of things in the archives: novel tools for environmental monitoring of archival collections. *Rec. Manag. J.* 30, 201–220. doi: 10.1108/rmj-08-2019-0046

Mahalakshmi, R., and Suseela, S. (2015). Big-SoSA: social sentiment analysis and data visualization on big data. *Int. J. Adv. Res. Comp. Commun. Eng.* 4, 304–306.

Patgiri, R., Hussain, S., and Nongmeikapam, A. (2020). Empirical study on airline delay analysis and prediction. *arXiv* [Preprint]. https://arxiv.org/abs/2002.10254

Paul, A., Ahmad, A., Rathore, M. M., and Jabbar, S. (2016). Smartbuddy: defining human behaviors using big data analytics in social internet of things. *IEEE Wirel. Commun.* 23, 68–74. doi: 10.1109/mwc.2016.7721744

Rodrigues, A. P., and Chiplunkar, N. N. (2018). Real-time twitter data analysis using hadoop ecosystem. *Cogent Eng.* 5:1534519. doi: 10.1080/23311916.2018.1534519

Rodrigues, A. P., Rao, A., and Chiplunkar, N. N. (2017). "Sentiment analysis of real time Twitter data using big data approach," in *Proceedings of the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, (Piscataway, NJ: IEEE), 1–6.

Salehi, N., and Bernstein, M. S. (2018). "Hive: collective design through network rotation," in *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 2, (New York, NY: ACM), 1–26. doi: 10.1145/3274420

Satish, K. V. R., and Kavya, N. P. (2017). "Hybrid optimization in big data: error detection and data repairing by big data cleaning using CSO-GSA," in *Proceedings of the International Conference on Cognitive Computing and Information Processing*, (Singapore: Springer), 258–273. doi: 10.1007/978-981-10-9059-2_24

Stieglitz, S., Mirbabaie, M., Ross, B., and Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.* 39, 156–168. doi: 10.1016/j.ijinfomgt.2017.12.002

Wang, J., Yang, Y., Wang, T., Sherratt, R. S., and Zhang, J. (2020). Big data service architecture: a survey. *J. Internet Technol.* 21, 393–405.

Xia, Q., Yin, X., He, J., and Chen, F. (2018). Real-time recognition of human daily motion with smartphone sensor. *Int. J. Performability Eng.* 14, 593–602.

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 Psychology

SYSTEMATIC REVIEW
published: 08 July 2021
doi: 10.3389/fpsyg.2021.686624

Achieving Operational Excellence Through Artificial Intelligence: Driving Forces and Barriers

Muhammad Usman Tariq, Marc Poulin and Abdullah A. Abonamah*

Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates

This paper presents an in-depth literature review on the driving forces and barriers for achieving operational excellence through artificial intelligence (AI). Artificial intelligence is a technological concept spanning operational management, philosophy, humanities, statistics, mathematics, computer sciences, and social sciences. AI refers to machines mimicking human behavior in terms of cognitive functions. The evolution of new technological procedures and advancements in producing intelligence for machines creates a positive impact on decisions, operations, strategies, and management incorporated in the production process of goods and services. Businesses develop various methods and solutions to extract meaningful information, such as big data, automatic production capabilities, and systematization for business improvement. The progress in organizational competitiveness is apparent through improvements in firm's decisions, resulting in increased operational efficiencies. Innovation with AI has enabled small businesses to reduce operating expenses and increase revenues. The focused literature review reveals the driving forces for achieving operational excellence through AI are improvement in computing abilities of machines, development of data-based AI, advancements in deep learning, cloud computing, data management, and integration of AI in operations. The barriers are mainly cultural constraints, fear of the unknown, lack of employee skills, and strategic planning for adopting AI. The current paper presents an analysis of articles focused on AI adoption in production and operations. We selected articles published between 2015 and 2020. Our study contributes to the literature reviews on operational excellence, artificial intelligence, driving forces for AI, and AI barriers in achieving operational excellence.

Keywords: operational excellence, artificial intelligence, driving forces, barriers, artificial intelligence operations

OPEN ACCESS

Edited by:
Davide La Torre,
SKEMA Business School,
Sophia Antipolis Campus, France

Reviewed by:
Jane Heather,
Eastern New Mexico University,
United States
Paola Adinolfi,
University of Salerno, Italy

****Correspondence:***
Muhammad Usman Tariq
m.tariq@adsm.ac.ae

Specialty section:
This article was submitted to Organizational Psychology, a section of the journal Frontiers in Psychology

Received: *29 March 2021*
Accepted: *09 June 2021*
Published: *08 July 2021*

Citation:
Tariq MU, Poulin M and Abonamah AA (2021) Achieving Operational Excellence Through Artificial Intelligence: Driving Forces and Barriers. Front. Psychol. 12:686624. doi: 10.3389/fpsyg.2021.686624

INTRODUCTION

Artificial intelligence is a technological concept in operational management, philosophy, humanities, statistics, mathematics, computer sciences, and social sciences. Artificial intelligence aims to create computers or machines to carry out jobs that generally need human intelligence. The sub-discipline of artificial intelligence is machine learning, which directs to statistical learning. Machine learning aims to create algorithms that can automatically manage information in actual-time and enhancing experience without being unequivocally customized. Supply chains

are experiencing advantages from investments and progressive interest in artificial intelligence technologies (Voronkova, 2019). The latest information systems, such as wireless technologies, the internet of things, affordable sensors, and cloud storage, act as the underpinning technologies of artificial intelligence. Today, it is quite feasible that business processes and value chains are connected within and across organizations. Organizations can use smart devices, mobile applications, and point of sale technologies to accumulate geographic, demographic, and behavioral customer data in real time that helps develop products and services. The applications can help improve business functions using robotics and automatic systems. They allow marketing to forecast and understand customer's demands more precisely. There is an importance of responsiveness within supply chain procedures as consumers demand custom-made products and peculiar services expeditiously. Supply chains and worldwide production network systems are experiencing effective modifications in the progressive business environment. The evolution of new technological procedures and advancements in production intelligence is producing positive impact on decisions, operations, strategies, and management. Businesses develop various methods and solutions to extract important information, such as big data with automatic production capabilities, and systematization for business affiliation. It helps to recognize barriers in enhancing organizational performance, such as equipment management, defect identification, time reduction of the cycle, speculation of demand, bioinformatics, and human resource (Deivanathan, 2019). Innovation in product development and smart technologies allows production intelligence to use soft computing, advanced algorithms, decision technologies, and findings. That can be in different information systems for advanced production systems, advanced equipment control, engineering data analysis, enterprise resource planning (ERP), manufacturing execution system, and supply chain management to improve decision excellence and effectivity of management. Smart production has become the trend with the logical incorporation of decision technologies and artificial intelligence to adopt the latest information technology (Mühlroth and Grottke, 2020).

Operational excellence is a concept from the eighteenth century covering a subject's productivity, labor division, and the free market. Any organization's success depends on operational excellence as it relates to the organization's functions serving consumers. This concept is at the first operating stage that is a short distance away from the resources involved in its functions but is crucially associated with its planning. The three important sections in operational excellence are effectiveness, right the first time, and efficiency of procedures (Mangla et al., 2020). The concept of operational excellence could entail adopting industrial tasks and theories, like Industry 4.0, Reverse Logistics, Lean Six Sigma, Business Procedure Reengineering, and the Internet of Things, which are enablers of acquiring accurate results. Improving organizational competitiveness is evident through firm's proper decisions and represents working efficiency (Danaher, 2018). Improvements and development resulting from operational excellence relate to revolution, the latest technologies, and solutions. Innovation typically permits small business operating expenses and increased revenues (Carvalho et al., 2019). Acquiring operational excellence relates to adopting individual management theories and techniques that grant an adequate cost level to be justified and innovativeness, leading to investments and persistent enhancement procedures through problem-solving techniques (Jamshidieini et al., 2017).

There are four drivers at the foundation of attaining operational excellence. The first one relates to the organization's vision that explains the requirements. The engagement of people in the strategy implementation is the second driver. The third driver involves creating the proper process performance metrics to support the organizational strategy. The last driver is the technology used to support the required processes.

Advanced algorithms are one important type of technology service as a driver for operational excellence. The development of an advanced algorithm helps forecast customer needs. If an algorithm is available, competitors may eventually obtain it or develop a similar tool to maintain their competitive position. The firms that do not adopt similar algorithms will be at a great disadvantage and may lose their market value. Advanced algorithms are essential for companies to staying in business within highly competitive markets. Integrating big data technologies enables the evolution of productive algorithms to understand the customer needs, which can be exceptionally beneficial for decision makers (Shehadeh et al., 2016).

A framework of the core functionalities for operational excellence is proposed in **Figure 1**. The figure depicts a better understanding of the role of artificial intelligence in operational excellence. Different operational excellence sections are distributed among the core functionalities, including performance management, employee engagement, process management, strategy development, organizational planning, and improvement initiatives. Artificial intelligence enhances the core functionalities of operational excellence.

FIGURE 1 | Operational excellence core functionalities.

The framework of artificial intelligence with operational excellence core functionalities is proposed in **Figure 2**. Artificial intelligence reshapes operational excellence by using different core functionalities to improve organization operations. Different automated intelligent algorithms find the patterns among the operations excellence's different functions to automatically process the information. The outcome of artificial intelligence processing and real-time data can improve the organization's operational decisions for achieving excellence. Artificial intelligence seeks to automatically select the right operational path connecting the various process within a business.

The current study focuses on the following research questions:

1. What is the connection between operational excellence and artificial intelligence?
2. What are the driving forces for achieving organizational performance by using artificial intelligence?
3. What are the barriers to achieving organizational performance using artificial intelligence?

This paper comprises six sections. The second section investigates the literature about artificial intelligence, operational excellence, and driving forces and barriers. In the third section, there is an explanation of the methodology used to perform this research. Outcomes and interpretations of this study are in the fourth and fifth sections, respectively. In the sixth section, we present the discussion, conclusion, and future research directions.

LITERATURE ANALYSIS

Information technology is constantly growing with its application in various areas, including the educational field, healthcare field, or human resource field. Artificial intelligence and virtual reality are branches of computer science and are significant tools for improving human life or sustaining lifetime learning procedures (Stanica et al., 2018). Technology plays a vital role in influencing the social, political, cultural, educational, and organizational sectors in this rapidly changing world. The advancements in technology have made their importance valuable for all the sectors, subsequently increasing productivity by practical training and learning methods (Abduljabbar et al., 2019; Karsenti, 2019). Artificial intelligence is embedded in a computer or device by programming software, which helps to perform complex and specific tasks that were previously possible only with human intelligence. Handling complexity is the key factor while adopting artificial intelligence to solve complex problems. Despite the complexity and being time-consuming, artificial intelligence can perform various jobs in seconds without human's assistance (Becker, 2017; Arrieta et al., 2020). Artificial intelligence is a technology that enhances the daily activities of social and economic life. It positively influences economic growth by solving various social obstacles. Artificial intelligence has recently magnetized the attention of many developed and developing countries, such as the United States, Europe, India, and China. The major focus is on the development of robotic

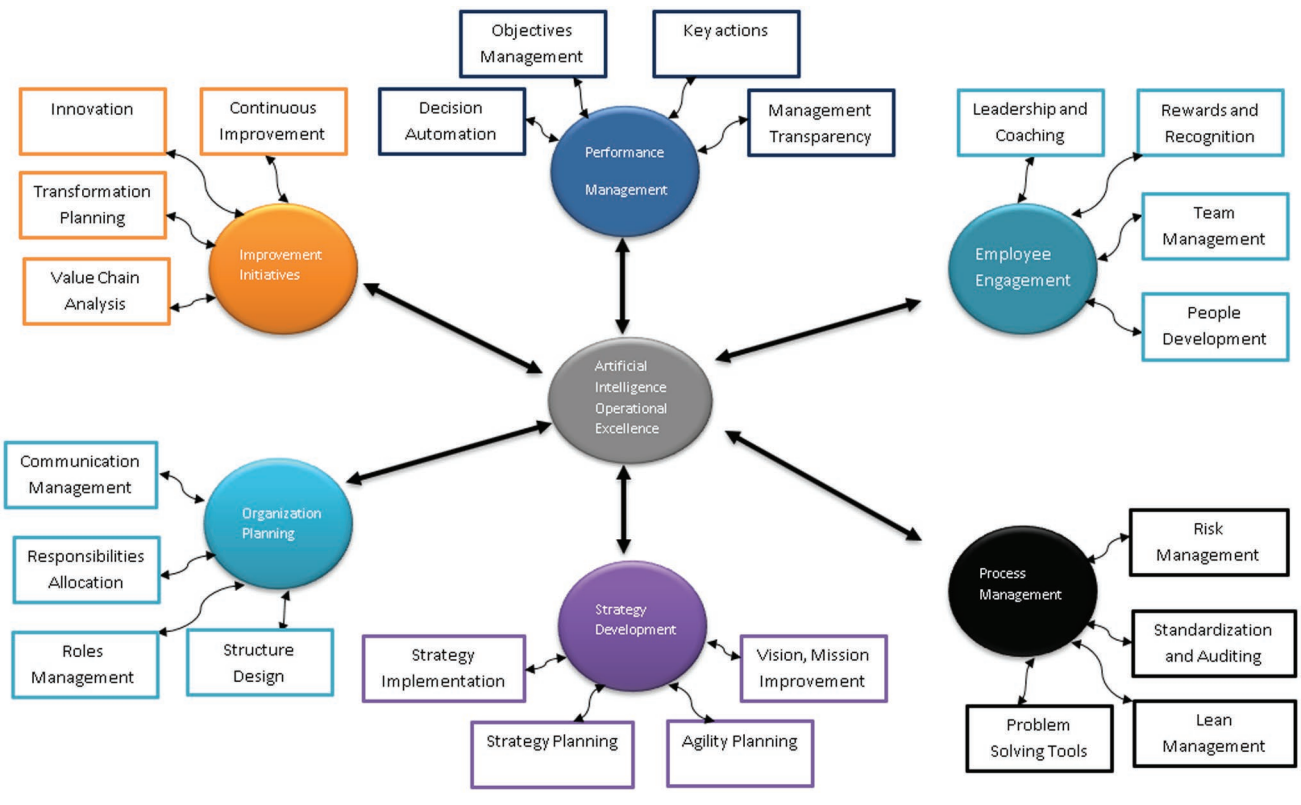

FIGURE 2 | Artificial intelligence-based operational excellence framework.

technology and intelligence information technology. Even though the latest artificial intelligence technology is proving its excellence in obtaining specific models with various barriers, many intelligence information technology models require a self-idea function, rely on big data, and are complex (Skurdauskaitė, 2020).

Adoption of Artificial Intelligence in Organizations

Artificial intelligence have shown to be useful in fulfilling the following requirements (Davenport and Ronanki, 2018).

Automation of Business Procedures

Digitization is captivating and modifies the market in the business sector. Manual workflows depending on paperwork lower firms' manufacturing efforts (Scheer, 2017). The business should organize its internal procedures in an excellent way (Paschek et al., 2017). "Robotic Process Automation" tends to be the software-based solution to regulate business procedures that include daily tasks, systemized data, and determining results (Aguirre and Rodriguez, 2017). Involving robotics within manufacturing is not recent, but their functions and abilities have advanced significantly, surpassing human skills in many circumstances.

Obtaining Insight Through Data Analytics

Information plays a significant role in the decision-making procedures on the operational, strategic, and tactical stages. However, the calculation and accumulation of data within enterprises are rising quickly. Essentially, big data analytics is the use of the latest statistics applied to any type of electronic communication, which may include "messages, updates, images posted to social networks, readings from sensors, and GPS signals from cell phones" (Wamba et al., 2018; Allam and Dhunny, 2019). Big data analytics allows to advance the way large quantity of data are processed.

Rational Customer Engagement

With the rapid advancements in the technological fields, businesses are maintaining personal connections with customers, and brands are gradually pursuing to maintain a connection with the customers on the digital mediums. There is a growth of a broad range of communication procedures on several platforms, such as consumer feedback, sharing videos of the brands on social media platforms, and creating blogs (Eigenraam et al., 2018). With the development in technology and digitization, social media platforms act as a medium to spread the information about the products in both businesses to business and business to consumer organizations (Pansari and Kumar, 2017; Choi et al., 2018).

Smart Agents

Many customers enjoy group-based online shopping. Smart agents based on advanced algorithms can negotiate to lessen the efforts while collecting the buyer's information, transaction costs, and sellers' negotiation. Smart agents can help other models other than C2B when there is a negotiation between buyers and sellers (Liang et al., 2019).

Recommendations About Products and Services

With the emergence of artificial intelligence, there is an evolution in product and service recommendation systems for organizations to increase sales, personalization, and engagement using easy-to-understand images and languages. With the rapid advancements in artificial intelligence, there is a development of concepts and priorities to enhance sales management (Singh et al., 2019).

Employee Engagement

Artificial intelligence can be used to improve employee management in two ways. First, firms can have easy access to a large amount of data related to their business functions to help achieve an effective decision-making procedure. Secondly, artificial intelligence's constant evolution allows organizations to manage and process the data in real time (Robert et al., 2020).

Benefits for Employees

AI can impact employee's physical and emotional engagement. It can provide a direct influence on employee benefits and improve organizational excellence (Alvi et al., 2020). For instance, in this technological era, online survey systems supported by AI can help recognize employee's needs regarding their organization (Kang et al., 2016).

Human Resource Strategies

In this technological era, the human resource information system has a vital role in the decision-making procedure for effective human resource management. A semi-structured and unstructured process of HR decisions is attainable by adopting an intelligent decision support system (IDSS) with the combination of the knowledge discovery database (KDD; Masum et al., 2018).

Safety and Quality Analysis

Safety and quality are the main concerns for governments and the automobile sector. The main technical problems involve validating inductive learning in the modern environment inputs and attaining good levels of reliability needed for complete fleet formation. Moreover, the significant challenge may be creating an end-to-end structure and formation procedure that enhances the safety concerns limitless technical specialties into a combined approach (Koopman and Wagner, 2017).

Operational Excellence

Operational excellence is a concept that focuses on problem-solving techniques and leadership skills as the main factor for continuous development. Firms are usually unsure how to proceed with operational excellence, and most organizations find it too broad or doubtful as it is a complicated concept to explain. The employee's and manager's attitudes are not simply a set of activities that organizations perform (Gólcher-Barguil et al., 2019).

Continuous Improvement vs. Operational Excellence

Continuous improvement is the ongoing attempt to enhance organizational procedures, services, and products. It often occurs gradually instead of occurring instantly through some advanced innovation or constant progression. With continuous improvement, a firm has more chances to sustain and progress on these improvements as processes are built in to assure continuity (Sánchez-Ruiz et al., 2019). Although continuous improvement is significant, it is not sufficient on its own to maximize a firm's improvement.

As a firm continues to clarify its procedures, services, and products, it requires a way to pursue progress (Benzaid and Taleb, 2020; van Assen, 2020). Operational excellence plays a vital part in this stage. Operational excellence is an outlook that accepts some regulations and tools to develop long-lasting improvement within a firm. Firms can achieve operational excellence when every individual in the firm can observe its value (Çaliş and Bulkan, 2015). Firms should diligently attempt to implement changes to seek the observed value. Essentially, operational excellence focuses on eliminating costs or enhancing firm's productivity. It is about developing an organizational culture that will allow the firms to manufacture valuable products and services for the consumers and attain sustainable progress. Operational excellence is a concept that includes adopting appropriate techniques for the right procedures. When this occurs perfectly, the absolute working environment prospers, motivates, and empowers the employees (Sehnem et al., 2019).

Principles of Operational Excellence

Respect Every Person

Every individual is valuable and has potential, so an employee deserves respect. The perfect way to exhibit with respect to the employees and organizations must involve them in requisite activities. It will positively enhance employee's motivational levels to feel more empowered to present their ideas (Sony, 2019).

Lead With Modesty

Modesty includes a desire to listen and accept every individual's suggestion, setting aside that individual's position or status within the organization. Leaders should always lead with modesty (Sony, 2019).

Seek Excellence

Managers must try to simplify the working procedures without compromising on quality. Managers and employees should look for sustainable solutions when any problem arises. It increases perfection that is one of the competitive advantages (Sony, 2019; Dogru and Keskin, 2020).

Accept Scientific Ideas

Continuous thinking and experimentation lead to innovation. It is essential to explore and encourage new ideas without fear of defeat (Sony, 2019).

Focus on the Procedure

When there is a negative outcome from any procedure, managers often blame employees for it. However, the source of the negative outcomes is often because of faults in the design of a procedure. Even the best employees cannot continuously provide exceptional outcomes with flawed procedures. So, instead of blaming employees, it is imperative to obtain an accurate picture of the real cause and make the proper adjustments to achieve the essential outcomes (Sony, 2019).

Ensure Quality

If monitoring is done on every part of a procedure, there is a possibility of achieving high quality. It helps to arrange work areas in a way that will make it possible to become visible. When there is a problem at any stage, it is important to pause the working procedure to solve it (Gólcher-Barguil et al., 2019).

Pull and Flow Value

Every firm's goal is to furnish the utmost value to its consumers. For this purpose, firms should ensure that the procedure and workflow are uninterrupted as disruption cause inefficiencies and waste. There is an importance of evaluating and integrating customer requirements in procedures to ensure a firm fulfills the proper requirements (Chiarini and Kumar, 2020).

Think Logically

There is an interconnection between all the parts working together. It is significant to recognize the connection between all the parts because it will allow them to make decisions. Organizations should avoid a localized vision and eliminate data flow and ideas (Postavaru et al., 2019).

Develop Purpose Constancy

Employees should understand the mission statement and objectives of the firm from day one. Firms should focus on these objectives continuously. Employees should be aware of the changes and goals and seek to achieve those goals. Understanding this will allow the employees to adjust their activities, objectives, and behavior to benefit the organization (Ivanov and Sokolov, 2019).

Value Creation for Customers

Firms must work to recognize the demands and expectations of their consumers. A firm that cannot create and deliver value to its customer does not remain sustainable (Heinonen et al., 2019).

Operational Excellence Methods

Firms can enhance their performance and culture through operational excellence, helping in sustainable progress. Organizations should observe traditional events and look forward to a sustainable change system. Several popular methodologies for achieving operational excellence (Chakraborty et al., 2020) follow:

Lean Manufacturing

This method concentrates on systematically reducing waste during production procedures. It recognizes that every procedure has some restrictions and stresses all the efforts for improvement on those restrictions are the fastest way to success. The basic principles of lean manufacturing concentrate on improving the quality of products and services, reducing anything that is not important, and eliminating overall costs (Jordan and Mitchell, 2015; Kamble et al., 2020). Conventional lean manufacturing recognizes seven areas of waste that are generally known as "seven deadly wastes."

- **Overproduction:** It occurs when employees create something before its requirement. It is an unfavorable form of waste in the form of inventory, which often hides substantive issues.
- **Waiting:** When employees have to wait for the next manufacturing stage, there is no additional value. It can be very surprising and informative to investigate each production stage and calculate the actual time adding value to a product versus non-value-added time.
- **Transport:** It is a waste because the movement of incomplete or undone products does not add value to a product.
- **Motion:** This stage considers all extra movement that does not add value to a product. This often occurs due to poor working processes.
- **Over-processing:** It occurs when there is excess time on processing rather than producing according to customer's requirements. It is one of the most challenging wastes to eliminate, especially when producing a low level of product variety.
- **Inventory:** This kind of waste can occur for several reasons, such as when supply is greater than demand, there are opportunities for quantity discounts from suppliers, and work-in-progress is created due to separation of the production processes, or smoothing of production levels. Although there are reasons to create inventory, the Lean philosophy still recognizes it as a waste.
- **Defects**: This obvious waste occurs when mistakes in production cause parts or products that cannot be used. Either time and resources are wasted to fix the part, or it must be thrown away. In any case, the defective part creates waste (Jarrahi, 2018; Alefari et al., 2020).

Six Sigma

It is a set of techniques and tools to enhance business procedures that help produce better services and products. Six Sigma's objective is to enhance the consistency of customer's experience by recognizing and reducing variation. A Six Sigma organization will seek to create not more than 3.4 defects for every million opportunities. The definition of defect is any product not meeting the standards accepted by customers. DMAIC implementation can help to build Six Sigma business. DMAIC stands for "define, measure, analyze, improve, and control." Following are the steps for this procedure:

- **Define:** In the first step, the organization clearly defines the problem in order to fix it. After identifying the problem, the organization can develop a strategy and evaluate the accessible resources.
- **Measure:** Organizations need to measure all accessible information and investigate the current processes closely with the current procedure. Where is the requirement for improvement? What is functioning correctly?
- **Analyze:** After the organization's measurement of data, they can analyze its findings and find the source of the problem.
- **Improve:** After analyzing the data, organizations need to develop solutions. Solutions are adopted on a low scale to examine the necessary modifications.
- **Control:** After implementing the new procedures, the organization must find a path to maintain the procedure. It is important to ensure continuous improvement to make the procedure effective (Niñerola et al., 2019).

Kaizen

In Japanese, Kaizen means "continuous improvement." It helps to adopt positive, continuing changes in the work environment. Kaizen's leading principles are that the enhanced procedure leads to positive outcomes, group work is significant for success, and some procedures need improvement. Firms adopt Kaizen to assist them in developing a continuous improvement culture. Employees work together to attain ongoing changes in the working environment. Small modifications will combine to develop major outcomes. This method does not certainly only focus on small changes. It also concentrates on all employees' involvement to impact actual change. Kaizen focuses on the importance of continuous improvement. Organizations need to continuously make improvements. Kaizen helps increase employees' efficiency, reduce costs, and enhance customer's experience (Shan et al., 2016).

Achieving Operational Excellence

The ultimate objective of organizations focusing on continuous improvement is operational excellence. Techniques and tools are a practical step to begin but are not enough to attain long-lasting change. There is often no difference of opinion between artificial intelligence and automation in normal discussions. However, both concepts are quite different from each other. Organizations who use them complementary can provide significant benefits for business productivity. With improved efficiency, organizations have room to reconfigure resources and progress. Like ERP, automation software will also help extract data for extensive comprehension, recognize the latest income flow to investigate, and assist organizations in accepting innovative technologies (Kibria et al., 2018). Achieving operational excellence enhances the efforts between employees and tools. Artificial intelligence performs tasks previously done by humans to improve their performance and productivity. Before adopting artificial intelligence and automation in business procedures, organizations need to understand their integration with advanced technology (Found et al., 2018). Skill improvement is the fundamental requirement of an advanced generation. Technology-based employees accept the integration of technological knowledge and strategic planning. In a competitive organizational

environment, there is a need for daily improvement. It defines the basic concept of organizational excellence. Organizations are adopting artificial intelligence to develop automatic responses to information technology's crucial functions. Automation procedure involves modifying the methods that both information technology specialists and business leaders apply in their daily activities (Hertz et al., 2018). By eliminating human work in fields where there is no necessity for human interaction, artificial intelligence can develop a foundation that improves processes. With the latest evaluation in artificial intelligence, organization of any size can automate their operational activities to develop fast machine-learning techniques (Kumar et al., 2018). It implies that advanced technology can detect and resolve issues as soon as they occur. These are the self-healing protocols of artificial intelligence that lead to improved information technology repair response time, reduced costs, networks, and business infrastructure with more reliability (Rusev and Salonitis, 2016). Many organizations are replacing the manual workforce with artificial intelligence and automation, permitting information technology specialists to spare more time on innovation instead of concentrating on daily problem solving from calculating and transferring large amounts of information while lessening actual-time RAM errors. CPU adjustments and controlling artificial intelligence probabilities are substantially limitless (Laird et al., 2017). Some latest algorithms can forecast problems before they occur, implementing a proactive strategy to discover suitable solutions. Solution based on artificial intelligence is modifying the business world continuously. The main advantages of self-hearing artificial intelligence solutions are the potential to lower the problem-solving time, discover the problems before they occur proactively, the critical ability for cost-reduction, and the potential to enhance customer's experience. IBM is developing the latest technologies to help the organizations align their procedure, lower operating costs, and develop innovative business ideas (Lee et al., 2018; Thürer et al., 2018).

Driving Forces for Using Artificial Intelligence for Achieving Organizational Excellence

Following are the driving forces for using artificial intelligence for achieving organizational performance:

Improvement in Computing Abilities of Machine

The power and speed of computers are improving with each passing day. Organizations can perform their daily activities using computers embedded with artificial intelligence for improving their performance (Chen et al., 2018).

Development of Data-Based Artificial Intelligence

The development from rule-based artificial intelligence to data-based artificial intelligence enables machine learning. Machines can learn and adopt procedures independently without if/then algorithms. It acts as a driving force for achieving operational excellence as machines continue the procedures with minimal human involvement (Li et al., 2017).

Advancements in Deep Learning

Deep learning allows machines to recognize and perceive the world in new manners. Traditionally, machines could not differentiate between similar products. Advanced technologies can extract the data and start performing functions independently. The term intelligent machines depicts this concept. From the organizational perspective, these technologies help the firms smartly detect the requirements and fulfill them (John et al., 2020).

Cloud Computing

Cloud computing also acts as a driving force that enables machines to communicate and complete specific tasks. Organizations must collect and process data in the present era, and cloud computing makes it easier and transparent for the procedures (Bolodurina and Parfenov, 2017).

Managing the Data

Artificial intelligence provides the opportunity to analyze, process, and act according to the data with exceptional speed levels. Data management is the basic requirement for embedding artificial intelligence in business procedures. Artificial intelligence allows us to analyze, manage, and process data according to requirements. Artificial intelligence permits enhanced storage capacities to manage and save data and information (Harrison et al., 2019).

Barriers in the Adoption of Artificial Intelligence for Achieving Organizational Excellence

Following are the barriers in the adoption of artificial intelligence for achieving organizational excellence:

Cultural Constraints

This constraint relates to intransigence to change. Humans follow habit patterns. Once humans discover a methodology of performing tasks that is effective or efficient, they like to stick to it. It mainly requires some confidence in organizations to say that the disturbances and costs involved in modifying procedures or adopting new procedures will be worthy of bringing profitability (Tarafdar et al., 2019). Resistance can be simple as reluctance to manage over control, whether that is straight away to machines or employees who manage the technological framework that makes artificial intelligence possible (Makridakis, 2017). Mostly, this observes the requirement for artificial intelligence and insufficient recognition of its benefits (Yigitcanlar et al., 2020). Education can help to overcome this barrier. People need to understand how advanced technologies from natural language processing to cloud computing can enhance productivity and lessen costs. Once people become aware of it, they repeatedly engage themselves to enhance the potential for productive change with artificial intelligence's adoption (Fountaine et al., 2019).

Fear

Fear is a natural and comprehensible response of humans. Fear about the unknown is the ancient and powerful

emotion of humans. Still, there are enormous things that humans cannot imagine about how artificial intelligence will change our future. Fear can be subsided by understanding how new jobs will be created. Decision-making procedure by computer algorithms is indeed difficult to understand. It creates a fear that humans are losing control over their tasks and are no longer experts in their work (Mata et al., 2018). If machines carry out tasks more effectively and efficiently, there could be a decrease in the demand for jobs in certain areas. It can lead to two different situations: Machines fulfill primary requirements, and humans can pursue innovative and leisure activities. Second, the dependency on artificial intelligence can cause unemployment and social disturbance. To avoid the barriers, the solution is to turn over to technological grounds to enhance human's work instead of replacing them (Ploder, 2019).

Lack of Skills

It is an actual and crucial issue for many organizations requiring the adoption of artificial intelligence and shifting to other data-based frameworks for automatic modification. When it comes to organizational performance and growth because of artificial intelligence, a barrier exists due to the lack of skills and technology specialists with the training and experience required to adopt the essential organizational change and infrastructure. Even though artificial intelligence has been applied for several years, it is only recently that this talent is in demand by the industrial sector (Lee et al., 2018). The massive progress in demand means that those with capabilities can demand higher salaries and promotional positions in the firms that appoint them. Furthermore, firms now understand the need for artificial intelligence and invest in implementing this technological (Moulin-Frier et al., 2017). Google and Facebook, for instance, are considered greatly advantages as businesses who have talent, considering other firms face fierce competition to hire new talent. However, there is a possibility that this barrier will be overcome by society closing the gap between demand and supply. With the need for skills, there is an opportunity for the talent to grow. Another resolution is enhancing skills among employees. With an increase in the number of artificial intelligence solutions accessible as a service progresses, there will be a lower need for thoroughly trained employees in conventional science to implement artificial intelligence solutions to many business issues. Thus, it will be more accessible and less difficult for growing internal employees in AI areas (Jarrahi, 2018).

Lack of Strategic Planning to Artificial Intelligence Adoption

Somehow or another, this is a blend of a few different obstructions – shortage of skills, cultural barriers, difficulty in management to understand the benefits and productivity of artificial intelligence, and technological transformation. According to the overall progress and development programs, the outcome is that artificial intelligence capabilities are not according to plans at a strategic level. The outcomes cannot address organizational benefits, overall progress, and development programs. The main reason is that when organizations gain awareness about the significance of adoption of artificial intelligence technology and the benefits it can provide, they are unable to consider it from a strategic point of view, or from a complete know-how about the goals and objectives of all characteristics of the operations of artificial intelligence, from data accumulation to uncovering the insights (Nguyen et al., 2019). The solution to this barrier is that firms must always assure a clear strategic plan before money and time spent on resource-intensive artificial intelligence ideas. They must have a clear understanding of the advantages they can bring forward. Companies should ensure their artificial intelligence is wholly associated with the business objectives and excellence, where every shareholder has complete knowledge about failure and success (Olsen and Tomlin, 2020; Raj et al., 2020).

METHODOLOGY

The present paper conducts an in-depth systematic literature review on the intersection of AI and operational excellence. This method helps to provide a theoretical background for the study. The literature review should be a "systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers." Conducting a systematic literature review is known as "Fundamental Scientific Activity." It also permits us to understand the research scope, present and learn about the previous literature, and discover the specific topic. We focused on English-language articles by following standard research procedures. The first step was a manual search of different relevant articles from EBSCO, ProQuest, Emerald Insight, Science direct, Taylor and Francis, Wiley, JSTOR, and IEEE. The methodology was designed in the following stages (**Figure 3**).

The articles were searched according to the search phrases and extracted from different databases. Initially, 1,854 articles were identified based on the generic keywords of artificial intelligence and operations excellence from 2015 to 2020. All results were stored and combined from all databases. Duplicate articles were omitted, leading to a net number of 850 articles compared according to the first English criteria. Articles were then documented in a "results" spreadsheet. An additional search was performed to ensure that all articles were found in the database to complete the process. Articles related to artificial intelligence and operational excellence were used in combination strings. Articles related to artificial intelligence and operations management were searched in the first set. The second set included articles about artificial intelligence, operations, operations research, operations excellence, and operations management. The strings were modified based on the different database types. The articles were based on quality and were filtered to only those in peer-reviewed journals and conferences, which led to a reduction of 350 articles. All articles were re-reviewed to ensure that they match the area of artificial

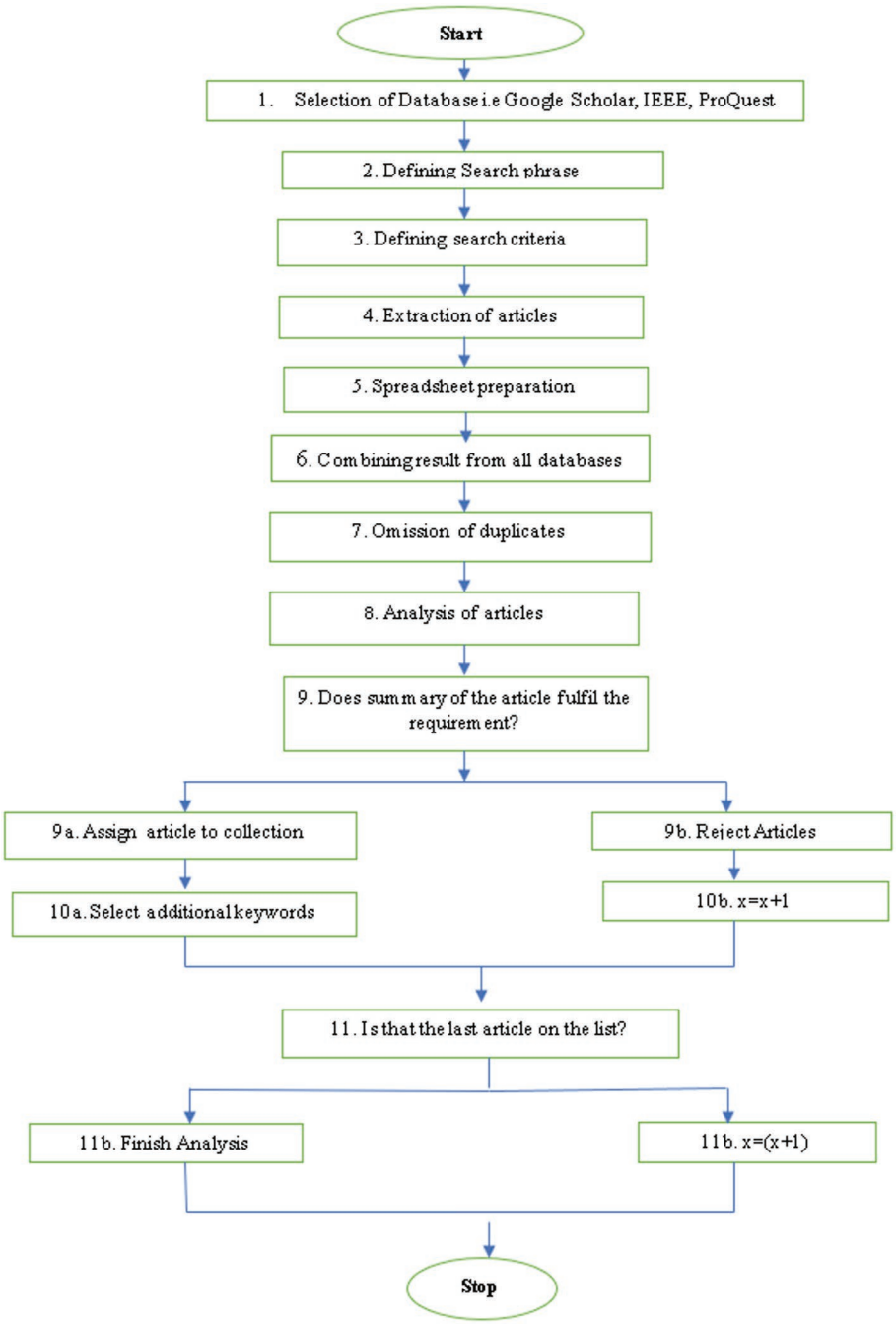

FIGURE 3 | Article selection methodology.

intelligence and operations excellence. Abstracts of different articles were reviewed to match the study area further. That final filtering shortlisted the articles to 53, which were compared and discussed for in-depth analysis. The results are given in **Table 1**, which describes the stats on articles during the initial database search and after article processing. The first number in each column represents the initial search, and the second number represents the quantity after filtering. It is shown in Initial Search and After Processing format.

FINDINGS AND RESULTS

The previous studies analysis proved that the artificial intelligence and operations excellence terms combined with other keywords identified a relatable need for coordination and commitment in the operation excellence field. The study findings are presented with tabulations and statistics to summarize the reviewed literature and discuss the research questions. **Figure 4** provides the article's numbers over the selected period sourced from

TABLE 1 | Database search and Processing results.

	EBSCO	ProQuest	Emerald Insight	Science Direct	Taylor & Francis	Wiley	JSTOR	IEEE	Total
	Artificial Intelligence								
Operations	51.2	48.0	62.1	35.1	23.0	43.0	21.0	18.0	301.4
Operations excellence	22.3	29.1	33.1	23.0	18.1	28.1	13.1	11.2	177.10
Operations management	31.0	32.0	28.1	19.1	22.0	22.0	14.0	5.2	173.4
Operations risk	19.0	9.0	19.0	7.1	6.1	12.0	9.0	6.0	87.2
Operations research	11.0	9.1	23.0	6.1	7.1	11.0	19.0	8.0	94.3
Operations model	32.1	19.1	22.0	9.0	6.0	15.1	11.0	3.0	107.3
Operations framework	15.1	5.1	14.1	11.1	5.0	12.1	9.1	7.0	78.6
Organization and operations	45.0	9.1	18.0	3.0	4.0	12.0	13.1	4.0	108.2
Operations and structure	23.0	9.0	18.0	7.0	3.0	12.0	7.1	3.0	82.1
Operations improvement	22.1	15.0	11.1	12.1	11.1	16.0	9.0	6.0	102.4
Operations failure	18.0	9.1	16.0	7.0	5.1	25.1	12.0	8.0	100.3
Operations standards	12.0	8.1	9.1	9.0	3.1	9.0	14.0	3.0	67.3
Operations strategy	21.0	18.0	18.1	8.0	2.0	13.0	21.0	3.0	104.1
Operations planning	22.1	21.0	21.1	11.1	2.1	17.0	11.0	6.0	111.4
Operation decisions	15.0	17.1	16.0	4.0	2.0	21.1	19.0	9.0	103.2
Operation teams	12.0	8.1	7.0	3.0	1.0	9.0	11.0	3.0	54.1
Total	371.8	255.10	335.8	174.7	120.7	277.5	213.4	105.4	1850.53

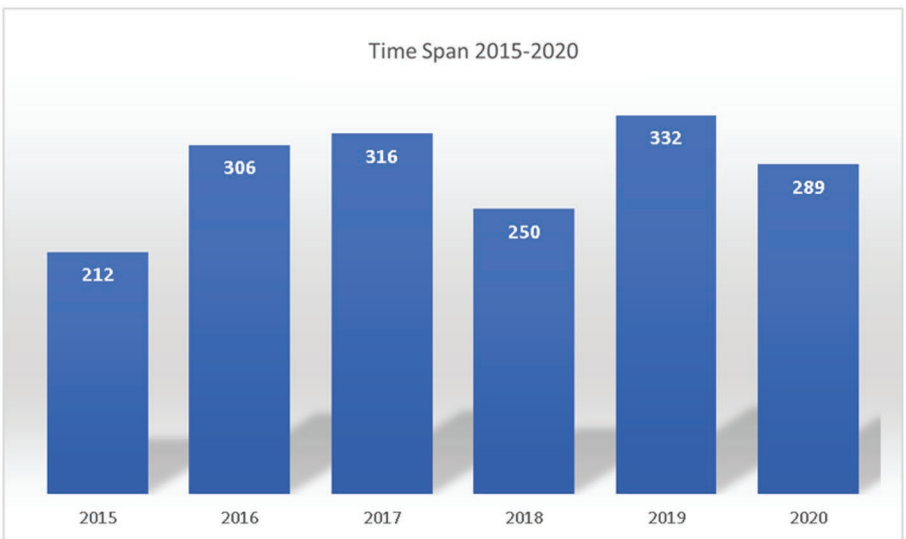

FIGURE 4 | Time span for identified articles.

peer-reviewed conferences and journals. **Figure 5** provides an overview of the distribution of articles from the database search. 75% of the literature review was from peer-reviewed journals and 25% from conferences. **Table 2** summarizes the artificial intelligence methods relevant to the operational excellence finalized for the 53 articles.

Table 3 provides the frequency of AI methods used based on reviewed literature. Various AI methods were used in operational excellence based on the implementation. Most articles were based on using neural networks with a highest-ranking method with 21, decision support system as the second highest method of 20, decision trees and modeling as 16 each, simulation as 14, and automation as 12. The rest of the methods varies among different approaches. **Figure 6** provides the overall summary of the methodologies used by the articles. 68% of the articles have used single methods, 21% have used double methods, and 11% have used multiple methods to solve operations problems for achieving excellence. **Table 4** provides the frequency of the analyzed articles in terms of the unique outcome that shows the importance of using artificial intelligence for operational excellence. **Figure 7** shows the summary of the articles reviewed in terms of a unique outcome. Most of the articles have used the framework as a unique outcome, nearly at 17%, the approach is being used as 15%, the model is

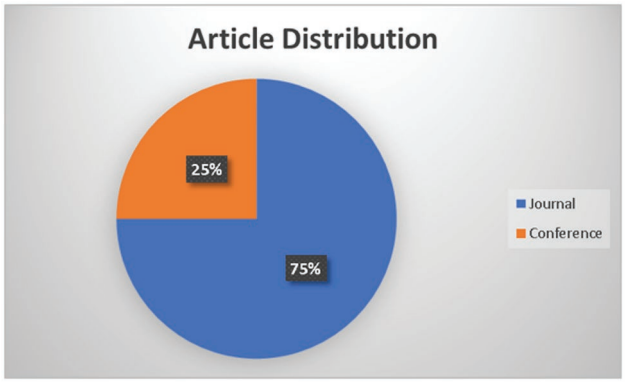

FIGURE 5 | Article distribution.

TABLE 2 | AI method usage with operational excellence.

Areas	Artificial intelligence methods/ Frequency
Operations	Automation (5) Neural networks (3) Decision trees (4) Modeling (6) Clustering (3) Support vector (2) Natural language processing (3) Fuzzy logic (4)
Operations management	Automation (7) Expert systems (1) Neural networks (3) Decision trees (2) Modeling (3) Simulations (5) Natural language processing (1)
Operations excellence	Decision support system (6) Modeling (5) Simulations (4) Automated planning (2) Neural networks (3) Deep learning (2)
Operations frameworks	Expert systems (2) Neural networks (5) Decision trees (6) Decision support system (7)
Operations models	Agent-based system (4) Expert systems (3) Modeling (2) Neural networks (4) Decision support system (3)
Operations decisions	Simulations (5) Decision trees (4) Neural networks (3) Image processing (2) Fuzzy logic (2) Decision support system (4) Natural language processing (4) Deep learning (4)

being used as 13%, case study, and method as 9% each. The rest of the outcomes varies among the selection of outcomes by the researchers.

TABLE 3 | Frequency of AI methods.

Method	Frequency
Neural networks	21
Decision support system	20
Decision trees	16
Modeling	16
Simulation	14
Automation	12
Natural language processing	8
Deep learning	6
Expert systems	6
Fuzzy logic	6
Agent-based system	4
Clustering	3
Support vector	2
Automated planning	2
Image processing	2

DISCUSSION

The results exhibit a connection between artificial intelligence and operational excellence that is proven by the selection of the 53 articles. It proves that 1 out of 12 articles about artificial intelligence refers to operational excellence. It also highlights that 12 percent of the 53 articles show a distinct connection between operational excellence and artificial intelligence. The results are significant because they depict that achieving operational excellence is dependent on the driving forces and resolving barriers. It is possible to achieve operational excellence by eliminating barriers. The analysis does not prove any specific methods to enhance operational excellence. It investigates the driving forces and barriers of using artificial intelligence to achieve operational excellence. The methodology allowed us to recognize how various keywords have a relationship and explain the research gaps.

What Is the Connection Between Operational Excellence and Artificial Intelligence?

The answer to the first research question is we conducted the analysis that explains artificial intelligence and virtual reality are the branches of computer studies and are significant tools for improving human life or sustaining their lifetime learning procedures. Technology plays a vital role in influencing the social, political, cultural, educational, and organizational sectors (Siryani et al., 2017).

Intelligent agents can help models other than C2B when negotiating between buyers and sellers. With the rapid advancements in artificial intelligence, there is a development in many concepts and priorities to enhance product selling management. In this technological era, online survey systems help recognize employee's needs regarding working in the organization (Wang et al., 2015). Operational excellence is a concept that focuses on problem-solving techniques and leadership skills as the main factor for continuous development. Firms are usually unsure how to proceed with operational excellence, and most organizations find it too broad or doubtful

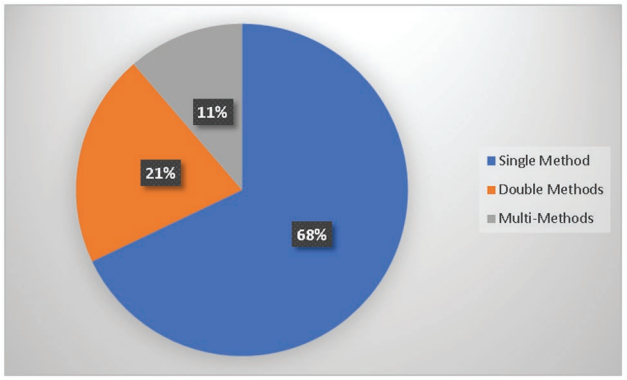

FIGURE 6 | Summary of methods.

TABLE 4 | Frequency of unique outcomes.

Unique Outcome	Frequency
Framework	9
Approach	8
Model	7
Case study	5
Method	5
Simulation	5
Literature review	4
Concept	3
Application	3
Exploratory	2
Tool	2

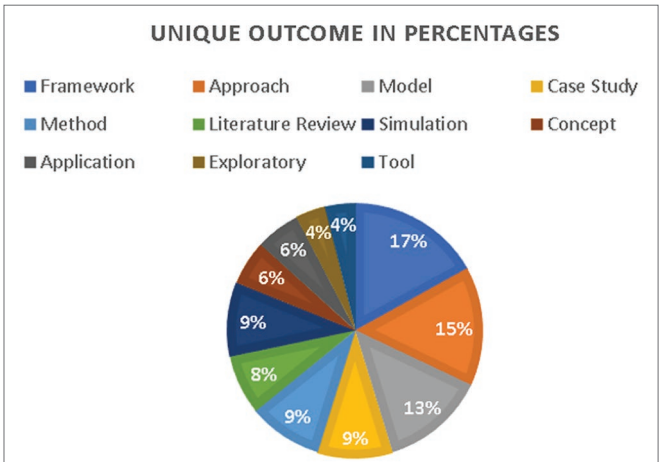

FIGURE 7 | Unique outcomes in percentages.

as it is a complicated concept to explain. The employee's and manager's attitude are not a set of activities that organizations perform. Firms can enhance their performance and culture through operational excellence, helping in sustainable progress. Organizations should observe past traditional events and look forward to a sustainable change system. Principles of operational excellence are: Respect every person, lead with modesty, seek excellence, accept scientific ideas, focus on the procedure, ensure quality, pull and flow value, think logically, develop purpose constancy, and value creation for customers.

What Are the Driving Forces for Achieving Organizational Performance by Using Artificial Intelligence?

The second research question addresses the question that the driving forces for achieving operational excellence through artificial intelligence are improved machine computing abilities (Wirtz, 2019). Organizations can perform their daily activities using computers embedded with artificial intelligence to improve their performance. The second driver is the development of data-based artificial intelligence: Machines can learn and adopt procedures independently without the algorithm of if/then. The third driver is advancements in deep learning, which means that deep learning allows machines to recognize and perceive the world differently (Zhao et al., 2019). From the organizational perspective, these technologies help the firms smartly detect the requirements and fulfill them. The fourth barrier is cloud computing, which enables machines to communicate and work to complete specific tasks. In the present era, organizations have to collect and process data, and cloud computing makes it easier and transparent for the procedures. The fifth driver is managing the data. Artificial intelligence provides the opportunity to analyze, process, and act according to the data with exceptional speed levels. Artificial intelligence allows us to analyze, manage, and process data according to requirements.

What Are the Barriers to Achieving Organizational Performance Using Artificial Intelligence?

The third research answer is the barriers in achieving organizational excellence through artificial intelligence that are cultural constraints; Once humans discover a methodology of performing tasks that is effective or efficient, they like to stick to it. It can be simple as reluctance to manage over control, whether that is straight away to machines or employees who manage the technological framework that makes artificial intelligence possible. Education can help to overcome this barrier. People need to understand how advanced technologies from natural language processing to cloud computing can enhance productivity and lessen costs. Once people become aware of it, they repeatedly engage themselves to enhance the potential for effective change with artificial intelligence's adoption. The second barrier is fear. Fear is a natural and comprehensible response of humans. Fear about the unknown is human's ancient and powerful emotion (Bottani et al., 2019).

For a quick understanding, the fear can circulate distance between the job and employees to get the salary. It creates a fear that humans are losing control over their tasks and are no longer experts in their work. To avoid this barrier, the solution is to turn over to technological grounds to enhance human work instead of replacing them. The third barrier is the lack of skills. It is an actual and crucial issue for many

organizations requiring the adoption of artificial intelligence and shifting to other data-based frameworks for automatic modification (Huo et al., 2020). When it comes to organizational performance and growth because of artificial intelligence, a barrier exists due to the lack of skills and technology specialists with the training and experience required to adopt the fundamental organizational change and infrastructure (Gray-Hawkins and Lăzăroiu, 2020). Although there is a possibility that this barrier will be overcome by managing the demand and supply (Marcos et al., 2020), with the need for skills, there is an opportunity for the talent to grow. Another resolution is enhancing skills among employees. The fourth barrier is the lack of strategic planning for artificial intelligence adoption. Somehow or another, this is a blend of a few different obstructions – shortage of skills, cultural barriers, difficulty in management that affect the benefits and productivity of artificial intelligence, and technological transformation. The solution to this barrier is that pretty direct firms must always assure a clear strategic plan before money and time spent on resource-intensive artificial intelligence ideas without a clear understanding of the advantages they can bring forward.

RESEARCH IMPLICATION

This study will help managers understand the barriers to adopting artificial intelligence to achieve operational excellence. This paper upgrades the previous understanding of operational excellence, artificial intelligence, and the drivers and barriers in adopting artificial intelligence. Knowing the drivers and barriers in adopting artificial intelligence for achieving operational excellence supports the organizations to maintain their competitive position. The focused literature review revealed that operational excellence could be implemented in a more efficient manner using artificial intelligence. The proposed framework for the operational excellence core functionalities can be extended according to organizational needs. Additionally, the in-depth framework for the artificial intelligence-based operational excellence framework provided the value-added areas that can help organizational leaders to focus on the essential domains as well-linked functionalities. Implementing the proposed frameworks will expedite the operational excellence in organizations with a clear roadmap to align the key performance indicators.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

There are some limitations to this study. First, the framework of this study focuses on prior literature. So, it is not practical to apply generalization. Furthermore, the study is restricted to operational excellence and artificial intelligence uses in the organization. Future researchers can investigate other factors that impact artificial intelligence's adoption in other sectors, such as banking, construction, and health sectors. Additionally, the proposed frameworks can be used for implementation in an organization for a pilot run. The outcome of the results can be further verified through the pre- and post-implementation of the artificial intelligence-based operational excellence framework. Moreover, a case study-based approach can be undertaken to provide complete procedures and processes for other organizations and better understand artificial intelligence and operational excellence.

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

MT contributed to the ideation, conceptualization systematic analysis, review, collection of articles, visualization, and formatting of the articles. MP and AA contributed to the literature review, tabular result analysis, proofreading, logical flow, visualization enhancement, and final formatting. All authors contributed to the article and approved the submitted version.

FUNDING

Abu Dhabi School of Management will fund the open access publication fees for the current research.

REFERENCES

Abduljabbar, R., Dia, H., Liyanage, S., and Bagloee, S. A. (2019). Applications of artificial intelligence in transport: an overview. *Sustain. For.* 11, 421–426. doi: 10.1016/j.radonc.2018.05.030

Aguirre, S., and Rodriguez, A. (2017). "Automation of a business process using robotic process automation (RPA): a case study" in *Workshop on Engineering Applications* (Cham: Springer), 65–71.

Alefari, M., Almanei, M., and Salonitis, K. (2020). Lean manufacturing, leadership and employees: the case of UAE SME manufacturing companies. *Prod. Manuf. Res.* 8, 222–243. doi: 10.1080/21693277.2020.1781704

Allam, Z., and Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities* 89, 80–91. doi: 10.1016/j.cities.2019.01.032

Alvi, A. K., Jawaid, A., Kaur, P., Safdar, U., and Bakht Yawar, R. (2020). Relationship between organizational benefits and employee job engagement. *Eur. Online J. Nat. Social Sci.* 9:339. doi: 10.1080/0142159X.2017.1359522

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI. *Info. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Becker, B. (2017). Artificial intelligence in education: what is it, where is it now, where is it going. *Irel. Yearb. Ed.* 2018, 42–46.

Benzaid, C., and Taleb, T. (2020). AI-driven zero touch network and service management in 5G and beyond: challenges and research directions. *IEEE Netw.* 34, 186–194. doi: 10.1109/MNET.001.1900252

Bolodurina, I., and Parfenov, D. (2017). Development and research of models of organization distributed cloud computing based on the software-defined infrastructure. *Proc. Comp. Sci.* 103, 569–576. doi: 10.1016/j.procs.2017.01.064

Bottani, E., Centobelli, P., Gallo, M., Mohamad, A. K., Jain, V., and Murino, T. (2019). Modelling wholesale distribution operations: an artificial intelligence framework. *Ind. Manag. Data Syst.* 119, 698–718. doi: 10.1108/IMDS-04-2018-0164

Çaliş, B., and Bulkan, S. (2015). A research survey: review of AI solution strategies of job shop scheduling problem. *J. Intell. Manuf.* 26, 961–973. doi: 10.1007/s10845-013-0837-8

Carvalho, A. M., Sampaio, P., Rebentisch, E., Carvalho, J. Á., and Saraiva, P. (2019). Operational excellence, organisational culture and agility: the missing link? *Total Qual. Manag. Bus. Excell.* 30, 1495–1514. doi: 10.1080/14783363.2017.1374833

Chakraborty, S., Sharma, A., and Vaidya, O. S. (2020). Achieving sustainable operational excellence through IT implementation in Indian logistics sector: an analysis of barriers. *Resour. Conserv. Recycl.* 152:104506. doi: 10.1016/j.resconrec.2019.104506

Chen, M., Herrera, F., and Hwang, K. (2018). Cognitive computing: architecture, technologies and intelligent applications. *IEEE Access* 6, 19774–19783. doi: 10.1109/ACCESS.2018.2791469

Chiarini, A., and Kumar, M. (2020). Lean six sigma and industry 4.0 integration for operational excellence: evidence from Italian manufacturing companies. *Prod. Plan. Control.* 1, 1–18. doi: 10.1080/09537287.2020.1784485

Choi, T. M., Wallace, S. W., and Wang, Y. (2018). Big data analytics in operations management. *Prod. Oper. Manag.* 27, 1868–1883. doi: 10.1111/poms.12838

Danaher, J. (2018). Toward an ethics of AI assistants: an initial framework. *Philos. Tech.* 31, 629–653. doi: 10.1007/s13347-018-0317-3

Davenport, T. H., and Ronanki, R. (2018). Artificial intelligence for the real world. *Harv. Bus. Rev.* 96, 108–116.

Deivanathan, R. (2019). "A review of artificial intelligence technologies to achieve machining objectives" in *Cognitive Social Mining Applications in Data Analytics and Forensics* (United States: IGI Global), 138–159.

Dogru, A. K., and Keskin, B. B. (2020). AI in operations management: applications, challenges and opportunities. *J. Data Info. Manage.* 2, 1–8. doi: 10.1007/s42488-020-00023-1

Eigenraam, A. W., Eelen, J., Van Lin, A., and Verlegh, P. W. (2018). A consumer-based taxonomy of digital customer engagement practices. *J. Interact. Mark.* 44, 102–121. doi: 10.1016/j.intmar.2018.07.002

Found, P., Lahy, A., Williams, S., Hu, Q., and Mason, R. (2018). Towards a theory of operational excellence. *Total Qual. Manag. Bus. Excell.* 29, 1012–1024. doi: 10.1080/14783363.2018.1486544

Fountaine, T., McCarthy, B., and Saleh, T. (2019). Building the AI-powered organization. *Harv. Bus. Rev.* 97, 62–73.

Gólcher-Barguil, L. A., Nadeem, S. P., and Garza-Reyes, J. A. (2019). Measuring operational excellence: an operational excellence profitability (OEP) approach. *Prod. Plan. Control* 30, 682–698. doi: 10.1080/09537287.2019.1580784

Gray-Hawkins, M., and Lăzăroiu, G. (2020). Industrial artificial intelligence, sustainable product lifecycle management, and internet of things sensing networks in cyber-physical smart manufacturing systems. *J. Self-Gov. Manage. Eco.* 8, 19–28. doi: 10.22381/JSME8420202

Harrison, T. F., Luna-Reyes, L., Pardo, T., De Paula, N., Najafabadi, M., and Palmer, J. (2019). "The data firehose and AI in government: why data management is a key to value and ethics." in *Proceedings of the 20th Annual International Conference on Digital Government Research*. June 2019; New York, NY: Association for Computing Machinery, 171–176.

Heinonen, K., Campbell, C., and Ferguson, S. L. (2019). Strategies for creating value through individual and collective customer experiences. *Business Horiz.* 62, 95–104. doi: 10.1016/j.bushor.2018.09.002

Hertz, H. S., Barker, S., and Edgeman, R. (2018). Current and future states: reinventing enterprise excellence. *Total Qual. Manag. Bus. Excell.* 2, 1–10. doi: 10.1080/14783363.2018.1444475

Huo, C., Hameed, J., Haq, I. U., Noman, S. M., and Sohail, R. C. (2020). The impact of artificial and non- artificial intelligence on production and operation of new products -an emerging market analysis of technological advancements a managerial perspective. *Rev. Argent. De Clín. Psicoló.* 29:69. doi: 10.24205/03276716.2020.1008

Ivanov, D., and Sokolov, B. (2019). Simultaneous structural–operational control of supply chain dynamics and resilience. *Ann. Oper. Res.* 283, 1191–1210. doi: 10.1007/s10479-019-03231-0

Jamshidieini, B., Rezaie, K., Eskandari, N., and Dadashi, A. (2017). Operational excellence in optimal planning and utilisation of power distribution network. *CIRED-Open Access Proc. J.* 2017, 2449–2452. doi: 10.1049/oap-cired.2017.1115

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61, 577–586. doi: 10.1016/j.bushor.2018.03.007

John, M. M., Olsson, H. H., and Bosch, J. (2020). "Developing ML/DL models: a design framework." in *Proceedings of the International Conference on Software and System Processes*. 1–10.

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415

Kamble, S., Gunasekaran, A., and Dhone, N. C. (2020). Industry 4.0 and lean manufacturing practices for sustainable organisational performance in Indian manufacturing companies. *Int. J. Prod. Res.* 58, 1319–1337. doi: 10.1080/00207543.2019.1630772

Kang, J. H., Matusik, J. G., Kim, T. Y., and Phillips, J. M. (2016). Interactive effects of multiple organizational climates on employee innovative behavior in entrepreneurial firms: a cross-level investigation. *J. Bus. Ventur.* 31, 628–642. doi: 10.1016/j.jbusvent.2016.08.002

Karsenti, T. (2019). Artificial intelligence in education: the urgent need to prepare teachers for tomorrow's schools. *Formation et Profession* 27, 112–116. doi: 10.18162/fp.2019.a167

Kibria, M. G., Nguyen, K., Villardi, G. P., Zhao, O., Ishizu, K., and Kojima, F. (2018). Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *IEEE Access* 6, 32328–32338. doi: 10.1109/ACCESS.2018.2837692

Koopman, P., and Wagner, M. (2017). Autonomous vehicle safety: an interdisciplinary challenge. *IEEE Intell. Transp. Syst. Mag.* 9, 90–96. doi: 10.1109/MITS.2016.2583491

Kumar, S., Mookerjee, V., and Shubham, A. (2018). Research in operations management and information systems interface. *Prod. Oper. Manag.* 27, 1893–1905. doi: 10.1111/poms.12961

Laird, J. E., Lebiere, C., and Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Mag.* 38, 13–26. doi: 10.1609/aimag.v38i4.2744

Lee, J., Davari, H., Singh, J., and Pandhare, V. (2018). Industrial artificial intelligence for industry 4.0-based manufacturing systems. *Manuf. Lett.* 18, 20–23. doi: 10.1016/j.mfglet.2018.09.002

Li, B. H., Hou, B. C., Yu, W. T., Lu, X. B., and Yang, C. W. (2017). Applications of artificial intelligence in intelligent manufacturing: a review. *Front. Info. Tech. Electron. Eng.* 18, 86–96. doi: 10.1631/FITEE.1601885

Liang, C. C., Liang, W. Y., and Tseng, T. L. (2019). Evaluation of intelligent agents in consumer-to-business e-commerce. *Comp. Stand. Interfaces* 65, 122–131. doi: 10.1016/j.csi.2019.03.002

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. *Futures* 90, 46–60. doi: 10.1016/j.futures.2017.03.006

Mangla, S. K., Kusi-Sarpong, S., Luthra, S., Bai, C., Jakhar, S. K., and Khan, S. A. (2020). Operational excellence for improving sustainable supply chain performance. *Resour. Conserv. Recycl.* 162:105025. doi: 10.1016/j.resconrec.2020.105025

Marcos, G. G., Victor Hugo Carlquist, D. S., Rodrigues Pinto, L. F., Centoamore, P., Digiesi, S., Facchini, F., et al. (2020). Economic, environmental and social gains of the implementation of artificial intelligence at dam operations toward industry 4.0 principles. *Sustain. For.* 12:3604. doi: 10.3390/su12093604

Masum, A. K. M., Beh, L. S., Azad, M. A. K., and Hoque, K. (2018). Intelligent human resource information system (i-HRIS): a holistic decision support framework for HR excellence. *Int. Arab. J. Inf. Technol.* 15, 121–130.

Mata, J., de Miguel, I., Duran, R. J., Merayo, N., Singh, S. K., Jukan, A., et al. (2018). Artificial intelligence (AI) methods in optical networks: a comprehensive survey. *Opt. Switch. Netw.* 28, 43–57. doi: 10.1016/j.osn.2017.12.006

Moulin-Frier, C., Puigbo, J. Y., Arsiwalla, X. D., Sanchez-Fibla, M., and Verschure, P. F. (2017). "Embodied artificial intelligence through distributed adaptive control: an integrated framework" in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. New York: IEEE, 324–330.

Mühlroth, C., and Grottke, M. (2020). Artificial intelligence in innovation: how to spot emerging trends and technologies. *IEEE Trans. Eng. Manag.* 5, 1–18. doi: 10.1109/TEM.2020.2989214

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., et al. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artif. Intell. Rev.* 52, 77–124. doi: 10.1007/s10462-018-09679-z

Niñerola, A., Sánchez-Rebull, M. V., and Hernández-Lara, A. B. (2019). Six sigma literature: a bibliometric analysis. *Total Qual. Manag. Bus. Excell.* 8, 1–22. doi: 10.1080/14783363.2019.1652091

Olsen, T. L., and Tomlin, B. (2020). Industry 4.0: opportunities and challenges for operations management. *Manuf. Serv. Oper. Manag.* 22, 113–122. doi: 10.1287/msom.2019.0796

Pansari, A., and Kumar, V. (2017). Customer engagement: the construct, antecedents, and consequences. *J. Acad. Mark. Sci.* 45, 294–311. doi: 10.1007/s11747-016-0485-6

Paschek, D., Luminosu, C. T., and Draghici, A. (2017). "Automated business process management–in times of digital transformation using machine learning or artificial intelligence" in *MATEC Web of Conferences. Vol. 121*. France: EDP Sciences, 04007.

Ploder, C. (2019). "Artificial intelligence tool penetration in business: adoption, challenges and fears" in *International Conference on Knowledge Management in Organizations. Vol. 1027*. Cham: Springer, 259.

Postavaru, N., Draghici, G., Filip, C., Mohammed, A. R., and Mohammed, S. M. (2019). Business management strategies for business development. Organization of the territory and planning of construction works. *Ovidius Univ. Ann. Const. Ser. Civil Eng.* 21, 45–50. doi: 10.2478/ouacsce-2019-0005

Raj, A., Dwivedi, G., Sharma, A., de Sousa Jabbour, A. B. L., and Rajak, S. (2020). Barriers to the adoption of industry 4.0 technologies in the manufacturing sector: an inter-country comparative perspective. *Int. J. Prod. Econ.* 224:107546. doi: 10.1016/j.ijpe.2019.107546

Robert, L. P., Pierce, C., Marquis, L., Kim, S., and Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Comp. Int.* 2, 1–31. doi: 10.1080/07370024.2020.1735391

Rusev, S. J., and Salonitis, K. (2016). Operational excellence assessment framework for manufacturing companies. *Proc. CIRP* 55, 272–277. doi: 10.1016/j.procir.2016.08.026

Sánchez-Ruiz, L., Blanco, B., and Gómez-López, R. (2019). Continuous improvement enablers: defining a new construct. *J. Ind. Eng. Manag.* 12, 51–69. doi: 10.3926/jiem.2743

Scheer, A. W. (2017). "Theses on digitalization," in *The Drivers of Digital Transformation*. ed. F. Abolhassan (Cham: Springer), 33–43.

Sehnem, S., Jabbour, C. J. C., Pereira, S. C. F., and de Sousa Jabbour, A. B. L. (2019). Improving sustainable supply chains performance through operational excellence: circular economy approach. *Resour. Conserv. Recycl.* 149, 236–248. doi: 10.1016/j.resconrec.2019.05.021

Shan, A. W., Ahmad, M. F., and Nor, N. H. M. (2016). "The mediating effect of kaizen between total quality management (TQM) and business performance" in *IOP Conference Series: Materials Science and Engineering. Vol. 160*. United Kingdom: IOP Publishing, 012012.

Shehadeh, R., Al-Zu'bi, Z. M. F., Abdallah, A. B., and Maqableh, M. (2016). Investigating critical factors affecting the operational excellence of service firms in Jordan. *J. Manag. Res.* 8, 18–49. doi: 10.5296/jmr.v8i1.8680

Singh, J., Flaherty, K., Sohi, R. S., Deeter-Schmelz, D., Habel, J., Le Meunier-Fitz Hugh, K., et al. (2019). Sales profession and professionals in the age of digitization and artificial intelligence technologies: concepts, priorities, and questions. *J. Pers. Sell. Sales Manag.* 39, 2–22. doi: 10.1080/08853134.2018.1557525

Siryani, J., Tanju, B., and Eveleigh, T. J. (2017). A machine learning decision-support system improves the internet of things' smart meter operations. *IEEE Internet Things J.* 4, 1056–1066. doi: 10.1109/JIOT.2017.2722358

Skurdauskaitė, I. (2020). 45 ways to look at benefits and risks of artificial intelligence: what to expect? *Politologija* 97, 123–129. doi: 10.15388/Polit.2020.97.5

Sony, M. (2019). Implementing sustainable operational excellence in organizations: an integrative viewpoint. *Prod. Manuf. Res.* 7, 67–87. doi: 10.1080/21693277.2019.1581674

Stanica, I., Dascalu, M. I., Bodea, C. N., and Moldoveanu, A. D. B. (2018). "VR job interview simulator: where virtual reality meets artificial intelligence for education" in *Zooming Innovation in Consumer Technologies Conference (ZINC)*. New York: IEEE, 9–12.

Tarafdar, M., Beath, C. M., and Ross, J. W. (2019). Using AI to enhance business operations. *MIT Sloan Manag. Rev.* 60, 37–44.

Thürer, M., Tomašević, I., Stevenson, M., Fredendall, L. D., and Protzman, C. W. (2018). On the meaning and use of excellence in the operations literature: a systematic review. *Total Qual. Manag. Bus. Excell.* 2, 1–28. doi: 10.1080/14783363.2018.1434770

van Assen, M. F. (2020). Empowering leadership and contextual ambidexterity–The mediating role of committed leadership for continuous improvement. *Eur. Manag. J.* 38, 435–449. doi: 10.1016/j.emj.2019.12.002

Voronkova, O. V. (2019). The impact of artificial intelligence technologies on society. *Rep. Sci. Soc.* 1, 7–9.

Wamba, S. F., Gunasekaran, A., Dubey, R., and Ngai, E. W. (2018). Big data analytics in operations and supply chain management. *Ann. Oper. Res.* 270, 1–4. doi: 10.1007/s10479-018-3024-7

Wang, X., Li, X., and Leung, V. C. (2015). Artificial intelligence-based techniques for emerging heterogeneous network: state of the arts, opportunities, and challenges. *IEEE Access* 3, 1379–1391. doi: 10.1109/ACCESS.2015.2467174

Wirtz, J. (2019). Organizational ambidexterity: cost-effective service excellence, service robots, and artificial intelligence. *Organ. Dyn.* 49:100719. doi: 10.1016/j.orgdyn.2019.04.005

Yigitcanlar, T., Desouza, K. C., Butler, L., and Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: insights from a systematic review of the literature. *Energies* 13:1473. doi: 10.3390/en13061473

Zhao, Y., Li, T., Zhang, X., and Zhang, C. (2019). Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. *Renew. Sust. Energ. Rev.* 109, 85–101. doi: 10.1016/j.rser.2019.04.021

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 Psychology

ORIGINAL RESEARCH
published: 12 July 2021
doi: 10.3389/fpsyg.2021.696770

Connecting Social Psychology and Deep Reinforcement Learning: A Probabilistic Predictor on the Intention to Do Home-Based Physical Activity After Message Exposure

Patrizia Catellani[1]*, *Valentina Carfora*[1] and *Marco Piastra*[2]

[1] Department of Psychology, Catholic University of Milan, Milan, Italy, [2] Department of Industrial, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

Previous research has shown that sending personalized messages consistent with the recipient's psychological profile is essential to activate the change toward a healthy lifestyle. In this paper we present an example of how artificial intelligence can support psychology in this process, illustrating the development of a probabilistic predictor in the form of a Dynamic Bayesian Network (DBN). The predictor regards the change in the intention to do home-based physical activity after message exposure. The data used to construct the predictor are those of a study on the effects of framing in communication to promote physical activity at home during the Covid-19 lockdown. The theoretical reference is that of psychosocial research on the effects of framing, according to which similar communicative contents formulated in different ways can be differently effective depending on the characteristics of the recipient. Study participants completed a first questionnaire aimed at measuring the psychosocial dimensions involved in doing physical activity at home. Next, they read recommendation messages formulated with one of four different frames (gain, non-loss, non-gain, and loss). Finally, they completed a second questionnaire measuring their perception of the messages and again the intention to exercise at home. The collected data were analyzed to elicit a DBN, i.e., a probabilistic structure representing the interrelationships between all the dimensions considered in the study. The adopted procedure was aimed to achieve a good balance between explainability and predictivity. The elicited DBN was found to be consistent with the psychosocial theories assumed as reference and able to predict the effectiveness of the different messages starting from the relevant psychosocial dimensions of the recipients. In the next steps of our project, the DBN will form the basis for the training of a Deep Reinforcement Learning (DRL) system for the synthesis of automatic interaction strategies. In turn, the DRL system will train a Deep Neural Network (DNN) that will guide the online interaction process. The discussion focuses on the advantages of the proposed procedure in terms of interpretability and effectiveness.

Keywords: probabilistic predictor, Dynamic Bayesian Network, message framing, home-based physical activity, intention change

OPEN ACCESS

Edited by:
Stefano Triberti, University of Milan, Italy

Reviewed by:
Davide Mazzoni, University of Milan, Italy
Juneman Abraham, Binus University, Indonesia

***Correspondence:**
Patrizia Catellani
patrizia.catellani@unicatt.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *17 April 2021*
Accepted: *16 June 2021*
Published: *12 July 2021*

Citation:
Catellani P, Carfora V and Piastra M (2021) Connecting Social Psychology and Deep Reinforcement Learning: A Probabilistic Predictor on the Intention to Do Home-Based Physical Activity After Message Exposure. Front. Psychol. 12:696770. doi: 10.3389/fpsyg.2021.696770

INTRODUCTION

Doing physical activity is essential for people's health and well-being (Hyde et al., 2013; Rhodes et al., 2017). During the lockdown due to the COVID-19 pandemic, this role of physical activity has become even more crucial and an increase in physical activity at home has become essential to keep in exercise despite the constraints of external mobility (Taylor et al., 2020; University of Virginia Health System, 2020). Even when we are aware of the benefits associated with physical activity, this awareness does not necessarily translate into consistent behavior. This is because the psychological factors related to physical activity are many and their relationships are complex. Understanding these relationships is essential to develop personalized and effective intervention strategies, which can be addressed to as many people as possible and be economically sustainable.

Some previous research has investigated how to promote physical activity using automatic interaction systems, such as artificial intelligence chatbot or personalized physical activity coaching based on machine learning (Dijkhuis et al., 2018; Aldenaini et al., 2020; Zhang et al., 2020). However, a full understanding of the theoretical guidance and practices on designing automatic interaction systems to support the increase in people's physical activity is still lacking (Zhang et al., 2020). Such understanding should include the development of empirically testable theoretical models, which consider the psychosocial processes related to behavior planning and how communication can influence it.

In the present study, we developed an empirically testable model to facilitate the promotion of physical activity thanks to the application of artificial intelligence. To do so, we first collected data on a sample of participants exposed to different messages promoting home-based physical activity during the first lockdown due to the Covid-19 epidemic in 2020. Participants were involved in an experimental procedure articulated in three phases: (a) filling out a first questionnaire aimed at identifying the psychosocial dimensions involved in the intention to do home-based physical activity; (b) reading persuasive messages aimed at promoting home-based physical activity and framed in different ways depending on the experimental condition; (c) filling out a second questionnaire aimed at detecting the evaluation of the messages received and any change in the intention to exercise at home.

We then developed a probabilistic graphical structure, i.e., a Dynamic Bayesian Network (DBN; Dagum et al., 1995; Murphy, 2012), as a first step in a process aimed at harnessing psychological models in the construction of automated interaction strategies via artificial intelligence. In doing this, we aimed at striking a balance between the *explanatory power* of the DBN, namely, its capacity of describing the causal connections among the psychological dimensions included in the theoretical model, and the *predictive capability* of the DBN, namely, its effectiveness in anticipating the effect of a specific interaction strategy. In other words, we aimed at achieving a good equilibrium between *what* can be predicted and *why* it can be predicted. The goal of achieving such a balance is relevant both for quantitative psychology (Yarkoni and Westfall, 2017) and for artificial intelligence (Adadi and Berrada, 2018).

To summarize, the main aim of our paper was to develop a probabilistic predictor in the form of a DBN, capable to explain and predict change in the intention to do physical activity at home after being exposed to messages on the subject. Such DBN is intended as the first step of an articulate process that has the ultimate goal of developing effective and automatic interaction strategies regarding behavior change.

In the rest of the paper, we first present the procedure and the measures employed in the empirical study, specifying the psychosocial theories we referred to in carrying it out. We then illustrate the main characteristics of the DBN, as structured predictor, and describe the methods adopted for its elicitation from the data collected in the study. The criteria to balance explanatory power and predictive capability, and the deterministic structure search of the DBN are also discussed. Then, in the Results section we illustrate the structure and parameters of the elicited DBN and its consistency with the psychosocial theoretical models. We finally discuss the advantages, limits, and future developments of our procedure, which will include a Deep Reinforcement Learning component for training a Deep Neural Network expected to drive online interactions with people.

METHODS

Participants and Procedure

The present study was conducted following receipt of ethical approval by the Catholic University of the Sacred Heart (Milan). In April 2020, a sample of Italian participants was recruited to participate in a university study on the effects of public communication regarding the benefits of home-based physical activity. Participants were recruited by students of psychology courses at the Catholic University of Milan and received an email with a link to an online survey developed through the Qualtrics platform.

An initial sample of 280 participants accessed the online survey developed through the Qualtrics platform. First, participants completed a questionnaire measuring psychosocial dimensions involved in doing home-based physical activity (Time 1). Then, they were automatically and randomly assigned to four different experimental conditions, which consisted in being asked to read differently framed messages regarding the physical and psychological outcomes of exercising at home (Message Intervention). Finally, they were required to fill in a second questionnaire measuring their evaluation of the messages and again the psychosocial dimensions involved in home-based physical activity, to assess whether they had changed after message exposure (Time 2).

After excluding participants who either failed to pass the attention check questions in the questionnaires or did not complete them ($N = 8$), the final sample consisted of 272 participants (126 males, 142 females, 4 other; mean age = 42.97, $SD = 14.98$, age range = 18–70).

All data presented in this study can be found in the open repository at https://bitbucket.org/unipv_cvmlab/connecting_social_psychology_and_drl/.

Theory-Based Measures

The theoretical starting point of our study was the integration of psychosocial models aimed at explaining behavior planning, its change through persuasive communication, and the matching effect between persuasive messages and recipients' characteristics (see also Di Massimo et al., 2019; Carfora et al., 2020a).

Regarding behavior planning, our reference model was the widely known Theory of Planned Behavior (TPB; Ajzen, 1991), according to which the *intention* to enact a certain behavior is predicted by the *attitude* toward the behavior (e.g., perceiving exercising at home as a useless activity), the *social norm* (e.g., feeling that others would approve of their regular exercising at home), and *perceived behavioral control* (e.g., being convinced to have internal and external resources to exercise at home). Over time, various researches have highlighted that the predictive capacity of TPB is further increased by the addition of two further dimensions, namely, *past behavior* (e.g., having exercised regularly in the past month) and *anticipated positive or negative emotions* concerning the outcome (e.g., anticipating that one will feel satisfied (or guilty) if one will (or will not) exercise at home).

Regarding the effects of persuasive communication, we referred to the Elaboration Likelihood Model (ELM; Petty and Cacioppo, 1986), according to which the long-term persuasiveness of a message largely depends on the *evaluation* and *systematic processing* of the message itself. Subsequent developments of this model have led to highlighting additional factors that can increase or vice versa decrease the persuasive effect of a message. Among the first, the perception of *trust* that the message arouses (Petty, 2018) and the positive *tone* of the message (Latimer et al., 2008a). Among the second, the perception of *threat* or *distress* activated by the message (Shen, 2015) and the negative *tone* of it (Latimer et al., 2008a).

Finally, in devising persuasive messages we referred to the Self-Regulatory Model of Message Framing (Cesario et al., 2013), according to which similar contents can be framed in different ways, for example by stressing either the positive or the negative outcomes of the recommended action. In a gain message the outcome of the action is formulated with a positive valence, whereas in a loss message the outcome is formulated with a negative valence. Gain messages can be further differentiated in messages describing an actual *gain* (e.g., "If you do home-based physical activity, you will improve your health") and messages describing a *non-loss* (e.g., "If you do home-based physical activity, you will avoid damaging your health"). Similarly, loss messages can be further distinguished in messages describing an actual *loss* (e.g., "If you do not do home-based physical activity, you will damage your health") and messages describing a *non-gain* (e.g., "If you do not do home-based physical activity, you will miss the opportunity to improve your health").

Finally, previous research has shown that the persuasiveness of a message increases when its framing matches the recipient's regulatory focus (e.g., Yi and Baumgartner, 2009; Bertolotti et al., 2020). According to the Regulatory Focus Theory (RFT; Higgins, 1997), self-regulation with a *prevention focus* involves the avoidance of losses and the fulfillment of duties and obligations, while self-regulation with a *promotion focus* involves the pursuit of gains and the achievement of an ideal desirable state. Messages framed in terms of non-loss are more persuasive with people who have a prevalent focus of prevention, while messages framed in terms of gain are more persuasive with people who have a prevalent focus of promotion (Yi and Baumgartner, 2009). In this study we therefore introduced the regulatory focus measures at Time 1, to assess whether they would have an impact on intention change at Time 2, after exposure to differently framed messages.

Time 1 Measures

At the beginning of the survey, participants provided their informed consent and read the following statement: "We are interested in understanding what drives people to do physical activity at home in the absence of alternatives (i.e., in the impossibility of accessing parks, gyms, and open spaces). By physical activity at home we mean, for example: bodyweight workout (such as stretching, aerobics, push-ups, and abs), walking for at least 30 min (6,000 steps per day), training with weights and machines (such as stationary bikes and treadmills)." After that, participants answered to a series of questions measuring the relevant psychosocial dimensions investigated in the study.

Prevention focus was assessed using five items on a 7-point Likert scale adapted from the Health Regulatory Focus scale [e.g., "I often imagine myself being ill in the future… (1) Strongly disagree—(7) Strongly agree"; Ferrer et al., 2017]. The five items were used to compute a single prevention regulatory focus index, with higher values indicating a higher prevention focus. Cronbach's α was 0.87.

Promotion focus was assessed using five items on a 7-point Likert scale adapted from the Health Regulatory Focus scale [e.g., "I frequently imagine how I can achieve a state of "ideal health… Strongly disagree (1)—Strongly agree (7)"; Ferrer et al., 2017]. The five items were used to compute a single promotion regulatory focus index, with higher values indicating a higher promotion focus. Cronbach's α was 0.83.

Past behavior, related to physical activity *at home,* was assessed by asking how often participants engaged in exercising at home before the COVID-19 restrictions: "Before this period of restrictions, on average how many times a week did you exercise at home?... Never (1)—Every day (7)." Higher scores indicated a higher frequency of home-based physical activity before the COVID-19 restrictions.

Past outdoor behavior, related to *outdoor* physical activity, was assessed by asking how often participants engaged in exercising outside home before the COVID-19 restrictions: "Before this period of restrictions, on average how many times a week did you exercise outside home?... Never (1)—Every day (7)." Higher scores indicated a higher frequency of outdoor physical activity before the COVID-19 restrictions.

Attitude toward home-based physical activity was assessed using eight items on a semantic differential scale ranging from "1" to "7" (e.g., "I believe that doing physical exercises at home regularly is… useless—useful"; Caso et al., 2021). The eight items

were used to compute a single attitude index, with higher values indicating a more positive attitude toward exercising at home. Cronbach's α was 0.93.

Subjective norm was assessed with three items using a Likert scale [e.g., "Most of the people important to me (partners, family, friends) think I should do physical exercises at home regularly… Strongly disagree (1)—Strongly agree (7)"; adapted from Carfora et al., 2020a,b]. The three items were used to compute a single subjective norm index, with higher scores indicating a higher level of it. Cronbach's α was 0.83.

Perceived behavioral control related to home-based physical activity was measured using five items on a seven-point Likert scale [e.g., "If I wanted, I would be able to do the physical activity regularly when I am feeling tired… (1) Strongly disagree—(7) Strongly agree"; adapted from Bandura, 1977]. The five items were used to compute a single index, with higher values indicating higher perceived behavioral control regarding exercising at home. Cronbach's α was 0.90.

Anticipated positive emotions for doing home-based physical activity were assessed with three items using a Likert scale [e.g., "If I do physical exercises at home regularly I will be satisfied… Strongly disagree (1)—Strongly agree (7)"; adapted from Carfora et al., 2018]. The three items were used to compute a single anticipated positive emotions index, with higher scores indicating a higher level of them. Cronbach's α was 0.92.

Anticipated negative emotions for not doing home-based physical activity were assessed with three items using a Likert scale [e.g., "If I do not do physical exercises at home regularly I will regret it… Strongly disagree (1)—Strongly agree (7)"; adapted from Carfora et al., 2018]. The three items were used to compute a single anticipated negative emotions index, with higher scores indicated a higher level of them. Cronbach's α was 0.89.

Intention at Time 1 toward doing home-based physical activity was measured using three items on a seven-point Likert scale [e.g., "I intend to do physical exercises at home regularly in the next month… Strongly disagree (1)—Strongly agree (7)"; Clark and Bassett, 2014]. The three items were used to compute a single intention at Time 1 index. Higher scores indicated a greater intention to exercise at home at Time 1. Cronbach's α was 0.97.

A list of the above dimensions with examples of the items employed to measure them can be found in **Figure 1**.

Message Intervention

After completing the first questionnaire, participants read an infographic with six messages describing the physical, psychological, and social consequences of doing home-based physical activity (**Figure 2**). All messages were formulated in prefactual terms (i.e., "If … then"; see Carfora and Catellani, 2021) and approximately consisted of 14 words each. Messages were formulated differently, according to the experimental condition to which participants had been randomly assigned. Participants in the *gain message condition* read messages emphasizing the positive consequences of doing home-based physical activity (e.g., "If you do physical activity at home, you will improve your fitness"). Participants in the *non-loss message condition* read messages informing how to avoid negative outcomes by doing home-based physical activity (e.g., "If you do physical activity at home, you will avoid worsening your fitness"). Participants in the *non-gain message condition* read messages emphasizing how doing home-based physical activity is associated with missing out positive consequences (e.g., "If you do not do physical activity at home, you will lose the chance to improve your fitness"). Finally, participants in the *loss message condition* read messages on the negative consequences of not doing home-based physical activity (e.g., "If you do not do physical activity at home, you will worsen your fitness").

Time 2 Measures

After reading the messages, participants completed the second questionnaire, which measured the evaluation of the messages and once again the intention to exercise at home.

Message-induced threat was measured with four items on a 7-point Likert scale related to how much reading messages had made participants feel their freedom threatened [e.g., "The messages have tried to pressure me… (1) Strongly disagree – (7) Strongly agree"; adapted from Shen, 2015]. The four items were used to compute a single message-induced threat index, with higher values indicating higher perceived threat. Cronbach's α was 0.89.

Message-induced distress was assessed with five items on a 7-point Likert scale, pertaining to the degree to which reading messages induced distress [e.g., "How far this message scared you? … (1) Not at all – (7) Completely"; adapted from Brown and Smith, 2007]. All items were used to compute a single message-induced distress index, with higher values indicating higher distress after reading the messages. Cronbach's α was 0.86.

Message tone was measured with one item asking participants to rate the tone of the messages along the positivity-negativity dimension ["Overall, how would you rate the tone of the information presented in the messages? (1) Extremely negative – (7) Extremely positive"; adapted from Godinho et al., 2016]. Higher values indicated a more positive perception of the message tone.

Message trust was assessed with three items on a 7-point Likert scale [e.g., "Do you think the information presented in the message is reliable? (1) Not at all – (7) Extremely"; adapted from Godinho et al., 2016]. The three items were used to compute a single message trust index, with higher values indicating a higher trust in the messages. Cronbach's α was 0.92.

Systematic processing was measured with five items on a 7-point Likert scale, asking participants to state how deeply they had processed the information presented in the messages [e.g., "I tried to think about the importance of the information presented in the message for my daily life… (1) Strongly disagree – (7) Strongly agree"; adapted from Smerecnik et al., 2012]. The five items were used to compute a single systematic processing index, with higher values indicating a deeper processing of the messages. Cronbach's alpha was 0.91.

Message evaluation was assessed with six items on a 7-point Likert scale, regarding how participants evaluated the messages [e.g., "Messages were very interesting… (1) Strongly disagree – (7) Strongly agree"; adapted from Godinho et al., 2016]. The three

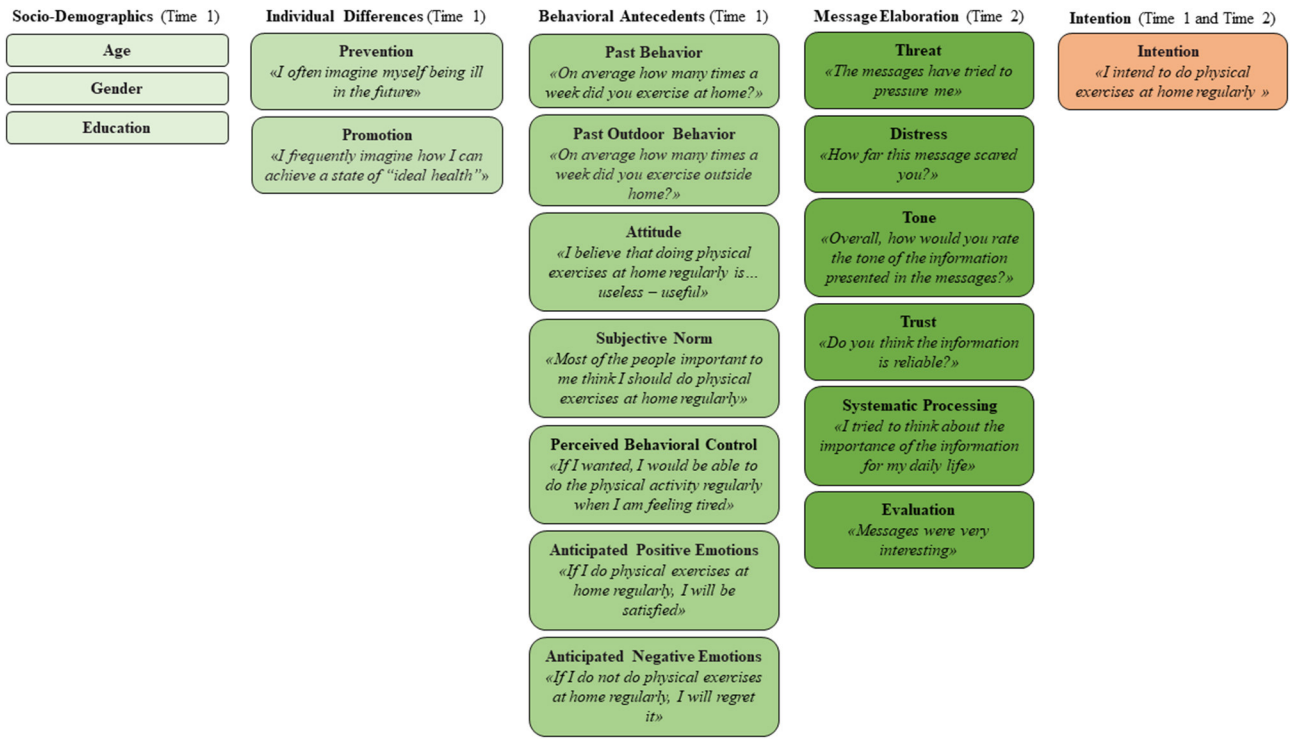

FIGURE 1 | Psychosocial predictors of change in the intention to exercise at home, with examples of the measures employed.

FIGURE 2 | Infographics proposed in the gain, non-loss, non-gain, and loss message conditions.

items were used to compute a single message evaluation index, with higher values indicating a more positive evaluation of the messages. Cronbach's α was 0.92.

Intention at Time 2 toward doing home-based physical activity was measured with the same three items employed at Time 1. Cronbach's α was 0.98.

Intention change was calculated subtracting the index *Intention at Time 1* from the index *Intention at Time 2*.

At the end of the questionnaire, participants reported their age, sex, and education.

A list of the above dimensions with examples of the items employed to measure them can be found in **Figure 1**.

Dynamic Bayesian Network

We now describe the theoretical framework adopted for defining the probabilistic predictor (sections Learning Structure and Parameters From Data and Explanatory Power vs. Predictive Capability) and then describe the method used for eliciting the predictor from collected data (section Deterministic Structure Search).

A Bayesian Network $\mathcal{B} = (V, A, p)$ (BN, Darwiche, 2009) is a directed acyclic graph where nodes V correspond to the random variables in the model, p is a joint probability distribution over the set of random variables, and each link $A \subseteq V \times V$ represents an oriented dependence relation among two random variables. Together, nodes and directed arcs represent the structure of p, in terms of independence and conditional independence conditions among random variables. More precisely, assuming that $\{X_1, \ldots, X_n\}$ is the set of all random variables in the model, the joint probability distribution p can be factorized as

$$p(X_1, \ldots, X_n) = \prod_i p(X_i \mid \pi(X_i))$$

where $\pi(X_i)$ is the set of *parents* of X_i, i.e., the set of random variables whose representing nodes have an arc directed toward the node representing X_i.

A *Dynamic* Bayesian Network (DBN; Dagum et al., 1995; Murphy, 2012) is a BN that also includes the representation of *time*, intended as a discrete sequence of instants. In a DBN:

- Each node is associated to a specific time instant.
- The same random variable may correspond to more than one node, at different times.
- All links must respect the orientation of time, either by connecting nodes at the same instant or by being oriented from a previous instant to a subsequent one.

As it can be seen in **Figure 3**, in our study the DBN was assumed to span across a sequence of three instants: Time 1, Message Intervention, and Time 2.

Being mean values of multi-item scales (**Table 1**), the indexes of the psychological dimensions calculated on the collected data can be assumed to be continuous. However, for computational simplicity, each corresponding random variable was assumed in this study to have values in the categorical scale {*low*, *medium*, *high*}, except for the target variable *Intention Change*, which was assumed to have values in the scale {*high-negative*, *low-negative*, *neutral*, *low-positive*, *high-positive*}. Indexes were discretized using *quantiles* (Nojavan et al., 2017): 20% quantiles for *Intention Change* and 33% quantiles for all the other variables.

Learning Structure and Parameters From Data

In general, once the structure of a DBN has been defined, the probability distribution p can be learned from experimental data, in a direct form. The learning process is an optimization aiming to compute the *maximum likelihood estimator* (MLE):

$$\theta_{MLE} := \underset{\theta}{\operatorname{argmax}}\, L(\theta, D)$$

where θ is the set of probability values, D are the collected data and L is the likelihood function. Omitting details, in the case of discrete Bayesian Networks the above optimization process could be solved analytically, by computing all required probabilities as frequency ratios in D (Murphy, 2012). However, such direct method is rarely used since it is vulnerable to missing data, a circumstance that occurs very often with limited datasets. In practice, other methods such as the EM algorithm (Dempster et al., 1977) are preferred since they are more robust and can deal with missing data.

A more complicate task, which has been subject to intense research, is eliciting from data the structure of the Bayesian Network (i.e., the acyclic graph) that best synthesizes the information collected in the experiments. In many commonly adopted approaches, a scoring function is used to evaluate candidate structures (Koller and Friedman, 2009). An obvious choice for this would be the likelihood function itself. One problem in doing so, however, is that the likelihood function is monotonically increasing with the number of nodes and arcs in the network. In other words, a Bayesian Network including one node per each measured variable and being a fully connected (acyclic) graph is due to attain the maximal likelihood in all cases. To counter this tendency, the *Bayesian Information Criterion* (BIC) includes another term that measures the complexity of the network:

$$BIC(\mathcal{B}, D) := l(\theta, D) - \frac{\log N}{2} |\mathcal{B}|$$

where $\mathcal{B}$ is the Bayesian Network, $l(\theta, D) := \log L(\theta, D)$ is the log-likelihood, N is the size of the dataset and $|\mathcal{B}|$ measures the number of nodes and arcs in the graph. The second term above is also called *description length*. In our work, however, we preferred a still different way to counter the tendency to structure growth induced by functions as the likelihood, as it will be explained in section Deterministic Structure Search.

Once a scoring function has been chosen, the subsequent step is defining a procedure for finding the graph structure of $\mathcal{B}$ that maximizes the given score. Unfortunately, this problem is NP-hard (Koller and Friedman, 2009) in general and therefore impervious to exhaustive search in almost all practical cases. Several heuristic search strategies have been proposed in the literature to circumvent this problem (e.g., see Cheng et al.,

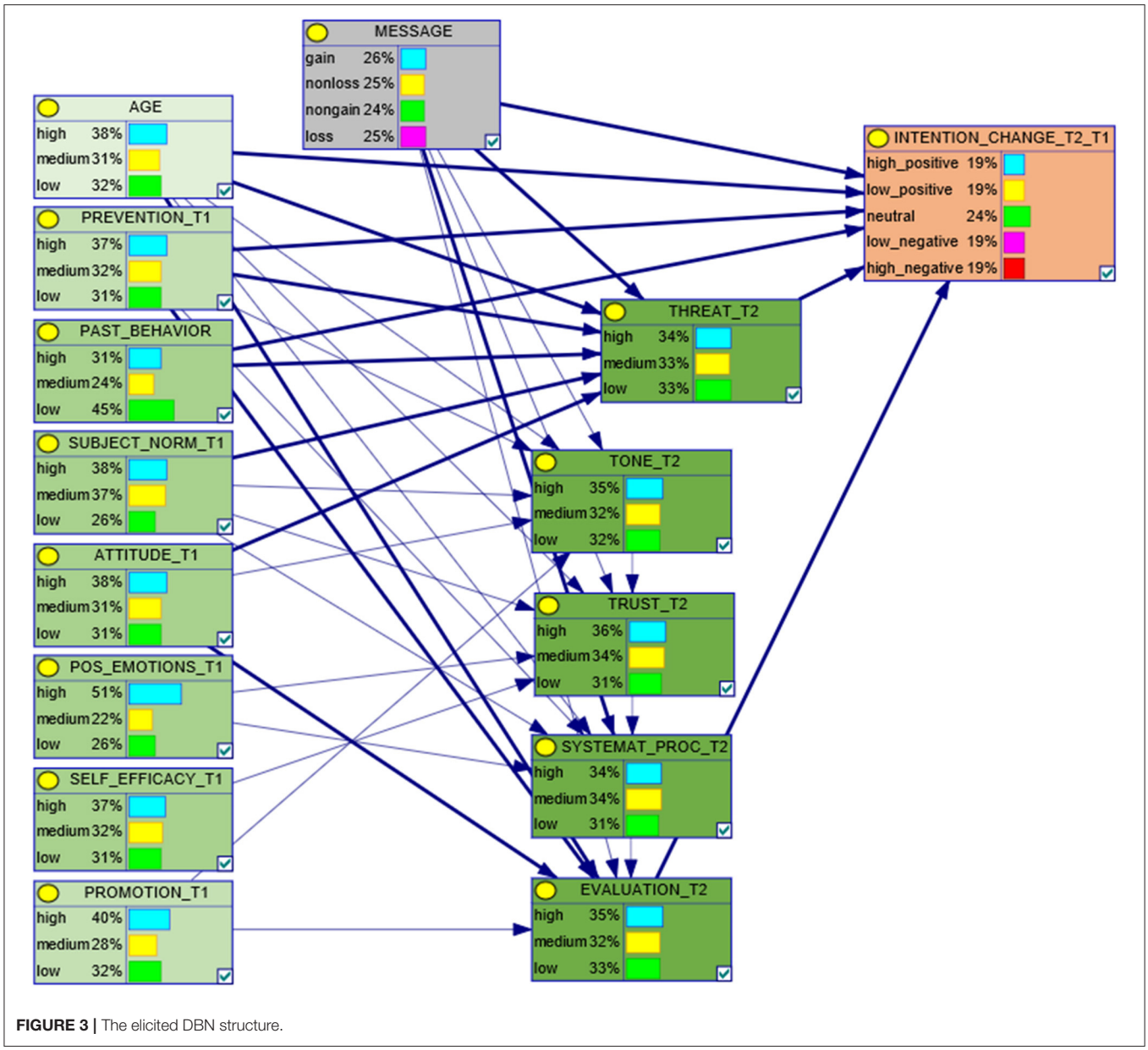

FIGURE 3 | The elicited DBN structure.

2002). In most cases, however, these strategies are stochastic, since they imply random choices of some sort (Scanagatta et al., 2019). In our study, we preferred adopting a more problem-specific and deterministic search strategy together with a suitable scoring function, as it will be explained in section Deterministic Structure Search.

Explanatory Power vs. Predictive Capability

Given the stated purposes, our objective was to achieve a DBN that could predict the value of the target variable *Intention Change* (whose index was computed subtracting *Intention* at Time 1 to *Intention* at Time 2) relying only on Time 1 observations and Message Intervention. In other words, the objective was estimating the conditional probability:

$$\mathrm{p}\left(\text{target variable} \,\middle|\, \text{Time 1 observations},\ \text{Message Intervention}\right)$$

for all message types considered. One possible way of evaluating the effectiveness of a categorical predictor of this kind is through *accuracy*. Calling X_t the target variable, for conciseness, the value predicted by the DBN will be:

$$v_{pred} := \underset{v}{\operatorname{argmax}}\, p\left(X_t = v \,\middle|\, Obs,\ Msg\right)$$

TABLE 1 | Means and standard deviations of the study measures.

Time 1			Time 2		
Measure	***M***	***SD***	**Measure**	***M***	***SD***
Prevention	3.68	1.39	Message-induced threat	5.79	1.52
Promotion	5.35	0.91	Message-induced distress	4.92	1.17
Past behavior	2.63	1.89	Message tone	5.19	1.29
Past outdoor behavior	4.39	1.76	Message trust	5.47	0.98
Attitude	1.21	0.43	Systematic processing	4.97	1.24
Perceived behavioral control	4.92	1.17	Message evaluation	4.60	1.24
Subjective norm	5.19	1.25	Intention	5.17	1.70
Anticipated positive emotions	5.43	1.46			
Anticipated negative emotions	4.36	1.76			
Intention	5.15	1.75			

where v is one of the categorical values of X_t and p is the probability computed by the DBN. Accuracy is computed by considering each participant in the data collection, computing the probability of each value v given Time 1 observations and the Message Intervention that has been delivered to the participant in point. Accuracy is defined as the ratio of how many times we succeed in having:

$$v_{pred} = v_{true}$$

where v_{true} is the value actually observed, over the size N of the dataset.

Given our objectives, the effectiveness of the DBN was intended as a balance between maintaining a clear connection with the theoretical background of reference and the generalization capability of predicting the target index for unseen subjects, given limited observations. In this perspective, accuracy could be evaluated both in-sample, for data explanation, and out-of-sample, to assess the predictive power of a DBN. In-sample accuracy can be evaluated by first learning the DBN parameters from the entire dataset, as described in section Learning Structure and Parameters From Data, and then predicting the target index in each record individually, in the same dataset, using partial observations only. Out-of-sample accuracy can be estimated via the *k-fold cross-validation* method (Allen, 1974). In our case, however, we preferred the *leave-one-out* method (Raschka, 2018): one participant d is removed from the dataset D, then probabilities θ are learnt from $(D - d)$ and accuracy is tested for d. The procedure is repeated for all participants in D and the resulting success ratio is computed.

Accuracy, however, is a somewhat crude measure in that it considers only the highest probability value, conditioned on known information, and not the entire distribution. A better metrics is *Area Under Curve* (AUC; Fawcett, 2006) which measures the area under the curve traced by points:

$$\left(p\left(FP \mid \gamma\right),\ p\left(TP \mid \gamma\right)\right)$$

where *FP* and *TP* are *False Positive* and *True Positive* value assignments, respectively, obtained when accepting a predicted value v whenever $p\left(X_t = v\right) \geq \gamma$, and γ varies in $[0, 1]$. Such curve is also called *Received Operating Characteristic* (*ROC*). Examples of ROC curves are shown in **Figure 4**. Given that the target variable in our case had five categorical values, in the present study the multiclass version of AUC (i.e., mAUC–Hand and Till, 2001) was used.

In summary, in our study we computed the mAUC values for both in-sample and out-of-sample (i.e., through leave-one-out) validation and we considered the average of the two as our main scoring function for selecting the best possible structure of the DBN.

Deterministic Structure Search

Despite its advantages, computing the mAUC is expensive (in particular for the leave-one-out validation) and this does not match well with the complexity of structure searching. This raises the need to pre-select candidate structures using a more conveniently computable scoring function.

In this perspective, as shown by Koller and Friedman (2009), the log-likelihood function can be expressed as:

$$l\left(\theta, D\right) = \mathrm{N}\left(\sum_i IG\left(X_i; \pi\left(X_i\right)\right) - \sum_i H\left(X_i\right)\right)$$

where N is the size of the dataset, H is the *entropy*:

$$H\left(X\right) := -\sum_X p\left(X\right) \log p\left(X\right)$$

and IG is the *information gain* (Jiang et al., 2015):

$$IG\left(X; Y_1, \ldots, Y_n\right) := H\left(X|Y_1, \ldots, Y_n\right) - H\left(X\right)$$

where the *conditional entropy* is defined as:

$$H\left(X|Y_1, \ldots, Y_n\right) := -\sum_{X, Y_1, \ldots, Y_n} p\left(X, Y_1, \ldots, Y_n\right) \log \frac{p\left(X, Y_1, \ldots, Y_n\right)}{p\left(Y_1, \ldots, Y_n\right)}$$

In all the above equations, p can be construed as the empirical probability distribution, estimated as frequency ratios in the dataset.

In other terms, in the above decomposition the log-likelihood score is shown to be proportional to information gain of the

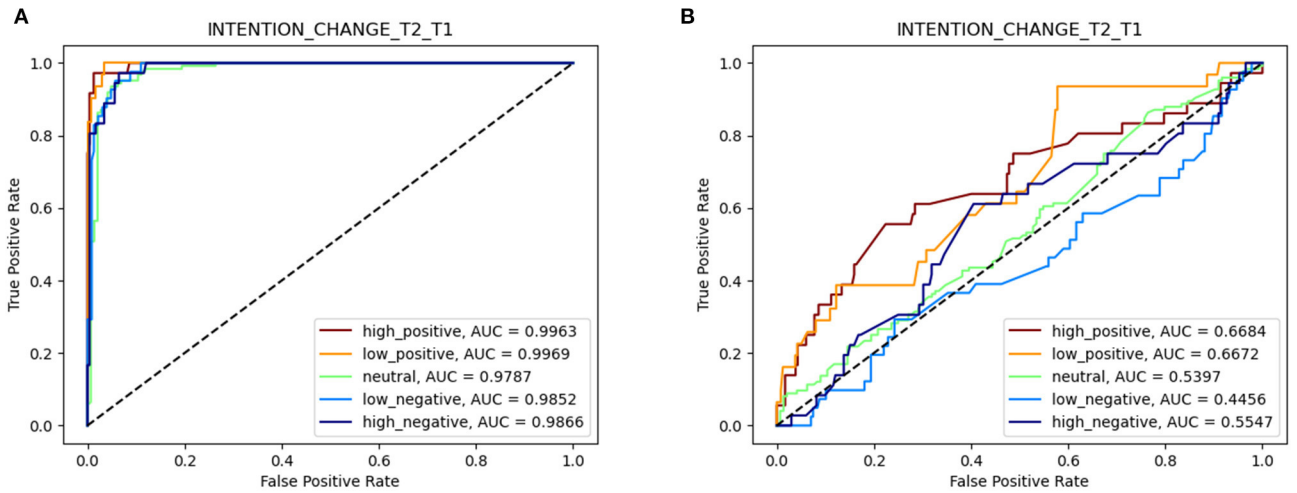

FIGURE 4 | Multivalued ROC curves obtained for the elicited DBN. elicited DBN. **(A)** In-sample ROCs. **(B)** Out-of-sample ROCs.

conditional probabilities in $\mathcal{B}$ minus a constant entropy term, i.e., which does not depend on the structure of $\mathcal{B}$. Furthermore, information gain values are terms in a sum and could be optimized separately, within the limit of not introducing cyclic dependencies in the graph.

In the light of the above, in our study we used information gain as a preliminary scoring function, to select the most promising structures. We then computed the combined mAUC metrics (i.e., in-sample and out-of-sample) of the later structures, to select the most effective one. Our procedure was as follow:

1. We first considered the target variable X_t and we computed the information gain for all possible subsets of parents of size in between 2 and 8, chosen among all other random variables (i.e., Time 1, message intervention, Time 2).
2. Having selected the best subsets of parents for X_t, one per each size in the above range, we expanded each Time 2 variable in each selected parenthood by measuring the information gain of all possible subsets of size in between 2 and 8, chosen among the remaining variables, avoiding cycles.
3. For each combination of sizes (i.e., one for the parenthood X_t and one for the parenthood of each Time 2 node), we pre-selected one structure, namely the one with the largest overall information gain, hence the highest likelihood.
4. For all the pre-selected structures we computed the combined in-sample and out-of-sample mAUC metrics, to select the most effective one.

Note that step 2 above was completed when all Time 2 nodes became expanded, so that all of them had a parenthood rooted in Time 1 nodes, either directly or indirectly. The need to do so derived from the objective of achieving a predictor of the target variable *Intention Change* that relies on Time 1 observations only.

To avoid a combinatorial explosion in the number of candidate structures, in the above procedures all parenthoods of Time 2 nodes in each structure were imposed to have the same size. For instance, in the structure that resulted as best in its combination of ranges (see **Figure 3**) all Time 2 nodes have 6 parents exactly. Clearly, this entails the risk of a certain redundancy in the structures produced. To evaluate this aspect, for all selected structures, we also computed the *interaction strength* (Zeng et al., 2016) on each set of parents:

$$IS\left(X;\, Y_1, \ldots, Y_n\right) := IG\left(X;\, Y_1, \ldots, Y_n\right) - \sum_i IG\left(X;\, Y_i\right)$$

Interaction strength measures the difference between the cumulative information gain of a subset of parents for a given variable over the sum of each individual information gains in the same subset. Unlike information gain, interaction strength is not monotonically increasing with the number of variables but has a peak that is expected to correspond to the strongest interacting parenthood. In our case, interaction strength was computed, for the selected structures, for all possible combinations of parents among the ones selected through the procedure described above.

The relevant advantage of the above chosen method is that the structure selection procedure is entirely deterministic and repeatable. The theoretical aspects of psychosocial models play a crucial role in the initial phase of dimensions and measures selection, whereas their interrelations are hypothesized only implicitly. Subsequently, starting from the analysis of the experimental data, structure and parameters of the probabilistic predictor are learned in an automatic way, by assuming the target variable *Intention Change* and the temporal sequence of events as the only constraints. The results thus obtained are in keeping with the implicit theoretical assumptions and this adds credibility to the proposed procedure.

RESULTS

The DBN structure described in **Figure 3** resulted as the best one among those generated via the procedure described in section Deterministic Structure Search, applied to the dataset of experimental measures. **Figure 4** describes the multivalued ROC curves obtained for the DBN in **Figure 3**, with in-sample and out-of-sample tests, respectively. The latter test was performed with the leave-one-out technique. In these tests, the DBN in point scored a combined mAUC value of 0.783 (with in-sample and out-of-sample values of 0.989 and 0.577, respectively).

As anticipated in the previous section, all parenthoods in the DBN were tested for interaction strength. The strongest interaction subsets in each parenthood are shown by thicker arrows in **Figure 3**. As it could be expected, the parenthood of the target variable *Intention Change* resulted as coincident with the strongest interacting subset. The same resulted for variable *Threat*. On the other hand, the strongest interacting subset for variable *Evaluation* included just 3 of 6 parents. Time 2 variables *Tone*, *Trust,* and *Systematic Processing* could not be found among the strongest interacting parenthoods.

Interestingly however, although to a minor extent, even marginal interactions were proven to have a role in determining the overall performance of the DBN in point. In fact, the reduced DBN structure obtained by considering only the thicker arrows in **Figure 3** and by discarding unconnected nodes, scored a combined mAUC value of 0.762 (0.960, 0.565). This result is also representative of the fact that, in our case, interaction strength did not prove to be as effective as the information gain for the pre-selection of candidate DBN structures.

For the results presented, the action of learning DBN parameters was performed, for both in-sample and out-of-sample tests, via the EM algorithm as implemented in the SMILE library, by BayesFusion[1]. All other computations were performed with custom code, made with Python and Numpy[2]. The complete definition of the DBN structure described in **Figure 3** can be found in the same open repository mentioned in section Participants and Procedure.

DISCUSSION

As part of an interdisciplinary project between social psychology and artificial intelligence, in this paper we presented a deterministic method for the elicitation of a DBN, starting from data on the psychosocial antecedents of the intention to exercise at home and intention change after being exposed to persuasive messages on the issue. This method constitutes a first step toward the development of deep reinforcement learning techniques which will allow devising personalized interaction strategies based on consolidated psychosocial models of behavior change. In this discussion, we will first focus on the theoretical consistency of the elicited DBN and we will then describe its strengths and limits.

[1]See https://www.bayesfusion.com/

[2]See https://numpy.org/

Theoretical Consistency of the Elicited DBN

The DBN structure that emerged from the analysis turned out to be largely consistent with the psychosocial literature of reference. It also highlighted the presence of interesting relationships between measures related to the different psychosocial theories we referred to when devising our integrated model. We will now illustrate the DBN structure analyzing the strongest links between the variables and interpreting them in the light of the psychosocial theories we referred to when selecting the variables to be included in the initial model.

We start by examining the direct predictors of *Intention Change*, i.e., change in the intention to exercise at home after reading the messages. Message framing directly predicted *Intention Change*, suggesting that the four different message frames employed in the study affected differently the observed changes in the behavioral intention of the recipients. Message-induced threat also had a direct impact on *Intention Change* and was in turn directly influenced by message framing. Therefore, different message frames triggered different levels of perceived threat in the recipients, which in turn influenced the change in the intention to exercise at home. This finding is consistent with previous research in the domain of the effects of communication on health. According to the psychological reactance theory, when individuals feel that a health message is prompting them to accept a certain behavior, they may not process it accurately and instead respond defensively, downplaying its recommendation and not changing their intention (Liberman and Chaiken, 1992; Falk et al., 2015; Howe and Krosnick, 2017). According to the theory of self-affirmation (Steele, 1988; Sherman and Cohen, 2006), this defensive reaction against threatening messages is based on the attempt to maintain the perception of being able to control the relevant results. When this defensive mechanism is activated, people can attempt to protect it by rejecting such threatening information (e.g., Strachan et al., 2020).

Message evaluation also had a direct influence on *Intention Change* and was directly influenced by message framing. Message evaluation was also influenced, albeit less strongly, by the systematic processing of the message, which in turn was influenced by trust in the message and the perceived positive or negative tone of the message itself. This chain of influences is consistent with previous literature on persuasive communication showing that intention changes depend upon the likelihood of a persuasive message being positively evaluated by the receiver (Petty and Cacioppo, 1986; Eagly and Chaiken, 1993). The positive evaluation of a message, in turn, depends on systematic processing (Chaiken, 1980), which implies cognitive effort in considering the content of a message. Previous literature also showed that people tend to evaluate the trustworthiness of a message before processing it (Schlegelmilch and Pollach, 2005). Finally, trust in a message is influenced by how receivers perceive its tone. A negative tone can more easily be perceived as an open persuasive attempt and can therefore induce lower trust toward the message (Yalch and Dempsey, 1978).

Intention Change was directly predicted not only by message framing and message-related variables, but also by three variables measured at Time 1, namely, participants' age, frequency of

past exercising at home, and prevention focus. Besides having a direct impact on *Intention Change*, participants' age had an indirect impact on it, through the mediation of message-induced threat and message evaluation. These results are consistent with a vast amount of past studies showing the effect of age on physical activity over lifespan (Varma et al., 2017), also during the COVID-19 pandemic (Alomari et al., 2020). Unlike age, gender and education did not have either a direct or indirect effect on *Intention Change*. This result is consistent with McCarthy et al. (2021), who found that socioeconomic group and gender were not associated with changes in physical activity during the COVID-19 restrictions. As to the frequency of past home exercising, it predicted *Intention Change* both directly and via the mediation of message-induced threat. This finding is strongly supported by past research, which offers wide evidence that past behavior is one of the largest contributors to the explanation of physical activity (Young et al., 2014). It is worth noting that the frequency of physical exercise outside home (which was also part of the initial model) did not enter in the final DBN and therefore did not turn out to be among the main predictors of *Intention Change*. This result may be explained by the fact that people do not perceive physical activity at home as equivalent to physical activity outside home, and therefore this latter activity may not play a significant role in predicting a change in the intention to train at home.

Prevention focus also directly predicted a change in the behavioral intention. It had both a direct influence on *Intention Change* and an indirect influence, via the mediation of message-induced threat and message evaluation. Avoidance of losses and the fulfillment of duties and obligations evidently influenced a change in recipients' intention after being exposed to differently framed messages fostering exercise at home. This result is consistent with previous research showing that the effect of differently framed messages may vary according to the recipient's regulatory focus (Latimer et al., 2008b; Pfeffer, 2013). In our study, the promotion focus also had a link, albeit only an indirect one, with *Intention Change*. However, it was a weaker link than the one of the prevention focus, mediated only by the evaluation of the message and not also by the threat induced by the message, as was the case with the prevention focus. Understanding why prevention focus had more impact on *Intention Change* than promotion focus would require analyses that go beyond the ones presented in this paper. For example, it may be the case that individuals with a high promotion focus are basically more oriented to do physical activity than individuals with a high prevention focus, to achieve an ideal of well-being and health. If so, their intention to do physical activity may be already high and therefore they would be less likely to be persuaded to further enhance this activity by messages focused on the issue.

As to the extended TPB variables measured at Time 1 (past behavior, attitude, subjective norm, perceived behavioral control, and anticipated emotions), as discussed above only past behavior had a direct impact on *Intention Change*. Attitude and subjective norm also had an influence on *Intention Change*, but this influence was mediated by message-related variables. Attitude had an influence on *Intention Change* via the mediation of message-induced threat and message evaluation. This result is consistent with previous studies on the influence of attitudes and message framing on intention change in health-related domains (e.g., Carfora and Catellani, 2021; Caso et al., 2021). Subjective norm had an impact on *Intention Change* via the mediation of message-induced threat. Previous research showed that subjective norm may exert its influence on intention through perceived threat (Becker and Maiman, 1975). Consistently, we can hypothesize that when people attach importance to the recommendations and expectations of others, they may tend to feel more threatened by the risks presented in persuasive messages. A confirmation of this link would, however, deserve further empirical support.

Overall, the DBN structure that emerged from our analysis was largely consistent with the psychosocial literature in the area. At the same time, it contributed to enrich it, showing the presence of interesting and plausible links between variables belonging to the three different psychosocial theories that we took as a reference when constructing the initial model.

Methodological Strengths of the Elicited DBN

The approach we followed in the elicitation of the DBN has several methodological strengths which can be traced back to three main points.

First, in our method the structure selection procedure was entirely deterministic and repeatable and nevertheless, as discussed above, led to a structure which was theoretically consistent. Notably, the adoption of the discretization of the values of the psychosocial measures on the one hand necessarily introduced approximations, but on the other hand simplified data analysis and allowed the identification of a significant structure from a small sample.

Second, the intention of balancing explanatory power with predictive capability led us to adopting a selection metric for eliciting the DBN which, albeit at the cost of increased computation complexity, effectively counteracted the tendency of the common likelihood metrics to reward the most complex structures. In this way, we believe it is also possible to prevent the overfitting, intended as the result of overestimating in-sample over out-of-sample performances, of structural models with respect to the sample of collected data (Yarkoni and Westfall, 2017). As a matter of fact, the in-sample and out-of-sample performances of the elicited DBN were divergent in the measured values (see **Figure 4**). Nevertheless, it is reasonable to expect that such gap could significantly decrease whenever the size and relevance of the sample could be made to increase.

Third, the DBN obtained was effective from both an explanatory and predictive point of view. In particular, the structure of the DBN was easy to interpret and relate to the psychological models that were assumed as the starting point. Its efficacy is a first important step for the creation of an artificial intelligence system that will translate the

results of psychological research into automatic interaction and interventions policies for improving many people's lives. Once fully operational, these systems will require less time and economic efforts to be operated, compared to those required by putting the same psychological models at work through human intervention alone.

Limits

Our research has some limitations, related to the quality of the data collected, data analysis and the development of the DBN. As for the data, these were collected on a non-representative sample of the population and with reference to the intention to carry out physical activity at home in a very particular historical moment, that of the first wave of the Covid-19 pandemic. This makes it difficult to extend our results to different populations and times. Furthermore, it should be noted that the measurement of the effectiveness of the messaging interventions employed was based on the change in the intention to carry out physical activity at home and not on measures relating to the actual performance of this activity, such as those that may be offered by bracelets or wearable sensors worn by participants. Regarding the intention measurement, we used a Likert scale that measured the participants' agreement with intending to do physical exercises at home. Future scale should instead use probability scales to reduce the likelihood of response-style biases (Morwitz and Munz, 2021).

As for data analysis and learning of structure and parameters of the DBN, the reduced size of data sample was definitely a limiting factor, as it can be observed in the divergence between in-sample and out-of-sample performances (see **Figure 4**). Therefore, the actual effectiveness of the predictor obtained should be further tested in a real-world application scenario.

Future Developments

The method for DBN elicitation described in this paper constitutes the first part of an articulated path. This same method is currently being tested within a purpose-specific framework based on Deep Reinforcement Learning (DRL; François-Lavet et al., 2018; Sutton and Barto, 2018) to train a Deep Neural Network component, which is intended to drive online interactions with actual people, by applying the psychosocial principles described.

Further on, the DRL software framework under construction is expected to evolve to include the capability to collect additional experience and allow the incremental improvement of the DBN itself. In this perspective, the DBN is intended to play a fundamental role, in guaranteeing the explainability of the behavior of the AI system, giving to both psychologists and experts of artificial intelligence the power to monitor and intervene in the learning procedure.

Thanks to the application of DRL techniques it will be possible to calculate the utility deriving from sending messages with different framing to people who differ from each other as regards the psychosocial dimensions underlying the behavior under study.

CONCLUSION

In conclusion, our results show that social psychology and artificial intelligence can usefully interact to develop automatic interaction strategies aimed at supporting behavior change in the direction of well-being. As we have seen, this interaction helps overcoming some of the constraints the two disciplines often encounter when developing models that are expected to find application in real life. The possibilities of applying a methodology such as the one tested here are many and concern various areas, virtually all those in which it is reasonable to think that sending personalized messages to the recipient through automatic systems can have positive effects for the well-being of the person. Much can therefore be done thanks to the integration of social psychology and artificial intelligence, moving from the assumptions that the wealth of processing and production of new data allowed by artificial intelligence systems can ultimately be a way to enrich and improve the experience of people, for whom artificial intelligence systems have reason to be.

DATA AVAILABILITY STATEMENT

All data presented in this study can be found in the open repository at https://bitbucket.org/unipv_cvmlab/connecting_social_psychology_and_drl/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Catholic University of the Sacred Heart. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PC proposed the research questions, planned the research design, and took responsibility for the manuscript. She also thoroughly revised the manuscript with regard to content and style. VC supervised data collection and analysis and participated in the interpretation of the results. MP designed the elicitation procedure for the probabilistic predictor, implemented the code, and carried out the computational experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This study was part of the project Re-HUB-ility: Rehabilitative Personalized Home System and Virtual Coaching for Chronic Treatment in elderly supported by Call HUB Ricerca e Innovazione, Regione Lombardia and by Athics s.r.l. (Grant Number: D.G.R. N. 727 of 5/11/2018; decreto 18854 del 14/12/2018).

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

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-t

Aldenaini, N., Alqahtani, F., Orji, R., and Sampalli, S. (2020). Trends in persuasive technologies for physical activity and sedentary behavior: a systematic review. *Front. Artif. Intell.* 3:7. doi: 10.3389/frai.2020.00007

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125–127. doi: 10.1080/00401706.1974.10489157

Alomari, M. A., Khabour, O. F., and Alzoubi, K. H. (2020). Changes in physical activity and sedentary behavior amid confinement: the bksq-covid-19 project. *Risk Manag. Healthc. Policy* 13:1757. doi: 10.2147/RMHP.S268320

Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Becker, M. H., and Maiman, L. A. (1975). Socio-behavioral determinants of compliance with health and medical care recommendations. *Med. Care* 13, 10–24. doi: 10.1097/00005650-197501000-00002

Bertolotti, M., Carfora, V., and Catellani, P. (2020). Regulatory focus and the effect of nutritional messages on health and well-being: the case of red meat intake. *Appl. Psychol. Health Well-being* 12, 212–230. doi: 10.1111/aphw.12180

Brown, S. L., and Smith, E. Z. (2007). The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. *Psychol. Health* 3, 255–268. doi: 10.1080/14768320600843127

Carfora, V., Caso, D., Palumbo, F., and Conner, M. (2018). Promoting water intake. The persuasiveness of a messaging intervention based on anticipated negative affective reactions and self-monitoring. *Appetite* 130, 236–246. doi: 10.1016/j.appet.2018.08.017

Carfora, V., and Catellani, P. (2021). The effect of persuasive messages in promoting home-based physical activity during covid-19 pandemic. *Front. Psychol.* 12:644050. doi: 10.3389/fpsyg.2021.644050

Carfora, V., Conner, M., Caso, D., and Catellani, P. (2020a). Rational and moral motives to reduce red and processed meat consumption. *J. Appl. Soc. Psychol.* 50, 744–755. doi: 10.1111/jasp.12710

Carfora, V., Di Massimo, F., Rastelli, R., Catellani, P., and Piastra, M. (2020b). Dialogue management in conversational agents through psychology of persuasion and machine learning. *Multimed. Tools. Appl.* 79, 35949–35971. doi: 10.1007/s11042-020-09178-w

Caso, D., Carfora, V., Capasso, M., Oliano, D., and Conner, M. (2021). Using messages targeting psychological versus physical health benefits to promote walking behaviour: a randomised controlled trial. *Appl. Psychol. Health Well-being* 13, 152–173. doi: 10.1111/aphw.12224

Cesario, J., Corker, K. S., and Jelinek, S. (2013). A self-regulatory framework for message framing. *J. Exp. Soc. Psychol.* 49, 238–249. doi: 10.1016/j.jesp.2012.10.014

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* 39, 752–766. doi: 10.1037/0022-3514.39.5.752

Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W. (2002). Learning Bayesian networks from data: an information-theory based approach. *Artif. Intell.* 137, 43–90. doi: 10.1016/S0004-3702(02)00191-1

Clark, H., and Bassett, S. (2014). An application of the health action process approach to physiotherapy rehabilitation adherence. *Physiother. Theor. Pract.* 30, 527–533. doi: 10.3109/09593985.2014.912710

Dagum, P., Galper, A., Horvitz, E., and Seiver, A. (1995). Uncertain reasoning and forecasting. *Int. J. Forecast* 11, 73–87, doi: 10.1016/0169-2070(94)02009-E

Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge, MA: Cambridge University Press.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

Di Massimo, F., Carfora, V., Catellani, P., and Piastra, M. (2019). "Applying psychology of persuasion to conversational agents through reinforcement learning: an exploratory study," in *CEUR – Workshop Proceedings, Vol. 2481*, 27.

Dijkhuis, T. B., Blaauw, F. J., Van Ittersum, M. W., Velthuijsen, H., and Aiello, M. (2018). Personalized physical activity coaching: a machine learning approach. *Sensors* 18:623. doi: 10.3390/s18020623

Eagly, A. H., and Chaiken, S. (1993). *The Psychology of Attitudes*. Fort Worth, TX: Harcourt, Brace, & Janovich.

Falk, E. B., O'Donnell, M. B., Cascio, C. N., Tinney, F., Kang, Y., Lieberman, M. D., et al. (2015). Self-affirmation alters the brain's response to health messages and subsequent behavior change. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1977–1982. doi: 10.1073/pnas.1500247112

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Ferrer, R. A., Lipkus, I. M., Cerully, J. L., McBride, C. M., Shepperd, J. A., and Klein, W. M. (2017). Developing a scale to assess health regulatory focus. *Soc. Sci. Med.* 195, 50–60. doi: 10.1016/j.socscimed.2017.10.029

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning. *Found. Trends Mach. Learn.* 11, 219–354. doi: 10.1561/2200000071

Godinho, C. A., Alvarez, M. J., and Lima, M. L. (2016). Emphasizing the losses or the gains: comparing situational and individual moderators of framed messages to promote fruit and vegetable intake. *Appetite* 96, 416–425. doi: 10.1016/j.appet.2015.10.001

Hand, D. J., and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.* 45, 171–186. doi: 10.1023/A:1010920819831

Higgins, E. T. (1997). Beyond pleasure and pain. *Am. Psychol.* 52, 1280–1300. doi: 10.1037/0003-066x.52.12.1280

Howe, L. C., and Krosnick, J. A. (2017). Attitude strength. *Ann. Rev. Psychol.* 68, 327–351. doi: 10.1146/annurev-psych-122414-033600

Hyde, A. L., Maher, J. P., and Elavsky, S. (2013). Enhancing our understanding of physical activity and wellbeing with a lifespan perspective. *Int. J. Com. Wellbeing* 3, 98–115. doi: 10.5502/ijw.v3i1.6

Jiang, X., Jao, J., and Neapolitan, R. (2015). Learning predictive interactions using information gain and bayesian network scoring. *PLoS ONE* 10:e0143247. doi: 10.1371/journal.pone.0143247

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: The MIT Press.

Latimer, A. E., Rench, T. A., Rivers, S. E., Katulak, N. A., Materese, S. A., Cadmus, L., et al. (2008a). Promoting participation in physical activity using framed messages: an application of prospect theory. *Br. J. Health Psychol.* 13, 659–681. doi: 10.1348/135910707X246186

Latimer, A. E., Rivers, S. E., Rench, T. A., Katulak, N. A., Hicks, A., Hodorowski, J. K., et al. (2008b). A field experiment testing the utility of regulatory fit messages for promoting physical activity. *J. Exp. Soc. Psychol.* 44, 826–832. doi: 10.1016/j.jesp.2007.07.013

Liberman, A., and Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Pers. Soc. Psychol. Bull.* 18, 669–679. doi: 10.1177/0146167292186002

McCarthy, H., Potts, H. W., and Fisher, A. (2021). Physical activity behavior before, during, and after COVID-19 restrictions: Longitudinal smartphone-tracking study of adults in the United Kingdom. *J. Med. Internet. Res.* 23:e23701. doi: 10.2196/23701

Morwitz, V. G., and Munz, K. P. (2021). Intentions. *Consum. Psychol. Rev.* 4, 26–41. doi: 10.1002/arcp.1061

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Approach*. Cambridge, MA: The MIT Press.

Nojavan, A. F., Qian, S. S., and Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environ. Modell. Softw.* 97, 64–71. doi: 10.1016/j.envsoft.2016.10.007

Petty, R.E. (2018). *Attitudes and Persuasion: Classic and Contemporary Approaches*. London: Routledge.

Petty, R. E., and Cacioppo, J. T. (1986). "The elaboration likelihood model of persuasion," in *Communication and Persuasion*, eds R. E. Petty and J. T. Cacioppo (New York, NY: Springer), 1–24.

Pfeffer, I. (2013). Regulatory fit messages and physical activity motivation. *J. Sport Exerc. Psychol.* 35, 119–131. doi: 10.1123/jsep.35.2.119

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv:1811.12808*.

Rhodes, R. E., Janssen, I., Bredin, S. S., Warburton, D. E., and Bauman, A. (2017). Physical activity: health impact, prevalence, correlates, and interventions. *Psychol. Health* 32, 942–975. doi: 10.1080/08870446.2017.1325486

Scanagatta, M., Salmerón, A. and Stella, F. (2019). A survey on Bayesian network structure learning from data. *Prog. Artif. Intell.* 8, 425–439. doi: 10.1007/s13748-019-00194-y

Schlegelmilch, B. B., and Pollach, I. (2005). The perils and opportunities of communicating corporate ethics. *J. Mark. Manage.* 21, 267–290. doi: 10.1362/0267257053779154

Shen, L. (2015). Antecedents to psychological reactance: the impact of threat, message frame, and choice. *Health Comm.* 30, 975–985. doi: 10.1080/10410236.2014.910882

Sherman, D. K., and Cohen, G. L. (2006). The psychology of self-defense: self-affirmation theory. *Adv. Exp. Soc. Psychol.* 38, 183–242. doi: 10.1016/S0065-2601(06)38004-5

Smerecnik, C. M., Mesters, I., Candel, M. J., De Vries, H., and De Vries, N. K. (2012). Risk perception and information processing: the development and validation of a questionnaire to assess self-reported information processing. *Risk Anal. Int. J.* 32, 54–66. doi: 10.1111/j.1539-6924.2011.01651.x

Steele, C. M. (1988). The psychology of self-affirmation: sustaining the integrity of the self. *Adv. Exp. Soc. Psychol.* 21, 261–302. doi: 10.1016/S0065-2601(08)60229-4

Strachan, S. M., Myre, M., Berry, T. R., Ceccarelli, L. A., Semenchuk, B. N., and Miller, C. (2020). Self-affirmation and physical activity messages. *Psychol. Sport Exerc.* 47:101613. doi: 10.1016/j.psychsport.2019.101613

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction, 2nd Edn*. Bradford Books.

Taylor, S., Landry, C., Paluszek, M., Fergus, T. A., McKay, D., and Asmundson, G. J. (2020). Development and initial validation of the COVID stress scales. *J. Anxiety Disord.* 72:102232. doi: 10.1016/j.janxdis.2020.102232

University of Virginia Health System (2020). *COVID-19: Exercise May Help Prevent Deadly Complication*. Available online at: https://newsroom.uvahealth.com/2020/04/15/covid-19-exercise-may-help-prevent-deadly-complication/ (accessed April 15, 2020).

Varma, V. R., Dey, D., Leroux, A., Di, J., Urbanek, J., Xiao, L., et al. (2017). Re-evaluating the effect of age on physical activity over the lifespan. *Prevent. Med.* 101, 102–108. doi: 10.1016/j.ypmed.2017.05.030

Yalch, R. F., and Dempsey, M. C. (1978). "Selling a city: an experimental study of the communication effects of message tone," in *NA - Advances in Consumer Research, Vol. 5*, ed K. Hunt (Ann Abor, MI: Association for Consumer Research), 5–11.

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Yi, S., and Baumgartner, H. (2009). Regulatory focus and message framing: a test of three accounts. *Motiv. Emot.* 33, 435–443. doi: 10.1007/s11031-009-9148-y

Young, M. D., Plotnikoff, R. C., Collins, C. E., Callister, R., and Morgan, P. J. (2014). Social cognitive theory and physical activity: a systematic review and meta-analysis. *Obes. Rev.* 15, 983–995. doi: 10.1111/obr.12225

Zeng, Z., Jiang, X., and Neapolitan, R. (2016). Discovering causal interactions using Bayesian network scoring and information gain. *BMC Bioinform.* 17:221. doi: 10.1186/s12859-016-1084-8

Zhang, J., Oh, Y. J., Lange, P., Yu, Z., and Fukuoka, Y. (2020). Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *J. Med. Int. Res.* 22:e22845. doi: 10.2196/22845

Conflict of Interest: The authors declare that this study received funding from Athics s.r.l.. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

frontiers in Psychology

PERSPECTIVE
published: 26 July 2021
doi: 10.3389/fpsyg.2021.697101

"The Flow in the Funnel": Modeling Organizational and Individual Decision-Making for Designing Financial AI-Based Systems

Alessandra Talamo, Silvia Marocco and Chiara Tricol*

Department of Social and Developmental Psychology, Sapienza University of Rome, Rome, Italy

Nowadays, the current application of artificial intelligence (AI) to financial context is opening a new field of study, named financial intelligence, in which the implementation of AI-based solutions as "financial brain" aims at assisting in complex decision-making (DM) processes as wealth and risk management, financial security, financial consulting, and blockchain. For venture capitalist organizations (VCOs), this aspect becomes even more critical, since different actors (shareholders, bondholders, management, suppliers, customers) with different DM behaviors are involved. One last layer of complexity is the potential variation of behaviors performed by managers even in presence of fixed organizational goals. The aim of this study is twofold: a general analysis of the debate on implementing AI in DM processes is introduced, and a proposal for modeling financial AI-based services is presented. A set of qualitative methods based on the application of cultural psychology is presented for modeling financial DM processes of all actors involved in the process, machines as well as individuals and organizations. The integration of some design thinking techniques with strategic organizational counseling supports the modeling of a hierarchy of selective criteria of fund-seekers and the creation of an innovative value proposition accordingly with goals of VCOs to be represented and supported in AI-based systems. Implications suggest that human/AI integration in the field can be implemented by developing systems where AI can be conceived in two distinct functions: (a) automation: treating Big Data from the market defined by management of VCO; and (b) support: creating alert systems that are coherent with ordered weighted decisional criteria of VCO.

Keywords: decision-making, financial intelligence, artificial intelligence, qualitative methods, organizational psychology, social ergonomics

OPEN ACCESS

Edited by:
Ilaria Durosini, European Institute of Oncology (IEO), Italy

Reviewed by:
Angela Sorgente, Catholic University of the Sacred Heart, Italy
Galena Pisoni, University of Nice Sophia Antipolis, France

****Correspondence:***
Silvia Marocco
silvia.marocco@uniroma1.it

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *18 April 2021*
Accepted: *29 June 2021*
Published: *26 July 2021*

Citation:
Talamo A, Marocco S and Tricol C (2021) "The Flow in the Funnel": Modeling Organizational and Individual Decision-Making for Designing Financial AI-Based Systems. Front. Psychol. 12:697101. doi: 10.3389/fpsyg.2021.697101

INTRODUCTION

Artificial intelligence applied to decision-making (DM) processes has already been implemented in many fields (Galiano et al., 2019; Triberti et al., 2020; Bayrak et al., 2021) where it proves to have great potential. The current application of artificial intelligence (AI) to finance is nowadays opening a new field of study, named *financial intelligence*, in which the implementation of AI-based solutions as a "financial brain" aims at assisting in complex DM processes as wealth and risk management, financial security, financial consulting, and blockchain (Zheng et al., 2019).

Making investment decisions is usually considered a challenging task for investors, because it is a process based on risky, complex, and consequential choices (Shanmuganathan, 2020) on which trust both funding organizations and fund-seekers should rely. Furthermore, investment decisions are frequently influenced by emotional and cognitive biases, such as, overconfidence, and limited cognitive abilities. If individual DM is already challenging, at the layer of organizational contexts, such as venture capitalist organizations (VCOs), this aspect becomes even more critical, since different actors (shareholders, bondholders, management, suppliers, customers) with varying behaviors of DM are involved (De Bondt and Thaler, 1995). One last layer of complexity is the potential variation of behaviors performed by managers even in the presence of fixed organizational goals (Socea, 2012).

AI Contribution To The Field can be implemented in several aspects of the financial DM process, such as information collection and analysis, standardization of criteria of investments, and automation of customer interaction services. Nevertheless, recent findings show that the acceptance of AI-based solutions in DM by management is still an open issue within financial organizations since attitudes of manager toward intelligent agents are still unbalanced regarding human intervention in DM (Haesevoets et al., 2021).

Within this scenario, a core role can be played by tools that support AI modeling in designing financial AI-based solutions, which blend human/machine contribution in DM in the emerging field of explainable financial AI. In these AI-based solutions, the process on which results rely is transparent and understandable to users. What follows is an analysis of the debate on implementing AI in DM processes is introduced, and a proposal for modeling financial AI-based services is presented.

AI ROLE IN DM: STATE OF THE ART

Although the potential impact of AI in DM is proved to be significant (Zheng et al., 2019; Lepri et al., 2021), it led many practitioners and researchers in the field to take divergent points of view (Duan et al., 2019). The debate on human/technology relationships, even in AI, is not new: since the end of the last century, prominent scholars in the field have started positioning on contrasting perspectives, so that we can distinguish *techno-enthusiasts*, the true believers and supporters of technology and post-humanity, and *techno-skeptics*, who are more cautious and critical about future AI implementation in DM. These two divergent positions can be differentiated by focusing on specific issues:

Objectivity of AI vs. Subjectivity of Human Beings

On one hand, techno-enthusiasts believe that the objectivity conferred by technology is an added value because it reduces the variability of human error. Specifically, they argue that algorithmic DM processes can lead to more objective decisions than those made by humans, which may be influenced by individual bias, conflicts of interest, or fatigue (Lepri et al., 2021). On the other hand, techno-skeptics firmly state that machines can only partially simulate but never duplicate the unique mental life of humans; in fact, machines cannot feel or understand the complexity of real-life situations (Postman, 1993). Furthermore, in this perspective, the objectivity of AI and other intelligent technologies fails in making decisions with uncertain circumstances. Although AI systems can assist human decision-makers with predictive analytics, they are less capable of understanding common-sense situations (Guszcza et al., 2017) and unpredictable environments, particularly outside of a predefined domain of knowledge (Brynjolfsson and McAfee, 2012).

The Lack of Transparency of AI

One of the main concerns of techno-skeptics is the lack of transparency of AI. In this regard, most skeptics criticize algorithmic DM processes for the threat of privacy invasion, information asymmetry, and discrimination (Lepri et al., 2021). Moreover, AI and algorithmic DM processes are increasingly challenged for their black-box nature: although AI systems enable robust predictions, most users have little awareness and knowledge of how these systems make decisions. Hence, the lack of transparency hinders comprehension and negatively affects trust (Shin, 2021). This issue has fostered a new research field on explainable AI (XAI), which aims to substantially improve the trust and transparency of AI-based systems (Adadi and Berrada, 2018).

Augmentation vs. Automation

On one side, techno-enthusiasts aim at demonstrating the use of AI software systems and machines for automating tasks to eliminate human input. On the other, techno-skeptics are becoming more apprehensive, fearing that intelligent machines may soon take them over. In this regard, Stephen Hawking has noted that "the development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014), and Bill Gates has also stressed that humans should be concerned about the threat caused by AI (Rawlinson, 2015; Duan et al., 2019). As a result, some researchers have reframed the threat of automation as an opportunity for augmentation, proving that augmented intelligence can supplement and amplify human capabilities for cognition and collaboration (Miller, 2018).

Despite the fear and skepticism of some scholars, it is evident that the potential of AI implementation cannot be denied. According to the vision of some AI practitioners and researchers, it seems more meaningful to see AI in DM processes as an augmentation tool, able to extend human capabilities and judgments, rather than as automation, able to replace them (Miller, 2018; Wilson and Daugherty, 2018; Duan et al., 2019). In line with this last position, Jarrahi asserts that "artificial intelligence systems should be designed with the intention of augmenting, not replacing, human contributions" (Jarrahi, 2018, p. 584).

Within this debate, a question arises: how can humans and AI act in a complementary way in DM? The position starts from

a specific perspective on innovation in technological advances. In the 1970s, Thierry Gaudin (1978) developed a human-centered theory of innovation that may help us in reasoning in a practical way on the human–technology relationship. In the proposal of Gaudin, it is not just the technological development that promotes or inhibits innovation processes, but rather the behavior of organizations, considered as vital beings, with their missions, their evolutionary paths, and their modes of functioning (Talamo et al., 2016).

FROM HUMAN/MACHINE INTERACTION TO HUMAN/AI INTEGRATION: A PSYCHOLOGICAL PERSPECTIVE

Since the 1980s, a growing body of literature on human/machine interaction has produced consolidated evidence on the "external side" of user experience, that is, the front-end layer of interacting with systems. However, the fast development of AI implementation pushes us to reason on different layers, focusing on automation and replication of contextualized human reasoning models to shape the "internal side of technologies."

In the last 20 years, research on organizational disasters has already demonstrated the risk of taking an ingenuous perspective on technology implementation where technical, rationale, automatic, and general were considered preferable to practical, socialized, and contingent (Heath and Luff, 2000). Additionally, some highlighted the crucial role of proper treatment of information to support organizations and individuals in avoiding organizational disasters due to mistakes in information management in personal and collective DM processes (Choo, 2008). There is also growing evidence of the relevance of including ecological criteria for designing technologies (Talamo et al., 2011, 2015, 2017; Giorgi et al., 2013), to capture the complexity and contingency of real-life actions in specific situations.

Therefore, research on human/AI integration could benefit by considering some reflections from Cultural Psychology and more specifically from scholars by Activity Theory (AT) (Leont'ev, 1974, 1978; Engeström, 1987, 2000) who focus on three central concepts in analyzing the relationship between persons and technologies:

An Asymmetrical Interaction Between the Subject and the Object

Activity theory conceives human activity as a form of doing, performed by a subject and directed to an object, whose outcome will satisfy the needs of the subject. This interaction between the subject and the object is not a symmetrical relationship between two components of a system, since it is initiated and executed by the subject to meet its needs (Pickering, 1993, 1995). AI, for example, follows a program written by an IT developer who wants to respond to a need: technology, in fact, only has "the ability to act but not the need to act" (Kaptelinin and Nardi, 2006, p.33).

Intentionality of Human Beings

Agency, "the ability to act in the sense of producing effects according to an intention" (Kaptelinin and Nardi, 2006, p.33), is another crucial concept covered by socio-cognitive theories. For Leont'ev (1974, 1978), the primary type of agency is that of individual human subjects because it is closely related to the concept of human intentionality (Stetsenko and Arievitch, 2004). According to AT, intentionality is considered as a property of sole individual subjects. As Rose et al. (2005) observed, humans have "self-awareness, social awareness, interpretation, intentionality, and the attribution of agency to others," which are not available to non-living things, such as technological systems.

Mediation of Tools

Finally, the above-mentioned asymmetrical interaction between the subject and the object can be mediated by a tool, a physical artifact, or an intangible tool (e.g., ideas and procedures), which allows the subject to reach the final goal (Leont'ev, 1974, 1978). For example, technological tools, as activity mediators, can facilitate the interaction that allows the subject to achieve his goals, but they can also create boundaries because of the way in which the technology is implemented in those specific tools (Kuutti, 1996). Mediation of tools can also support the creation of interobjectivity among team members (Talamo and Pozzi, 2011).

PRECEDING AI DEVELOPMENT IN FINANCIAL SYSTEMS: MODELING AND INTEGRATING MULTI-ACTOR DM PROCESSES

As previously illustrated, while the contribution of AI in the financial sector requires the automation of DM processes to collect and analyze information, standardize investment criteria, and automate customer interaction services, Cultural Psychology ascribes to humans the primacy in DM processes, supporting the intentionality of human beings and the mediation role of tools. Hence, in order to enable functional human/AI integration and to guarantee proper functioning of AI-based systems, it is necessary to study in depth the DM model in the field.

Considering this, we propose a possible set of qualitative tools for the design of an AI-based financial DM support system. The tools we chose can be divided into two categories according to their objectives:

> **Tools to produce knowledge** (e.g., narrative interviews, maieutic clinical dialogs): to fully understand DM processes of all the actors involved in the financial field before AI implementation.
> **Tools to model processes** (e.g., user journey map, activity diagram): to model both DM processes of the provider and the user in order to create a bridge that can offer efficacy and efficiency to the provider and satisfaction of needs to the users.

To show how these qualitative tools can help in modeling DM processes, we refer to a case study on which we applied the

methods for the design of an AI platform to support decision-makers from a VCO. The original explicit request by VCO managers was to help information technology (IT) developers in modeling the AI system by tracing management DM, focusing on a specific DM step: comprehend which criteria and sub-criteria would "filter" the fund-seekers into the funnel flow in order to be judged for funding opportunities. The main issue was to figure out which aspects of human DM processes could be automated, and which human DM processes could be supported but not substituted.

To this aim, on one hand, we explored expected goals of VCO enhancing its awareness; on the other, we studied the fund-seekers to model their decision flow.

Enhancing VCO Awareness: Strategic Organizational Counseling

In most cases, when dealing with introducing AI in financial DM, attention is placed on modeling individual DM processes to be replicated and automated.

Nevertheless, we faced a lack of methods for modeling organizational DM in the financial field, not just in terms of formal declaration, rather with a necessary focus on shared implicit managerial criteria for providing funding. The method we implemented, SOC, consists of dialogical sessions with different managers guided by a psychologist implementing maieutic clinical techniques oriented to make implicit criteria arise in explicit talk. This tool made it possible to:

> Identify and model DM processes at an organizational layer (transversal to different managers).
> Differentiate the potential value proposition by the organization to different target among fund-seekers.
> Model scouting criteria on fund-seekers to orient financial decisions.

The result of this activity is twofold: on one side, it produced increased awareness in management on the complexity of DM processes they have to deal with and highlighted the need to share even implicit criteria used by different managers in different contextual circumstances; on the other, it produced descriptive charts of DM flows to be implemented in the platform. This process made it possible to align the system development team by highlighting two distinct roles of AI as potential support of

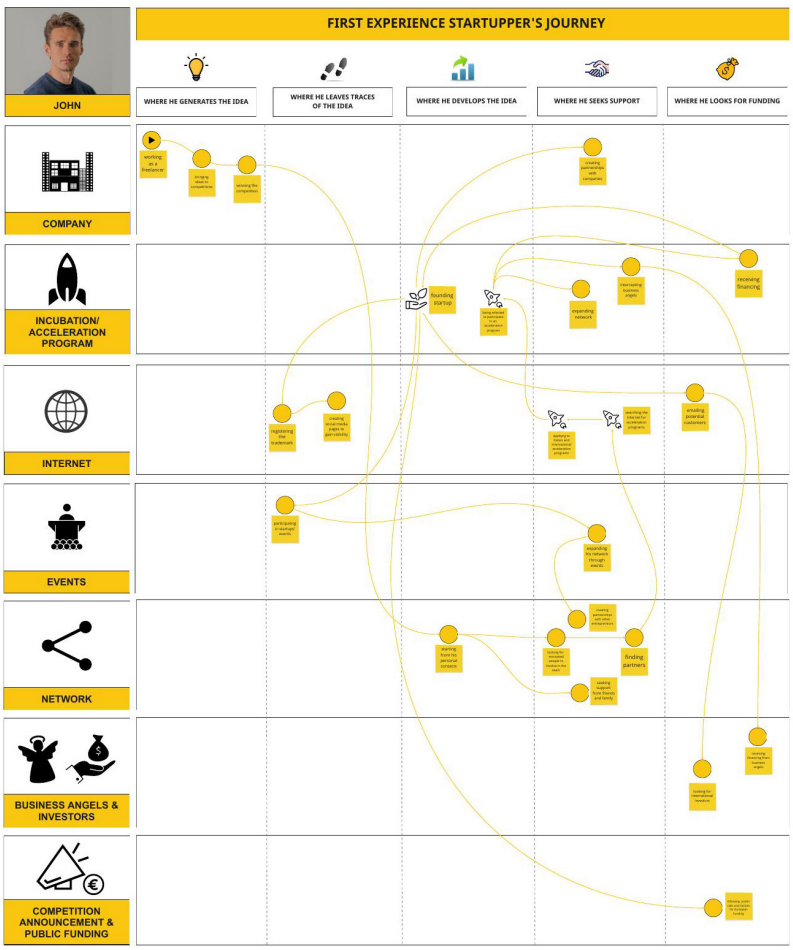

FIGURE 1 | User journey map.

the organizational DM: first, AI would automatically analyze Big Data from the market to suggest which trends should be preferentially funded; and second, AI could support managers in DM by signaling which inputs by fund-seekers would fit better with goals of the VCO. From this process, it was possible to shape ordered weighted averaging (OWA) operators (Merigò and Gil-Lafuente, 2010), to support VCO defining cases in which fund-seekers fit with funding criteria.

Exploring Decision Flow of Fund-Seekers

As Suchman (1987) demonstrated, human behavior in complex DM tasks is often complicated and can sometimes look chaotic. Nevertheless, once we observe it using descriptive tools from anthropology and psychology, we find more easily rationales that explain those behaviors in terms of personal and organizational contextualized objectives that actors are pursuing. For these reasons, design and usability practitioners elaborated over time different sets of qualitative methods for collecting valuable data on user behaviors (Cooper et al., 2007; Talamo et al., 2016; Recupero et al., 2018). The method we chose for the user research is the narrative interviews (Atkinson, 2002). This tool allowed us to collect perspectives of fund-seekers and the meanings they attribute to different steps of financial decision flow in terms of feelings, cognitions, representations of gain, and pain.

Modeling Activities of Fund-Seekers and DM Processes

Data collected were then employed to model activities of fund-seekers and DM processes. The tools we used belong to the Design Thinking approach, and some of these have been customized *ad hoc* to meet the scope of the VCO management. A tool that proved to be crucial in this process was the user journey map (Stickdorn and Schneider, 2011). This tool, configured as an oriented graph, provides a vivid, concise, structured, and precise visualization of the user experience, according to the decision flow of fund-seekers. It also enabled IT developers to understand the contexts and channels through which the platform could intercept fund-seekers and the moments and the kind of operations in which AI contribution could be most appropriate (see **Figure 1**).

Bridging Funders and Fund-Seekers

A crucial tool for creating a bridge between DM processes of fund-seekers and VCO was the activity diagram (Young, 2008). Consistently with AT, this tool, by structuring *activities, actions,* and *operations* of fund-seekers (Kaptelinin and Nardi, 2006) and matching them with services of VCO, proved to be very useful in identifying problems, developing potential bridging solutions, and recognizing spaces for innovation to create *ad hoc* AI-supported services (see **Figure 2**).

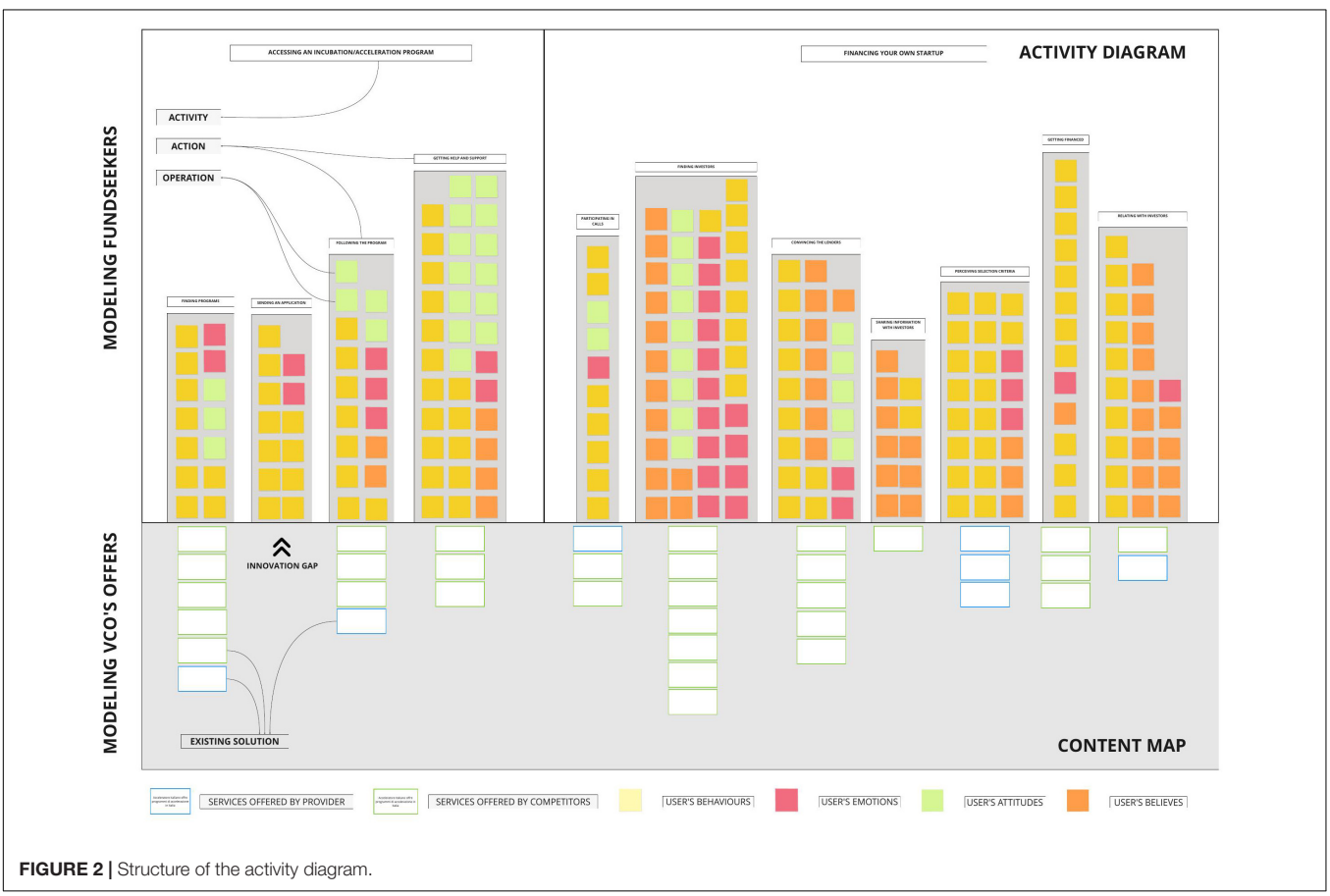

FIGURE 2 | Structure of the activity diagram.

As shown in **Figure 2**, the critical added value of modeling prospective users in the activity diagram relies on the potential comparison that the tool offers to match existing solutions (by the funding organization and its competitors). The content map section of this tool indicates some innovation gaps between actual financial services and activities and needs of fund-seekers, which are still not satisfied. Therefore, this modeling method can foster in funding organizations the capacity of creating innovative services to be implemented in AI-based systems.

DISCUSSION

Since DM in the financial field is a complex multilayer and multi-actor process, we propose a specific sequence of action-research activities aimed at modeling specific phases of DM processes by different actors:

> **Enhancing organization's DM awareness:** This step aimed at producing an increased awareness in management on their own intentions and funding criteria to finalize the different ways AI will support their decision-making.
> **Exploring fund-seekers:** This step aimed at studying the potential fund-seekers and their psychological world to collect data on which the modeling activity can be based.
> **Modeling activities of fund-seekers and DM processes:** This step led to a full-fledged view of the fund-seekers. The collected data were beneficial to the developers of systems and the VCO, providing insights about the contexts and the channels through which the platform could intercept fund-seekers.
> **Bridging funders and fund-seekers:** This last step, matching DM flow of fund-seekers with services of VCO, proved to be very useful in identifying problems, developing potential bridging solutions, and recognizing new spaces for innovation.

As the financial field is a promising multi-actor research area, the contribution of sociocultural approach from psychology and proposed methods can play a crucial role. In fact, Design Thinking combined with maieutic techniques, typical of expertise of psychologists, fosters modeling the complexity of DM systems emerging from different actors around funding decisions. Within the development team, the psychologist then becomes a mediator between IT developers and the VCO for which the system is developed.

Implications of this case study suggest that human/AI integration in the financial field can be successfully implemented by developing systems where AI can be conceived in two distinct functions: (a) automation/augmentation: treating Big Data from the market defined by VCO management; and (b) human/AI integration: creating OWA-based alert systems that support managers in taking decisions coherently with criteria of VCO.

Finally, we argue that, to achieve effective results in the design of complex IT systems that use AI in DM, technology development, albeit providing an enormous contribution, cannot disregard a deep comprehension of real practices by human actors. Therefore, as Kelly (2012) says: "*This is not a race against machines this is a race with machines.*"

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AT conceptualized the ideas presented in the article, defined the theoretical framework, and supervised the whole process. SM and CT wrote a first draft and helped to edit the manuscript. All authors contributed to revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

This research was possible thanks to a broader project promoted by Archangel Adventure in which the IDeACT laboratory was involved by Teleconsys.

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

Atkinson, R. (2002). *L'intervista Narrativa. Raccontare la Storia di sé Nella Ricerca Formativa, Organizzativa e Sociale*. Milano: Raffaello Cortina.

Bayrak, A. E., McComb, C., Cagan, J., and Kotovsky, K. (2021). A strategic decision-making architecture toward hybrid teams for dynamic competitive problems. *Dec. Support Syst.* 144:113490. doi: 10.1016/j.dss.2020.113490

Brynjolfsson, E., and McAfee, A. (2012). Winning the race with ever-smarter machines. *MIT Sloan Manag. Rev.* 53, 53–60.

Cellan-Jones, R. (2014). *Stephen Hawking Warns Artificial Intelligence Could End Mankind*. London: BBC Interview.

Choo, C. W. (2008). Organizational disasters: why they happen and how they may be prevented. *Manag. Dec.* 46, 32–45. doi: 10.1108/00251740810846725

Cooper, A., Reimann, R., and Cronin, D. (2007). *About Face 3 - The Essentials of Interaction Design*. Indianapolis, IND: Wiley Publishing, Inc.

De Bondt, W., and Thaler, R. H. (1995). "Financial decision-making in markets and firms: a behavioral perspective," in *Handbooks in OR & MS*, ed. R. Jarrow (Cambridge, MA: National Bureau of Economic Research, Inc).

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *Int. J. Inform. Manag.* 48, 63–71. doi: 10.1016/j.ijinfomgt.2019.01.021

Engeström, Y. (1987). *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Helsinki: Orienta-Konsultit.

Engeström, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics* 43, 960–974. doi: 10.1080/001401300409143

Galiano, A., Leogrande, A., Massari, S. F., and Massaro, A. (2019). I processi automatici di decisione: profili critici sui modelli di analisi e impatti nella relazione con i diritti individuali. *Rivista Italiana di Informatica e Diritto* 2, 41–61. doi: 10.32091/RIID0010

Gaudin, T. (1978). *L'écoute des Silences*. Paris: Union Généraled'Editions.

Giorgi, S., Ceriani, M., Bottoni, P., Talamo, A., and Ruggiero, S. (2013). Keeping "InTOUCH": an ongoing co-design project to share memories, skills and demands through an interactive table. *Lecture Notes Comput. Sci.* 7946, 633–640. doi: 10.1007/978-3-642-39062-3_43

Guszcza, J., Lewis, H., and Greenwood, P. E. (2017). Cognitive collaboration: why humans and computers think better together. *Deloitte Rev. Issue* 20, 7–29.

Haesevoets, T., De Cremer, D., Dierckx, K., and Van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Comp. Hum. Behav.* 119:106730. doi: 10.1016/j.chb.2021.106730

Heath, C., and Luff, P. (2000). *Acting with Technology*. Cambridge: Cambridge University Press.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus. Horizons* 61, 577–586. doi: 10.1016/j.bushor.2018.03.007

Kaptelinin, V., and Nardi, B. A. (2006). *Acting with Technology: Activity Theory and Interaction Design*. Cambridge, MA: MIT press.

Kelly, K. (2012). *Better than human: Why robots will and must take our jobs*. Germany: Wired.

Kuutti, K. (1996). "Activity theory as a potential framework for human-computer interaction research," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, ed. B. Nardi (Boston, MA: MIT Press), 17–44.

Leont'ev, A. (1978). *Activity, Consciousness, and Personality*. Englewood Cliffs, N.J: Prentice-Hall.

Leont'ev, A. N. (1974). The problem of activity in psychology. *Sov. psychol.* 13, 4–33. doi: 10.2753/RPO1061-040513024

Lepri, B., Oliver, N., and Pentland, A. (2021). Ethical machines: the human-centric use of artificial intelligence. *iScience* 24:102249. doi: 10.1016/j.isci 2021.102249

Merigò, J. M., and Gil-Lafuente, A. M. (2010). New decision making-techniques and their application in the selection of financial products. *Inform. Sci.* 180, 2085–2094. doi: 10.1016/j.ins.2010.01.028

Miller, S. (2018). AI: augmentation, more so than automation. *Asian Manag Insights* 5, 1–20.

Pickering, A. (1993). The mangle of practice: agency and emergence in the sociology of science. *Am. J. Sociol.* 99, 559–589. doi: 10.1086/230316

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Postman, N. (1993). *Technopoly: The Surrender of Culture to Technology*. New York, NY: Vintage.

Rawlinson, K. (2015). *Microsoft's Bill Gates Insists AI is a Threat*. London: BBC News.

Recupero, A., Triberti, S., Modesti, C., and Talamo, A. (2018). Mixed reality for cross-cultural integration: using positive technology to share experiences and promote communication. *Front. Psychol.* 9:1223. doi: 10.3389/fpsyg.2018.01223

Rose, J., Jones, M., and Truex, D. (2005). Socio-theoretic accounts of IS: the problem of agency. *Scand. J. Inform. Syst.* 17, 133–152.

Shanmuganathan, M. (2020). Behavioural finance in an era of artificial intelligence: longitudinal case study of robo-advisors in investment decisions. *J. Behav. Exp. Finance* 27:100297. doi: 10.1016/j.jbef.2020.100297

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Human-Computer Stud.* 146:102551. doi: 10.1016/j.ijhcs.2020.102551

Socea, A. D. (2012). Managerial decision-making and financial accounting information. *Procedia – Soc. Behav Sci.* 58, 47–55. doi: 10.1016/j.sbspro.2012.09.977

Stetsenko, A., and Arievitch, I. M. (2004). The self in cultural-historical activity theory. *Theory Psychol.* 14, 475–503. doi: 10.1177/0959354304044921

Stickdorn, M., and Schneider, J. (2011). *This is Service Design Thinking. Basic - Tools - Cases*. Amsterdam: BIS Publisher.

Suchman, L. (1987). *Plans and Situated Actions: the Problem of Humane-machine Communication*. Cambridge: Cambridge University Press.

Talamo, A., Camilli, M., Di Lucchio, L., and Ventura, S. (2017). Information from the past: how elderly people orchestrate presences, memories and technologies at home. *Univ. Access. Inf. Soc.* 16, 739–753. doi: 10.1007/s10209-016-0508-6

Talamo, A., Giorgi, S., and Mellini, B. (2011). "Designing technologies for ageing: is simplicity always a leading criterion?," in *Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity*, (New York, NY: ACM), 33–36.

Talamo, A., Mellini, B., Ventura, S., and Recupero, A. (2015). "Studying practices to inform design: organizational issues and local artifacts," in *Designing Technology, Work, Organizations and Vice Versa*, eds A. Bruni, L. L. Parolin, and C. Schubert (Wilmington: Vernon Press), 71–113.

Talamo, A., and Pozzi, S. (2011). The tension between dialogicality and interobjectivity in cooperative activities. *Cult. Psychol.* 17, 302–318. doi: 10.1177/1354067x11408131

Talamo, A., Recupero, A., Mellini, B., and Ventura, S. (2016). Teachers as designers of GBL scenarios: fostering creativity in the educational settings. *Interact. Des. Architecture(s) J.* 29, 10–23.

Triberti, S., Durosini, I., and Pravettoni, G. (2020). A "Third Wheel" effect in health decision making involving artificial entities: a psychological perspective. *Front. Public Health* 8:117. doi: 10.3389/fpubh.2020.00117

Wilson, J., and Daugherty, P. R. (2018). Collaborative intelligence humans and Al are joining forces. *Harvard Bus. Rev.* 96, 115–123.

Young, I. (2008). *Mental Models. Aligning Design Strategy with Human Behavior*. New York, NY: Rosenfeld Media.

Zheng, X.-L., Zhu, M.-Y., Li, Q.-B., Chen, C.-C., and Tan, Y.-C. (2019). FinBrain: when finance meets AI 2.0. *Front. Inform. Technol. Electron. Eng.* 20:914–924. doi: 10.1631/FITEE.1700822

Conflict of Interest: This article is based on a use case which has been part of a broader project based on the idea of Archangel AdVenture, a seed capital investment firm based in Italy (https://www.archangeladventure.it), of adopting AI to leverage open-source intelligence techniques applied to technology scouting for the purpose of investing in new ventures. The authors were involved by Teleconsys as experts in User-centered design research and were free to decide how to conduct the research.

Some key persons from funders' organizations were interviewed (as described in the paper) during the SOC collection of data as research subjects. The design of research procedures and tools, data analysis and interpretation, and the writing of this article were made by the authors. The authors declare to have informed the funders about the decision to submit the article for scientific publication.

frontiers in Computer Science

HYPOTHESIS AND THEORY
published: 12 August 2021
doi: 10.3389/fcomp.2021.696682

Adaptation Mechanisms in Human–Agent Interaction: Effects on User's Impressions and Engagement

Beatrice Biancardi[1]*, *Soumia Dermouche*[2] *and Catherine Pelachaud*[2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France, [2]CNRS-ISIR, Sorbonne University, Paris, France

Adaptation is a key mechanism in human–human interaction. In our work, we aim at endowing embodied conversational agents with the ability to adapt their behavior when interacting with a human interlocutor. With the goal to better understand what the main challenges concerning adaptive agents are, we investigated the effects on the user's experience of three adaptation models for a virtual agent. The adaptation mechanisms performed by the agent take into account the user's reaction and learn how to adapt on the fly during the interaction. The agent's adaptation is realized at several levels (i.e., at the behavioral, conversational, and signal levels) and focuses on improving the user's experience along different dimensions (i.e., the user's impressions and engagement). In our first two studies, we aim to learn the agent's multimodal behaviors and conversational strategies to dynamically optimize the user's engagement and impressions of the agent, by taking them as input during the learning process. In our third study, our model takes both the user's and the agent's past behavior as input and predicts the agent's next behavior. Our adaptation models have been evaluated through experimental studies sharing the same interacting scenario, with the agent playing the role of a virtual museum guide. These studies showed the impact of the adaptation mechanisms on the user's experience of the interaction and their perception of the agent. Interacting with an adaptive agent vs. a nonadaptive agent tended to be more positively perceived. Finally, the effects of people's *a priori* about virtual agents found in our studies highlight the importance of taking into account the user's expectancies in human–agent interaction.

Keywords: human–agent interaction, adaptation mechanisms, engagement, impressions, embodied conversational agent (ECA)

OPEN ACCESS

Edited by:
Stefano Triberti,
University of Milan, Italy

Reviewed by:
Sandra Cano,
Pontificia Universidad Católica de Valparaiso, Chile
Benjamin Lok,
University of Florida, United States

***Correspondence:**
Beatrice Biancardi
beatrice.biancardi@telecom-paris.fr

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *17 April 2021*
Accepted: *07 July 2021*
Published: *12 August 2021*

Citation:
Biancardi B, Dermouche S and Pelachaud C (2021) Adaptation Mechanisms in Human–Agent Interaction: Effects on User's Impressions and Engagement. Front. Comput. Sci. 3:696682. doi: 10.3389/fcomp.2021.696682

1 INTRODUCTION

During an interaction, we communicate through multiple behaviors. Not only speech but also our facial expressions, gestures, gaze direction, body orientation, etc. participate in the message being communicated (Argyle, 1972). Both interactants are active participants in an interaction and adapt their behaviors to each other. This adaptation arises on several levels: we align ourselves linguistically (vocabulary, syntax, and level of formality), but we also adapt our nonverbal behaviors (e.g., we respond to the smile of our interlocutor, and we imitate their posture and their gestural expressiveness), our conversational strategies (e.g., to be perceived as warmer or more competent), etc. (Burgoon et al., 2007). This multilevel adaptation can have several functions, such as reinforcing engagement in the interaction, emphasizing our relationship with others, showing empathy, and managing the impressions we give to others (Lakin and Chartrand, 2003;

Gueguen et al., 2009; Fischer-Lokou et al., 2011). The choice of verbal and nonverbal behaviors and their temporal realization are markers of adaptation.

Embodied conversational agents (ECAs) are virtual entities with a humanlike appearance that are endowed with communicative and emotional capabilities (Cassell et al., 2000). They can display a wide range of multimodal expressions to be active participants in the interaction with their human interlocutors. They have been deployed in various human–machine interactions where they can act as a tutor (Mills et al., 2019), health support (Lisetti et al., 2013; Rizzo et al., 2016; Zhang et al., 2017), a companion (Sidner et al., 2018), a museum guide (Kopp et al., 2005; Swartout et al., 2010), etc. Studies have reported that ECAs are able to take into account their human interlocutors and show empathy (Paiva et al., 2017), display backchannels (Bevacqua et al., 2008), and build rapport (Huang et al., 2011; Zhao et al., 2016). Given its relevance in human–human interaction, adaptation could be exploited to improve natural interactions with ECAs. It thus seems important to investigate whether an agent adapting to the user's behaviors could provoke similar positive outcomes in the interaction.

The majority of works in this context developed models learnt from existing databases of human–human interaction and did not consider the dynamics of adaptation mechanisms during an interaction. We are interested in developing an ECA that exploits how the interaction is currently going and is able to learn in real time what the best adaption mechanism for the interaction is.

In this article, we report three studies where an ECA adapts its behaviors by taking into account the user's reaction and by learning how to adapt on the fly during the interaction.

The goal of the different studies is to answer two broad research questions:

> "*Does adapting an ECA's behavior enhance user's experience during interaction*?"
>
> "*How does an ECA which adapts its behavior in real-time influence the user's perception of the agent?*"

A user's experience can involve many factors and can be measured by different dimensions, such as the user's engagement and the user's impressions about the ECA (Burgoon et al., 2007). In our three studies that we report in this article, we implemented three independent models where the agent's adaptation is realized at several levels and focuses on improving the user's experience along different dimensions as follows:

1) the agent's adaptation at a behavioral level: the ECA adapts its behaviors (e.g., gestures, arm rest poses, and smiles) in order to maximize the user's impressions about the agent's warmth or competence, the two fundamental dimensions of social cognition (Fiske et al., 2007). This model is described in **Section 7**;
2) the agent's adaptation at a conversational level: the ECA adapts its communicative strategies to elicit different levels of warmth and competence, in order to maximize the user's engagement. This model is described in **Section 8**; and
3) the agent's adaptation at a signal level: the ECA adapts its head and eye rotation and lip corner movement in function of the user's signals in order to maximize the user's engagement. This model is described in **Section 9**.

Each adaptation mechanism has been implemented in the same architecture that allows an ECA to adapt to the nonverbal behaviors of the user during the interaction. This architecture includes a multimodal analysis of the user's behavior using the Eyesweb platform (Camurri et al., 2004), a dialogue manager (Flipper (van Waterschoot et al., 2018)), and the ECA GRETA (Pecune et al., 2014). The architecture has been adapted to each model and evaluated through experimental studies. The ECA played the role of a virtual guide at the Science Museum of Paris. The scenario used in all the evaluation studies is described in **Section 6**.

Even though these three models have been implemented in the same architecture and tested on the same scenario, they have not been developed in order to do comparative studies. The main goal of this paper is to frame them in the same theoretical framework (see **Section 2**) and have insights into each of these different adaptation mechanisms to better understand what the main challenges concerning these models are and to suggest further improvements for an adaptation system working on multiple levels.

This article is organized as follows: in **Section 2**, we review the main theories about adaptation which our work relies on, in particular Burgoon and others' work; in **Section 3**, we present an overview of existing models that focus on adapting the ECA's behavior according to the user's behavior; in **Section 4**, we specify the dimensions we focused on in our adaptation models; in **Section 5**, we present the general architecture we conceived to endow our ECA with the capability of adapting its behavior to the user's reactions in real time; in **Section 6**, we describe the scenario we conceived to test the different adaptation models; in **Sections 7–9**, we report the implementation and evaluation of each of the three models. More details about them can be found in our previous articles (Biancardi et al., 2019b; Biancardi et al., 2019a; Dermouche and Pelachaud, 2019). We finally discuss the results of our work and possible improvements in **Sections 10**, **11**, respectively.

2 BACKGROUND

Adaptation is an essential feature of interpersonal relationships (Cappella, 1991). During an effective communication, people adapt their interaction patterns to one another's (e.g., dancers synchronize their movements and people adapt their conversational style in a conversation). These patterns contribute to defining and maintaining our interpersonal relationships, by facilitating smooth communication, fostering attraction, reinforcing identification with an in-group, and increasing rapport between communicators (Bernieri et al.,

1988; Giles et al., 1991; Chartrand and Bargh, 1999; Gallois et al., 2005).

There exist several adaptation patterns, differing according to their behavior type (e.g., the modality, the similarity to the other interlocutor's behavior, etc.), their level of consciousness, whether they are well decoded by the other interlocutor, and their effect on the interaction (Toma, 2014). Cappella and others (Cappella, 1981) considered an additional characteristic, that is, adaptation can be asymmetrical (unilateral), when only one partner adapts to the other, or symmetrical (mutual), like in the case of interaction synchrony.

In line with these criteria, in some examples of adaptation, people's behaviors become more similar to one another's. This type of adaptation is often unconscious and reflects reciprocity or convergence. According to Gouldner (Gouldner, 1960), reciprocity is motivated by the need to maintain harmonious and stable relations. It is contingent (i.e., one person's behaviors are dependent upon the other's) and transactional (i.e., it is part of an exchange process between two people).

In other cases, adaptation can include complementarity or divergence; this occurs when the behavior of one person differs from but complements that of the other person.

Several theories focus on one or more specific characteristics of adaptation and highlight different factors that drive people's behaviors. They can be divided into four main classes according to the perspective they follow to explain adaptation.

The first class of theories includes biologically based models (e.g., (Condon and Ogston, 1971), (Bernieri et al., 1988)). These theories state that individuals exhibit similar patterns to one another. These adaptation patterns have an innate basis, as they are related to satisfaction of basic needs like bonding, safety, and social organization. Their innate bases make them universal and involuntary, but they can be influenced by environmental and social factors as well.

Following a different perspective, arousal-based and affect-based models (e.g., (Argyle and Dean, 1965), (Altman et al., 1981), (Cappella and Greene, 1982)) support the role of internal emotional and arousal states as driving factors of people's behaviors. These states determine approaching or avoiding behaviors. This group of theories explains the balance between compensation and reciprocity.

Social-norm models (e.g., (Gouldner, 1960), (Dindia, 1988)) do not consider the role of physiological or psychological factors but argue for the importance of social phenomena as guiding forces. These social phenomena are, for example, the in-group or out-group status of the interactants, their motivation to identify with one another, and their level of affiliation or social distance.

The last class of theories includes communication- and cognition-based models (e.g., (Andersen, 1985), (Hale and Burgoon, 1984)), which focus on the communicative purposes of the interactants and on the meaning that the behavioral patterns convey. While adaption happens mainly unconsciously, it may happen that the process of interpersonal adaptation may be strategic and conscious (Giles et al., 1991; Gallois et al., 2005).

The majority of these theories have been studied by Burgoon and others (Burgoon et al., 2007). In particular, they examined fifteen previous models and considered the most important conclusions from the previous empirical research. From this analysis, they came out with a broader theory, the interaction adaptation theory (IAT). This theory states that we alter our behavior in response to the behavior of another person in conversations (Infante et al., 2010). IAT takes into account the complexities of interpersonal interactions by considering people's needs, expectations, desires, and goals as precursors of their degree and form of adaptation. IAT is a communication theory made of multiple theories, which focuses on the sender's and the receiver's process and patterns.

Three main interrelated factors contribute to IAT. Requirements (Rs) refer to the individual beliefs about what is necessary in order to have a successful interaction. Rs are mainly driven by biological factors, such as survival, safety, and affiliation. Expectations (Es) refer to what people expect from the others based on social norms or knowledge coming from previous interactions. Es are mainly influenced by social factors. Finally, desires (Ds) refer to the individual's goals and preferences about what to get out of the interaction. Ds are mainly influenced by person-specific factors, such as temperament or cultural norms. These three factors are used to predict an individual's interactional position (IP). This variable derives from the combination of Rs, Es, and Ds and represents the individual's behavioral predisposition that will influence how an interaction will work. The IP would not necessarily correspond to the partner's actual behavior performed in the interaction (A). The relation between the IP and A will determine the type of adaptation during the interaction. For example, when the IP and A almost match, IAT predicts behavioral patterns such as reciprocity and convergence. When A is more negatively valenced than the IP, the model predicts compensation and avoiding behaviors.

In the work presented in this article, we rely on Burgoon's IAT. Indeed, our adapting ECA has an interactional position (IP), resulting from its desires (Ds) and expectations (Es). In particular, the agent's desire (D) is to maximize the user's experience, and its expectations (Es) are about the user's reactions to its behaviors. In our different models of adaptation mechanisms, the agent's desire (D) refers either to giving the best impression to the user or to maximizing the user's engagement (see **Section 4**). Consequently, the expectations (Es) refer to the user's reaction reflecting their impressions or engagement level in response to the agent's behavior. The behavior that will be performed by the ECA depends on the relation between the agent's IP and the user's reaction (actual behavior A).

In addition, we explore different ways in which the ECA can adapt to the user's reactions. On one hand, we focus on theories that consider adaptive behaviors more broadly than a mere matching, that is, adaptation as responding in appropriate ways to a partner. The ECA will choose its behaviors according to the effect they have on the user's experience (see **Section 7**). In Study 2 (see **Section 8**), our adaptive agent follows the same perspective but by adapting its communicative strategies. On the other hand, we try to simulate a more unconscious and automatic process working at a motoric level; the agent adapts at a signal level (see Study 3, **Section 9**).

3 STATE OF THE ART

In this section, we present an overview of existing models that focused on adapting ECAs' behavior according to the user's behavior in order to enhance the interaction and the user's experience along different dimensions such as engagement, rapport, interest, liking, etc. These existing models predicted and generated different forms of adaptation, such as backchannels, mimicry, and voice adaptation, and were applied on virtual agents or robots.

Several works were interested in understanding the impact of adaptation on the user's engagement and rapport building. Some of them did so through the production of backchannels. Huang et al. (2010) developed an ECA that was able to produce backchannels to reinforce the building of rapport with its human interlocutor. The authors used conditional random fields (CRFs) (Lafferty et al., 2001) to automatically learn when listeners produce visual backchannels. The prediction was based on three features: prosody (e.g., pause and pitch), lexical (spoken words), and gaze. Using this model, the ECA was perceived as more natural; it also created more rapport with its interlocutor during the interaction. Schröder et al. (2015) developed a sensitive artificial listener that was able to produce backchannels. They developed a model that predicted when an ECA should display a backchannel and with which intention. The backchannel could be either a smile, nod, and vocalization or an imitation of a human's smile and head movement. Participants who interacted with an ECA displaying backchannels were more engaged than they were when no backchannels were shown.

Other works focused on modeling ECAs that were able to mimic their interlocutors' behaviors. Bailenson and Yee (2005) studied the social influence of mimicry during human–agent interaction (they referred to this as the chameleon effect). The ECA mimicked the user's head movements with a delay of up to 4 s. An ECA showing mimicry was perceived as more persuasive and more positive than an ECA showing no mimicry at all. Raffard et al. (2018) also studied the influence of ECAs mimicking their interlocutors' head and body posture with some delay (below 4 s). Participants with schizophrenia and healthy participants interacted with an ECA that either mimicked them or not. Both groups showed higher behavior synchronization and reported an increase in rapport in the mimicry condition. Another study involving mimicry was proposed by (Verberne et al., 2013) in order to evaluate if an ECA mimicking the user's head movements would be liked and trusted more than a non-mimicking one. This research question was investigated by running two experiments in which participants played a game involving drivers handing over the car control to the ECA. While results differed depending on the game, the authors found that liking and trust were higher for a mimicking ECA than for a non-mimicking one.

Reinforcement learning methods for optimizing the agent's behaviors according to the user's preference have been used in different works. For example, Liu et al. (2008) endowed a robot with the capacity to detect, in real time, the affective states (liking, anxiety, and engagement) of children with autism spectrum disorder and to adapt its behavior to the children's preferences of activities. The detection of children's affective states was done by exploiting their physiological signals. A large database of physiological signals was explored to find their interrelation with the affective states of the children. Then, an SVM-based recognizer was trained to match the children's affective state to a set of physiological features. Finally, the robot learned the activities that the children preferred to do at a moment based on the predicted liking level of the children using QV-learning (Wiering, 2005). The proposed model led to an increase in the reported liking level of the children toward the robot. Ritschel et al. (2017) studied the influence of the agent's personality on the user's engagement. They proposed a reinforcement learning model based on social signals for adapting the personality of a social robot to the user's engagement level. The user's engagement was estimated from their multimodal social signals such as gaze direction and posture. The robot adapted its linguistic style by generating utterances with different degrees of extroversion using a natural language generation approach. The robot that adapted its personality through its linguistic style increased the user's engagement, but the degree of the user's preference toward the robot depended on the ongoing task. Later on, the authors applied a similar approach to build a robot that adapts to the sense of humor of its human interlocutor (Weber et al., 2018).

Several works have been conducted in the domain of education where an agent, being physical as a robot or virtual as an ECA, adapted to the learner's behavior. These works reported that adaptation is generally linked with an increase in the learner's engagement and performance. For example, Gordon et al. (2016) developed a robot acting as a tutor for children learning a second language. To favor learning, the robot adapted its behaviors to optimize the level of the children's engagement, which was computed from their facial expressions. A reinforcement learning algorithm was applied to compute the robot's verbal and nonverbal behavior. Children showed higher engagement and learned more second-language words with the robot that adapted its behaviors to the children's facial expression than they did with the nonadaptive robot. Woolf et al. (2009) manually designed rules to adapt the facial expressions of a virtual tutor according to the student's affective state (e.g., frustrated, bored, or confused). For example, if the student was delighted and sad, respectively, the tutor might look pleased and sad, respectively. Results showed that when the virtual tutor adapted its facial expressions in response to the student's ones, the latter maintained higher levels of interest and reduced levels of boredom when interacting with the tutor.

Other works looked at adapting the activities undertaken by an agent during an interaction to enhance knowledge acquisition and reinforce engagement. In the study by (Ahmad et al., 2017), a robot playing games with children was able to perform three different types of adaptations, game-based, emotion-based, and memory-based, which relied, respectively, on the following: 1) the game state, 2) emotion detection from the child's facial expressions, and 3) face recognition mechanisms and remembering the child's performance. In the first category of adaptation, a decision-making mechanism was used to generate a supporting verbal and nonverbal behavior. For example, if the

child performed well, the robot said "Wow, you are playing extraordinary" and showed positive gestures such as a thumbs-up. The emotion-based adaptation mapped the child's emotions to a set of supportive dialogues. For example, when detecting the emotion of joy, the robot said, "You are looking happy, I think you are enjoying the game." For memory adaptation, the robot adapted its behavior after recognizing the child and retrieving the child's game history such as their game performance and results. Results highlighted that emotion-based adaptation resulted in the highest level of social engagement compared to memory-based adaptation. Game adaptation did not result in maintaining long-term social engagement. Coninx et al. (2016) proposed an adaptive robot that was able to change activities during an interaction with children suffering from diabetes. The aim of the robot was to reinforce the children's knowledge with regard to managing their disease and well-being. Three activities were designed to approach the diabetes-learning problem from different perspectives. Depending on the children's motivation, the robot switched between the three proposed activities. Adapting activities in the course of the interaction led to a high level of children's engagement toward the robot. Moreover, this approach seemed promising for setting up a long-term child–robot relationship.

In a task-oriented interaction, Hemminahaus and Kopp (2017) presented a model to adapt the social behavior of an assistive robot. The robot could predict when and how to guide the attention of the user, depending on the interaction contexts. The authors developed a model that mapped interactional functions such as motivating the user and guiding them onto low-level behaviors executable by the robot. The high-level functions were selected based on the interaction context and the attentive and emotional states of the user. Reinforcement learning was used to predict the mapping of these functions onto lower-level behaviors. The model was evaluated in a scenario in which a robot assisted the user in solving a memory game by guiding their attention to the target objects. Results showed that users were able to solve the game faster with the adaptive robot.

Other works focused on voice adaptation during social interaction. Voice adaptation is based on acoustic–prosodic entrainment that occurs when two interactants adapt their manner of speaking, such as their speaking rate, tone, or pitch, to each other's. Levitan (2013) found that voice adaptation improved spoken dialogue systems' performance and the user's satisfaction. Lubold et al. (2016) studied the effect of voice adaptation on social variables such as rapport and social presence. They found that social presence was significantly higher with a social voice-adaptive speech interface than with purely social dialogue.

In most previous works, the adaptation mechanisms that have been implemented measured their influence on the user's engagement through questionnaires. They did not include them as a factor of the adaptation mechanisms. In our first two studies reported in this article, we aimed to learn the agent's multimodal behaviors and conversational strategies to dynamically optimize the user's engagement and their impressions of the ECA, by taking them as input during the learning process.

Moreover, in most existing works, the agent's predicted behavior depended exclusively on the user's behavior and ignored the interaction loop between the ECA and the user. In our third study, we took into account this interaction loop, that is, our model takes as input both the user's and the agent's past behavior and predicts the agent's next behavior. Another novelty presented in our work is to include the agent's communicative intentions along with its adaptive behaviors.

4 DIMENSIONS OF STUDY

In our studies, we focused on adaptation in human–agent interaction by using the user's reactions as the input for the agent's adaptation. In particular, we took into account two main dimensions, which are the user's impressions of the ECA and the user's engagement during the interaction.

These two dimensions play an important role during human–agent interactions, as they influence the acceptability of the ECA by the user and the willingness to interact with it again (Bergmann et al., 2012; Bickmore et al., 2013; Cafaro et al., 2016). In order to engage the user, it is important that the ECA displays appropriate socio-emotional behaviors (Pelachaud, 2009). In our case, we were interested in whether and how the ECA could affect the user's engagement by managing the impressions it gave to them. In particular, we considered the user's impressions of the two main dimensions of social cognition, that is, warmth and competence (Fiske et al., 2007). Warmth includes traits like friendliness, trustworthiness, and sociability, while competence includes traits like intelligence, agency, and efficacy. In human–human interaction, several studies have showed the role of nonverbal behaviors in conveying different impressions of warmth and competence. In particular, communicative gestures, arm rest poses, and smiling behavior have been found to be associated with different degrees of warmth and/or competence (Duchenne, 1990; Cuddy et al., 2008; Maricchiolo et al., 2009; Biancardi et al., 2017a). In the context of human–agent interaction, we can control and adapt the nonverbal behaviors of the ECA during the flow of the interaction.

Following Burgoon's IAT theoretical model, our adapting ECA thus has the desire D to maintain the user's engagement (or impressions) during the interaction. Since the ECA aims to be perceived as a social entity by its human interlocutor, the agent's expectancy E is that adaptation can enhance the interaction experience. In our work, we are interested in whether adapting at a behavioral or conversational level (i.e., the agent's warmth and competence impressions) and/or at a low level (i.e., the agent's head and eye rotation and lip corner movement) could affect the user's engagement. Even though the impact of the agent's adaptation on the user's engagement has already been the object of much research (see **Section 3**), here we use the user's engagement as a real-time variable given as input for the agent's adaptation.

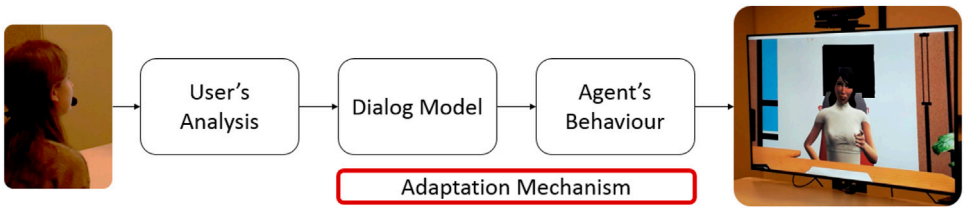

FIGURE 1 | System architecture: in the User's Analysis module, the user's nonverbal and verbal signals are extracted and interpreted and the user's reaction is sent to the Dialogue Model module, which computes the dialogue act to be communicated by the ECA. The Agent's Behavior module instantiates the dialogue act into multimodal behaviors to be displayed by the ECA. The Adaptation Mechanism module adapts the agent's behavior to the user's behavior. Its placement in the architecture depends on the specific adaptation mechanism that is implemented.

FIGURE 2 | Interaction space in the experiment room. The participants were sitting in front of the TV screen displaying the ECA. On the left, two screens separated the interaction space from the control space.

5 ARCHITECTURE

In this section, we present the architecture we conceived to endow the ECA with the capability of adapting its behavior to the user's reactions in real time. The architecture consists of several modules (see **Figure 1**). One module extracts information about the user's behaviors using a Kinect and a microphone. This information is interpreted in terms of speech (what the user has uttered) and the user's state (e.g., their engagement in the interaction). This interpreted information is sent to a dialogue manager that computes the communicative intentions of the ECA, that is, what it should say and how. Finally, the animation of the ECA is computed on the fly and played in real time. The agent's adaptation mechanisms are also taken into account when computing its verbal and nonverbal behaviors. The architecture is general enough to allow for the customization of its different modules according to the different adaptation mechanisms and goals of the agent.

In more detail, the four main parts of the architecture are as follows:

1) User's Analysis: the EyesWeb platform (Camurri et al., 2004) allows the extraction in real time of the following: 1) the user's nonverbal signals (e.g., head and trunk rotation), starting from the Kinect depth camera skeleton data; 2) the user's facial muscular activity (action units or AUs (Ekman et al., 2002)), by running the OpenFace framework (Baltrušaitis et al., 2016); 3) the user's gaze; and 4) the user's speech, by executing Microsoft Speech Platform[1].

   These low-level signals are processed using EyesWeb and other external tools, such as machine learning pretrained models (Dermouche and Pelachaud, 2019; Wang et al., 2019), to extract high-level features about the user, such as their level of engagement.

2) Dialogue Model: in this module, the dialogue manager Flipper (van Waterschoot et al., 2018) selects the dialogue act that the agent will perform and the communicative intention of the agent (i.e., how to perform that dialogue act).
3) Agent's Behavior: the agent's behavior generation is performed using GRETA, a software platform supporting the creation of socio-emotional embodied conversational agents (Pecune et al., 2014). The Agent's Behavior module is made of two main modules: the Behavior Planner receives the communicative intentions of the ECA from the Dialogue Model module as input and instantiates them into multimodal behaviors and the Behavior Realizer transforms the multimodal behaviors into facial and body animations to be displayed on a graphics screen.
4) Adaptation Mechanism: since the ECA can adapt its behaviors at different levels, the Adaptation Mechanism module is implemented in different parts of the architecture, according to the type of adaptation that the ECA performs. That is, the adaptation can affect the communicative intentions of the ECA, or it can occur during the behavior realization at the animation level. In the first two models presented in this article, the Adaptation Mechanism module is connected to the Dialogue Model module, while for the third model, it is connected to the Agent's Behavior module.

[1]https://www.microsoft.com/en-us/download/details.aspx?id=27225

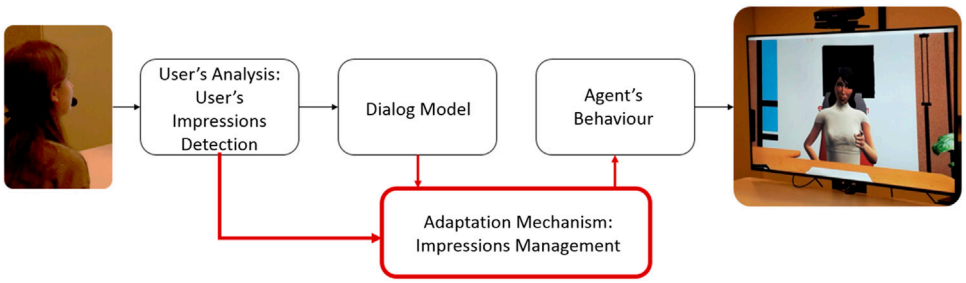

FIGURE 3 | Modified system architecture used in Study 1. In particular, the User's Analysis module contains the model to detect the user's impressions from facial signals. The Impressions Management module contains the Q-learning algorithm.

6 SCENARIO

Each type of adaptation has been investigated by running human–agent interaction experiments at the Science Museum of Paris. In the scenario conceived for these experiments, the ECA, called Alice, played the role of a virtual guide of the museum.

The experiment room included a questionnaire space, including a desk with a laptop and a chair; an interaction space, with a big TV screen displaying the ECA, a Kinect Two placed on the top of the TV screen, and a black tent behind the chair where the participant sat; and a control space, separated from the rest of the room by two screens, including a desk with the computer running the system architecture. The interaction space is shown in **Figure 2**.

The experiments were completed in three phases as follows:

1) before the interaction began, the participant sat at the questionnaire space, read and signed the consent form, and filled out the first questionnaire (NARS, see below). Then they moved to the interaction space, where the experimenter gave the last instructions [5 min];
2) during the interaction phase, the participant stayed right in front of the TV screen, between it and the black tent. They wore a headset and were free to interact with the ECA as they wanted. During this phase, the experimenter stayed in the control space, behind the screens [3 min]; and
3) after the interaction, the participant came back to the questionnaire space and filled out the last questionnaires about their perception of the ECA and of the interaction. After that, the experimenter proceeded with the debriefing [5 min].

Before the interaction with the ECA, we asked participants to fill out a questionnaire about their *a priori* about virtual characters (NARS); an adapted version of the NARS scale from the study by Nomura et al. (2006) was used. Items of the questionnaire included, for example, how much participants would feel relaxed talking with a virtual agent, or how much they would like the idea that virtual agents made judgments.

The interaction with the ECA lasted about 3 min. It included 26 steps. A step included one dialogue act played by the ECA and the participant's potential reaction/answer. The dialogue scenario was built so that the ECA drove the discussion. The virtual guide provided information on an exhibit that was currently happening in the museum. It also asked some questions about participants' preferences. Purposely, we limited the possibility for participants to take the lead in the conversation as we wanted to avoid any error due to automatic speech understanding. More details about the dialogue model can be found in the study by (Biancardi et al., 2019a).

7 STUDY 1: ADAPTATION OF AGENT'S BEHAVIORS

At this step, we aim to investigate adaptation at a high level, meant as convergence of the agent's behaviors according to the user's impressions of the ECA.

The goal of this first model is to make the ECA learn the verbal and nonverbal behaviors to be perceived as warm or competent by measuring and using the user's impressions as a reward.

7.1 Architecture

The general architecture described in **Section 5** has been modified in order to contain a module for the detection of the user's impressions and a specific set of verbal and nonverbal behaviors from which the ECA could choose.

The modified architecture of the system is depicted in **Figure 3**. In the following section, we give more details about the modified modules.

7.1.1 User's Analysis: User's Impression Detection

The user's impressions can be detected from their nonverbal behaviors, in particular, their facial expressions. The User's Analysis module is integrated with a User's Impression Detection module that takes as input a stream of the user's facial action units (AUs) (Ekman et al., 2002) and outputs the potential user's impressions about the level of warmth (or competence) of the ECA.

A trained multilayer perceptron (MLP) regression model is implemented in this module to detect the impressions formed by users about the ECA. The MLP model was previously trained using a corpus including face video recordings and continuous self-report annotations of warmth and competence given by

participants watching the videos of the NoXi database (Cafaro et al., 2017). The self-report annotations being considered separately, the MLP model was trained twice, one for warmth and one for competence. More details about this model can be found in the study by (Wang et al., 2019).

7.1.2 Adaptation Mechanism: Impression Management

In this model, the adaptation of the ECA concerns the impressions of warmth and competence given to the user. The inputs of the Adaptation Mechanism module are the dialogue act to be realized (coming from the Dialogue Model module) and the user's impression of the agent's warmth or competence (coming from the User's Analysis module). The output is a combination of behaviors to realize the dialogue act, chosen from a set of possible verbal and nonverbal behaviors to perform.

To be able to change the agent's behavior according to the detected participant's impressions, a machine learning algorithm is applied. We follow a reinforcement learning approach to learn which actions the ECA should take (here, verbal and nonverbal behaviors) in response to some events (here, the user's detected impressions). We rely on a Q-learning algorithm for this step. More details about it can be found in the study by (Biancardi et al., 2019b).

The set of verbal and nonverbal behaviors, from which the Q-learning algorithm selects a combination to send to the Behavior Planner of the Agent's Behavior module, includes the following:

- Type of gestures: the ECA could perform ideational (i.e., related to the content of the speech) or beat (i.e., marking speech rhythm, not related to the content of the speech) gestures or no gestures.
- Arm rest poses: in the absence of any kind of gesture, these rest poses could be performed by the ECA: akimbo (i.e., hands on the hips), arms crossed on the chest, arms along its body, or hands crossed on the table.
- Smiling: during the animation, the ECA could decide whether or not to perform smiling behavior, characterized by the activation of AU6 (cheek raiser) and AU12 (lip puller-up).
- Verbal behavior: the ECA could modify the use of you- and we-words, the level of formality of the language, and the length of the sentences. These features have been found to be related to different impressions of warmth and competence (Pennebaker, 2011; Callejas et al., 2014).

7.2 Experimental Design

The adaptation model described in **Subsection 7.1.2** has been evaluated by using the scenario described in **Section 6**. Here, we describe the experimental variables manipulated and measured during the experiment.

7.2.1 Independent Variable

The independent variable manipulated in this experiment, called Model, concerns the use of the adaptation model and includes three conditions:

- Warmth: when the ECA adapts its behaviors according to the user's impressions of the agent's warmth, with the goal to maximize these impressions;
- Competence: when the ECA adapts its behaviors according to the user's impressions of the agent's competence, with the goal to maximize these impressions; and
- Random: when the adaptation model is not exploited and the ECA randomly chooses its behavior, without considering the user's reactions.

7.2.2 Measures

The dependent variables measured after the interaction with the ECA are as follows:

- User's perception of the agent's warmth (*w*) and competence (*c*): participants were asked to rate their level of agreement about how well each adjective described the ECA (4 adjectives concerning warmth and four concerning competence, according to Aragonés et al. (2015)). Even though only one dimension was manipulated at a time, we measured the user's impressions about both of them in order to check whether the manipulation of one dimension can affect the impressions about the other (as already found in the literature (Rosenberg et al., 1968; Judd et al., 2005; Yzerbyt, 2005)).
- User's experience of the interaction (*exp*): participants were asked to rate their level of agreement about a list of items adapted from the study by (Bickmore et al., 2011).

7.2.3 Hypotheses

We hypothesized the following scenarios:

H1: when the ECA is in the Warmth condition, that is, when it adapts its behaviors according to the user's impressions of the agent's warmth, it will be perceived as warmer than it is in the Random condition;
H2: when the ECA is in the Competence condition, that is, when it adapts its behaviors according to the user's impressions of the agent's competence, it will be perceived as more competent than it is in the Random condition;
H3: when the agent ECA adapts its behaviors, that is, in either the Warmth or Competence conditions, this will improve the user's experience of the interaction, compared to that in the Random condition.

7.3 Analysis and Results

The visitors (24 women and 47 men) of the Carrefour Numérique of the Cité des sciences et de l'industrie of Paris were invited to take part in our experiment. 28% of them were in the range of 18–25°years old, 18% were in the range of 25–36, 28% were in the range of 36–45, 15% were in the range of 46–55, and 11% were over 55°years old. Participants were randomly assigned to each condition, with 25 participants assigned to the Warmth condition, 27 to the Competence condition, and 19 to the Random one.

We computed Cronbach's alphas on the scores of the four items about *w* and the four about *c*: good reliability was found for

TABLE 1 | Mean and standard deviation of *w* and *c* scores for each level of Model.

Model	Warmth	Competence
Warmth	3.48 ± 0.8	3.2 ± 0.75
Competence	3.51 ± 0.96	3.3 ± 0.69
Random	3.26 ± 0.93	2.76 ± 0.73

both ($\alpha = 0.85$ and $\alpha = 0.81$, respectively). Then, we computed the mean of these items in order to have one *w* score and one *c* score for each participant, and we used them for our analyses.

Since *NARS* scores got an acceptable degree of reliability ($\alpha = 0.69$), we computed the overall mean of these items for each participant and divided them into two groups, "high" and "low," according to whether they obtained a score higher than the overall mean or not, respectively. Participants were almost equally distributed into the two groups (35 in the "high" group and 36 in the "low" group). Chi-square tests for Model, age, and sex were run to verify that participants were equally distributed across these variables, too (all $p > 0.5$).

7.3.1 Warmth Scores

The *w* means were normally distributed (the Shapiro test's $p = 0.07$), and their variances were homogeneous (the Bartlett tests' *p*s for each variable were > 0.44). We run a 3 x 5 x 2 x 2 between-subjects ANOVA, with Model, age, sex, and NARS as factors.

No effects of age or sex were found. A main effect of NARS was found ($F(1, 32) = 4.23, p < 0.05$). A *post hoc* test specified that the group who got high scores in NARS gave higher ratings about the agent's *w* ($M = 3.65, SD = 0.84$) than the group who got low scores in NARS ($M = 3.24, SD = 0.96$).

Although we did not find any significant effect, *w* scores were, on average, higher in the Warmth and Competence conditions than in the Random condition. The mean and standard error of *w* scores are shown in **Table 1**.

7.3.2 Competence Scores

The *c* means were normally distributed (the Shapiro test's $p = 0.22$), and their variances were homogeneous (the Bartlett tests' *p*s for each variable were > 0.25). We run a 3 x 5 x 2 x 2 between-subjects ANOVA, with Model, age, sex, and NARS scores as factors.

We did not find any effect of age, sex, or NARS. A significant main effect of Model was found ($F(2, 32) = 3.22, p = 0.047, \eta 2 = 0.085$). In particular, *post hoc* tests revealed that participants in the Competence condition gave higher scores about the agent's *c* than participants in the Random condition ($M_C = 3.3, M_R = 2.76, p\text{-}adj = 0.05$).

7.3.3 User's Experience Scores

The *exp* items' means were not normally distributed, but their variances were homogeneous (the Bartlett tests' *p*s for each variable were > 0.17). We run nonparametric tests for each item and each variable.

Even if we did not find any statistically significant effect, on average, items' scores tended to be higher in the Warmth and Competence conditions than in the Random condition.

7.3.4 Performance of the Adaptation Model

The Q-learning algorithm ended up selecting (for each participant) one specific combination of verbal and nonverbal behaviors from the 84% ± 7 and 82% ± 7 of the interaction, for the Warmth and Competence conditions, respectively. In the Warmth condition, the rest pose Akimbo was the most selected one ($\chi 2 = 8.05, p < 0.01$), and we found a tendency to use Ideational gestures ($p > 0.05$). In the Competence condition, the Verbal Behavior aiming at eliciting low warmth and high competence (formal language, long sentences, and use of you-words) was the most selected one ($\chi 2 = 3.86, p < 0.01$).

7.4 Discussion

The results show that participants' ratings tended to be higher in the conditions in which the ECA used the adaptation model than when it selected its behavior randomly. In particular, the results indicate that we successfully manipulated the impression of competence when using our adaptive ECA. Indeed, higher competence was reported in the Competence condition than in the Random one. No *a priori* effect was found.

On the other hand, we found an *a priori* effect on warmth but no significant effect of our conditions (just a positive trend for both the Competence and Warmth conditions). People with high *a priori* about virtual agents gave higher ratings about the agent's warmth than people with low *a priori*.

We could hypothesize some explanations for these results. First, we did not get the effects of our experimental conditions on warmth ratings since people were more anchored into their *a priori*, and it was hard to change them. Indeed, people's expectancies have already been found to have an effect on the user's judgments about ECAs (Burgoon et al., 2016; Biancardi et al., 2017b; Weber et al., 2018). The fact that we found this effect only for warmth judgments could be related to the primacy of warmth judgments over competence (Wojciszke and Abele, 2008). Then, it could have been easier to elicit impressions of competence since we found no *a priori* effect on competence. This could be explained as follows: people might expect that it is easier to implement knowledge in an ECA rather than social behaviors.

The user's experience of the interaction was not affected by the agent's adaptation. During the debriefing, many participants expressed their disappointment about the agent's appearance, the quality of the voice synthesizer and the animation, described as "disturbing" and "creepy," and the limitations of the conversation (participants could only answer the ECA's questions). These factors could have reduced any other effect of the independent variables. Indeed, the agent's appearance and the structure of the dialogue were the same across conditions. If participants mainly focused on these elements, they could have paid less attention to the ECA's verbal and nonverbal behavior (the variables that were manipulated and that we were interested in), which thus did not manage to affect their overall experience of the interaction.

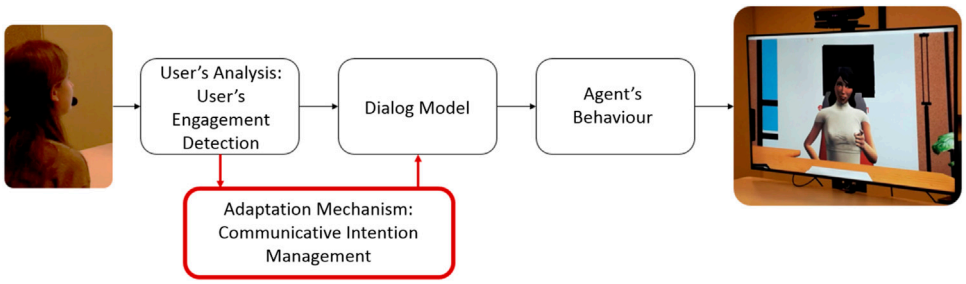

FIGURE 4 | Modified system architecture used in Study 2. In particular, the User's Analysis module contains the model to detect the user's engagement from facial and head/trunk signals. The Communicative Intention module uses reinforcement learning to select the agent's self-presentational strategy.

8 STUDY 2: ADAPTATION OF COMMUNICATIVE STRATEGIES

At this step, we investigate adaptation at a higher level than the previous one, namely, the communicative strategies of the ECA. In particular, we focus on the agent's self-presentational strategies, that is, different techniques to convey different levels of warmth and competence toward the user (Jones and Pittman, 1982). Each strategy is realized in terms of the verbal and nonverbal behavior of the ECA, according to the studies by (Pennebaker, 2011; Callejas et al., 2014; Biancardi et al., 2017a).

While in the previous study, we investigated whether and how adaptation could affect the user's impressions of the agent, we here focus on whether and how adaptation can affect the user's engagement during the interaction.

The goal of this second model is thus to make the ECA learn the communicative strategies that improve the user's engagement, by measuring and using the user's engagement as a reward.

8.1 Architecture

The general architecture described in **Section 5** has been modified in order to contain a module for the detection of the user's engagement and a communicative intention planner for the choice of the agent's self-presentational strategy.

The modified architecture of the system is depicted in **Figure 4**. In the following subsection, we give more details about the modified modules.

8.1.1 User's Analysis: User's Engagement Detection

The User's Analysis module is integrated with a User's Engagement Detection module that continuously computes the overall user's engagement at the end of every speaking turn. The computational model of the user's engagement is based on the detection of facial signals and head/trunk signals, which are indicators of engagement. In particular, smiling is usually considered an indicator of engagement, as it may show that the user is enjoying the interaction (Castellano et al., 2009). Eyebrows are equally important: for example, Corrigan et al. (2016) claimed that "frowning may indicate effortful processing suggesting high levels of cognitive engagement." Head/trunk signals are detected in order to measure the user's attention level. According to Corrigan et al. (2016), attention is a key aspect of engagement; an engaged user continuously gazes at relevant objects/persons during the interaction. We approximate the user's gaze using the user's head and trunk orientation.

8.1.2 Adaptation Mechanism: Communicative Intention Management

During its interaction with the user, the agent has the goal of selecting its self-presentational strategy (e.g., to communicate verbally and nonverbally a given dialogue act with high warmth and low competence). The agent can choose its strategy from a given set of four strategies inspired from Jones and Pittman's taxonomy (Jones and Pittman, 1982):

- Ingratiation: the ECA has the goal to convey positive interpersonal qualities and elicit impressions of high warmth toward the user, without considering its level of competence;
- Supplication: the ECA has the goal to present its weaknesses and elicit impressions of high warmth and low competence;
- Self-promotion: the ECA has the goal to focus on its capabilities and elicit impressions of high competence, without considering its level of warmth; and
- Intimidation: the ECA has the goal to elicit impressions of high competence by decreasing its level of warmth.

The verbal behavior characterizing the different strategies is inspired by the works of Pennebaker (2011) and Callejas et al. (2014). In particular, we took into account the use of you- and we-words, the level of formality of the language, and the length of the sentences.

The choice of the agent's nonverbal behavior is based on our previous studies (Biancardi et al., 2017a; Biancardi et al., 2017b). So, for example, if the current agent's self-presentational strategy is Supplication and the next dialogue act to be spoken is introducing a topic, then the agent would say "I think that while you play there are captors that measure tons of stuffs!" accompanied by smiling and beat gestures. Conversely, if the current agent's self-presentational strategy is Intimidation and the next dialogue act to be spoken is the same, then the agent would say "While you play at video games, several captors

TABLE 2 | Mean and standard deviation values of warmth scores for each level of Communicative Strategy. The mean score for Intim_static is significantly lower than that for all the other conditions.

Communicative Strategy	Warmth
Ingr_static	3.77 ± 0.57
Supp_static	3.54 ± 0.999
Self_static	3.81 ± 0.70
Intim_static	2.63 ± 0.93
Random	3.71 ± 0.80
Adaptation	3.89 ± 0.38

TABLE 3 | Mean and standard deviation values of competence scores for each level of Communicative Strategy. No significant differences among the conditions were found.

Communicative Strategy	Competence
Ingr_static	3.6 ± 0.62
Supp_static	2.98 ± 0.77
Self_static	3.75 ± 0.63
Intim_static	3.65 ± 0.79
Random	3.5 ± 0.70
Adaptation	3.43 ± 0.76

measure your physiological signals," accompanied by ideational gestures without smiling.

To be able to change the agent's communicative strategy according to the detected participant's engagement, we applied a reinforcement learning algorithm to make the ECA learn what strategy to use. Specifically, a multiarmed bandit algorithm (Katehakis and Veinott, 1987) was applied. This algorithm is a simplified setting of reinforcement learning which models agents evolving in an environment where they can perform several actions, each action being more or less rewarding for them. The choice of the action does not affect the state (i.e., what happens in the environment). In our case, the actions that the ECA could perform are the verbal and nonverbal behaviors corresponding to the self-presentational strategy that the ECA aims to communicate. The environment is the interaction with the user, while the state space is the set of dialogue acts used at each speaking turn. The choice of the action does not change the state (i.e., the dialogue act used during the actual speaking turn), but rather, it acts on how this dialogue act is realized by verbal and nonverbal behavior. More details about the multiarmed bandit function used in our model can be found in the study by (Biancardi et al., 2019a).

8.2 Experimental Design

The adaptation model described in **Section 8.1.2** was evaluated by using the scenario described in **Section 6**. Here, we describe the experimental variables manipulated and measured during the experiment.

8.2.1 Independent Variable

The design includes one independent variable, called Communicative Strategy, with six levels determining the way in which the ECA chooses the strategy to use:

1) Adaptation: the ECA uses the adaptation model and thus selects one self-presentational strategy at each speaking turn, by using the user's engagement as a reward;
2) Random: the ECA chooses a random behavior at each speaking turn;
3) Ingr_static: the ECA always adopts the Ingratiation strategy during the whole interaction;
4) Suppl_static: the ECA always adopts the Supplication strategy during the whole interaction;
5) Self_static: the ECA always adopts the Self-promotion strategy during the whole interaction; and
6) Intim_static: the ECA always adopts the Intimidation strategy during the whole interaction.

8.2.2 Measures

The dependent variables measured after the interaction with the ECA are the same as those described in **subsection 7.2.2**.

In addition to these measures, during the interaction, for people who agreed with audio recording of the experiment, we collected quantitative information about their verbal engagement, in particular, the polarity of the user's answer when the ECA asked if they wanted to continue to discuss and the number of any verbal feedback produced by the user during a speaking turn.

8.2.3 Hypotheses

We hypothesized that each self-presentational strategy would elicit the right degree of warmth and competence, in particular, the following:

H1ingr: the ECA in the Ingr_static condition would be perceived as warm by users;
H1supp: the ECA in the Suppl_static condition would be perceived as warm and not competent by users;
H1self: the ECA in the Self_static condition would be perceived as competent by users; and
H1intim: the ECA in the Intim_static condition would be perceived as competent and not warm by users.

Then, we hypothesized the following scenarios:

H2a: an ECA adapting its self-presentational strategies according to the user's engagement would improve the user's experience, compared to a non-adapting ECA and H2b: the ECA in the Adaptation condition would influence how it is perceived in terms of warmth and competence.

8.3 Analysis and Results

75 participants (30 females) took part in the evaluation, equally distributed among the six conditions. The majority of them were in the 18–25 or 36–45 age range and were native French speakers. In this section, we briefly report the main results of our analyses.

A more detailed report can be found in the study by (Biancardi et al., 2019a).

8.3.1 Warmth Scores

A 4 × 2 between-subjects ANOVA revealed a main effect of Communicative Strategy ($F(5, 62) = 4.75$, $p < 0.001$, $\eta 2 = 0.26$) and NARS ($F(1, 62) = 5.74$, $p < 0.05$, $\eta 2 = 0.06$). The *w* ratings were higher from participants with a high NARS score ($M = 3.74$, $SD = 0.77$) than from those with a low NARS score ($M = 3.33$, $SD = 0.92$).

Table 2 shows the mean and SD of *w* scores for each level of Communicative Strategy. Multiple comparisons t-test using Holm's correction shows that the *w* mean for Intim_static is significantly lower than that for all the others. As a consequence, the other conditions are rated as warmer than Intim_static. H1ingr and H1supp are thus validated, and H1intim and H2b are validated for the warmth component.

8.3.2 Competence Scores

No significant results emerged from the analyses. When looking at the means of *c* for each condition (see **Table 3**), Supp_static is the one with the lower score, even if its difference with the other scores does not reach statistical significance (all *p*-values > 0.1). H1supp and H1intim (for the competence component) are not validated.

8.3.3 User's Experience of the Interaction

Participants in the Ingr_static condition were more satisfied from the interaction than those in Suppl_static ($z = 2.88$, p-adj < 0.05) and in Intim_static ($z = 2.56$, p-adj < 0.05). Participants in the Ingr_static condition also liked the ECA more than participants in the Intim_static condition ($z = 2.87$, p-adj < 0.05). No differences were found between the scores of the participants in the Adaptation condition and those of the other participants for any of the items measuring *exp*.

The *exp* scores are also affected by participants' *a priori* about virtual agents (measured through the NARS questionnaire). In particular, participants who got high scores in the NARS questionnaire were more satisfied by the interaction ($U = 910.5$, $p < 0.05$), were more motivated to continue the interaction ($U = 998$, $p = 0.001$), and perceived the agent as less closed to a computer ($U = 1028$, $p < 0.001$) than people who got low scores in the NARS questionnaire.

Another interesting result concerns the effect of age on participants' satisfaction ($H(4) = 15.05$, $p < 0.01$); people in the age range of 55+ were more satisfied than people of any other age range (all *p*-adj < 0.05).

On the whole, these results do not allow us to validate H2a, but the agent's adaptation was found to have at least an effect on its level of warmth (H2b).

8.3.4 Verbal Cues of Engagement

During each speaking turn, the user was free to reply to the agent's utterances. We consider as a user's verbal feedback any type of verbal reply to the ECA, from a simple backchannel (e.g., "ok" and "mm") to a longer response (e.g., giving an opinion about what the ECA said). In general, participants who did not give much verbal feedback (i.e., less than 13 replies to the agent's utterances over all the speaking turns) answered positively to the ECA when it asked whether they wanted to continue to discuss with it, compared to the participants who gave more verbal feedback ($OR = 4.27$, $p < 0.05$). In addition, we found that the participants who did not give much verbal feedback liked the ECA more than those who talked a lot during the interaction ($U = 36.5$, $p < 0.05$). However, no differences in any of the dependent variables were found according to Communicative Strategy.

8.4 Discussion

First of all, regarding H1, the only statistically significant results concern the perception of the agent's warmth. The ECA was rated as colder when it adopted the Intim_static strategy than when it adopted the other conditions. This supports the thesis of the primacy of the warmth dimension (Wojciszke and Abele, 2008), and it is in line with the positive–negative asymmetry effect described by (Peeters and Czapinski, 1990), who argued that negative information generally has a higher impact on person perception than positive information. In our case, when the ECA displayed cold (i.e., low warmth) behaviors (i.e., in the Intim_static condition), it was judged by participants with statistically significant lower ratings of warmth. Regarding the other conditions (Ingr_static, Supp_static, Self_static, Adaptation, and Random), they elicited warmer impressions in the user, but there was not one strategy that was better than the others in this regard. The fact that Self_static also elicited the same level of warmth as the others reflected a halo effect (Rosenberg et al., 1968); the behaviors displayed to appear competent influenced its warmth perception in the same direction.

Regarding H2, the results do not validate our hypothesis (H2a) that the interaction would be improved when the ECA managed its impressions by adapting its strategy according to the user's engagement. When analyzing scores for *exp* items, we found that participants were more satisfied by the interaction and they liked the ECA more when the ECA wanted to be perceived as warm (i.e., in the Ingr_static condition) than when it wanted to be perceived as cold and competent (i.e., in the Intim_static condition). A hypothesis is that since the ECA was perceived warmer in the Ingr_static condition, it could have positively influenced the ratings of the other items, like the user's satisfaction. Concerning H2b with regard to a possible effect of the agent's adaptation on the user's perception of its warmth and competence, it is interesting to see that when the ECA adapted its self-presentational strategy according to the user's overall engagement, it was perceived as warm. This highlights a link between the agent's adaptation, the user's engagement, and a warm impression; the more the ECA adapted its behaviors, the more the user was engaged and the more she/he perceived the ECA as warm.

9 STUDY 3: ADAPTATION AT A SIGNAL LEVEL

At this step, we are interested in low-level adaptation at the signal level. We aim to model how the ECA can adapt its signals to the user's signals. Thus, we make the ECA predict the signals to

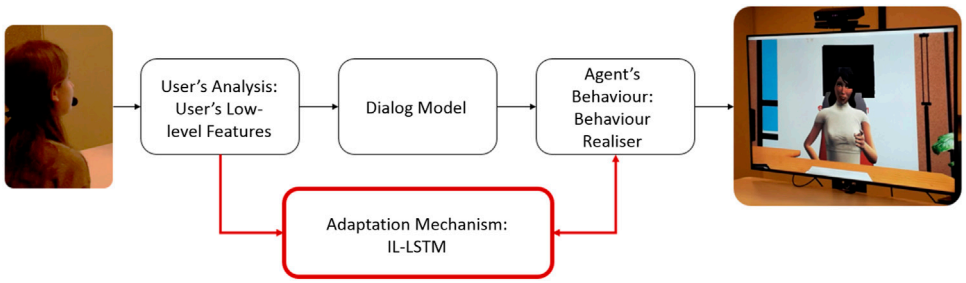

FIGURE 5 | Modified system architecture used in Study 3. In particular, the User's Analysis module detects the user's low-level signals such as head and eye rotations and lip corner activity. The Adaptation Mechanism module exploits the IL-LSTM model for selecting the agent's low-level signals. In the Agent's Behavior module, the Behavior Realizer is customized in order to take into account the agent's communicative behaviors and signals coming from the IL-LSTM module in real time.

display at each time step, according to those displayed by both the ECA and the user during a given time window. For the sake of simplicity, we consider a subset of signals, namely, lip corner movement (AU12), gaze direction, and head movement. To reach our aim, we follow a two-step approach. At first, we need to predict which signals that are due to adaptation to the user's behaviors should be displayed by the ECA at each time step. The prediction of signal adaptation is learned on human–human interaction. The ECA ought to communicate its intentions to adapt to the user's signals. Then, the second step of our approach consists in blending the predicted signals linked to the adaptation mechanism with the nonverbal behaviors corresponding to the agent's communicative intentions. We describe our algorithm in further detail in **subsection 9.1.2**.

9.1 Architecture

The general architecture described in **Section 5** has been modified in order to contain a module for predicting the next social signal to be merged with the agent's other communicative ones. The modified architecture of the system is depicted in **Figure 5**. In the following subsection, we explain the modified modules. More details about these modules can be found in the study by (Dermouche and Pelachaud, 2019).

9.1.1 User's Analysis: User's Low-Level Features

Low-level features of the user are obtained from the User's Analysis module using EyesWeb of the general architecture. In this model, we consider a subset of these features, namely, the user's head direction, eye direction, and AU12 (upper lip corner activity). At every frame, the EyesWeb module extracts these features and sends the last 20 analyzed frames to the Adaptation Mechanism module IL-LSTM (see **Section 9.1.2**). It also sends the user's conversational state (speaking or not) computed from the detection of the user's voice activity (done in EyesWeb) and from the agent-turn information provided by the dialogue manager Flipper.

9.1.2 Adaptation Mechanism: Interaction Loop–LSTM

In this version of the architecture, the adaptation mechanism is based on a predictive model trained on data of human–human interactions. We used the NoXi database (Cafaro et al., 2017) to train a long short-term memory (LSTM) model that takes as input sequences of signals of two interactants over a sliding window of n frames to predict which signal(s) should display one participant at time $n+1$. We call this model IL-LSTM, which stands for interaction loop–LSTM. LSTM is a kind of recurrent neural network. It is mainly used when "context" is important, that is, decisions from the past can influence the current ones. It allows us to model both sequentiality and temporality of nonverbal behaviors.

We apply the IL-LSTM model to the human–agent interaction. Thus, given the signals produced by both, the human and the ECA, over a time window, the model outputs which signals should display the ECA at the next time step (here, a frame). The predicted signals are sent to the Behavior Realizer of the Agent's Behavior module where they are merged with the behaviors of the ECA related to its communicative intents.

9.1.3 Agent's Behavior: Behavior Realizer

We have updated the Behavior Realizer so that the ECA not only communicates its intentions but also adapts its behaviors in real time to the user's behaviors. This module blends the predicted signals linked to the adaptation mechanism with the nonverbal behaviors corresponding to its communicative intentions that have been outputted using the GRETA agent platform (Pecune et al., 2014). More precisely, the dialogue module Flipper sends the set of communicative intentions to the Agent's Behavior module. This module computes the multimodal behavior of the ECA and sends it to the Behavior Realizer that computes the animation of the ECA's face and body. Then, before sending each frame to be displayed by the animation player, the animation computed from the communicative intentions is merged with the animation predicted by the Adaption Mechanism module. This operation is repeated at every frame.

9.2 Experimental Design

The adaptation model described in the previous section was evaluated by using the scenario described in **Section 6**. Here, we describe the experimental variables manipulated and measured during the experiment.

9.2.1 Independent Variable

We manipulated the type of low-level adaptation of the ECA by considering five conditions:

TABLE 4 | Mean ± standard deviation of each dimension of the questionnaires (each row of the table), for each of the five conditions (each column).

	Random	Head	Lip Corners	Eyes	All
Competence	2.98 ± 1.22	3.45 ± 0.81	3.73 ± 0.73	3.65 ± 1.06	3.61 ± 1.12
Distance	2.5 ± 1.12	2.6 ± 1.03	1.76 ± 0.97	2 ± 1.12	1.47 ± 1.03
Friendliness	3.03 ± 1.12	3.22 ± 0.86	3.80 ± 0.83	3.33 ± 0.86	4.09 ± 0.90[a]
Involvement	2.65 ± 1.22	2.65 ± 1.15	3.52 ± 1.00[a]	2.83 ± 1.33	3.60 ± 1.07
Realism	1.7 ± 0.92	1.95 ± 0.82	2.52 ± 1.12	2.08 ± 0.90	1.73 ± 0.86
Relevance	2.95 ± 1.38	3.86 ± 0.72	3.97 ± 0.79[a]	3.5 ± 1.24	3.80 ± 1.01
Satisfaction	2.46 ± 1.21	2.84 ± 0.77	3.39 ± 0.06	3.27 ± 1.08	3.39 ± 0.93

[a]*indicates that the score is significantly different compared to that in the Random condition (p – adj <.05).*

- Random: when the ECA did not adapt its behavior;
- Head: when the ECA adapted its head rotation according to the user's behavior;
- Lip Corners: when the ECA adapted its lip corner puller movement (AU12) according to the user's behavior;
- Eyes: when the ECA adapted its eye rotation according to the user's behavior; and
- All: when the ECA adapted its head and eye rotation and lip corner movement, according to the user's behavior.

We tested these five conditions using a between-subjects design.

9.2.2 Measures

The dependent variables measured after the interaction with the ECA were the user's engagement and the perceived friendliness of the ECA.

The user's engagement was evaluated using the I-PEFiC framework (van Vugt et al., 2006) that encompasses the user's engagement and satisfaction during human–agent interaction. This framework considers different dimensions regarding the perception of the ECA (in terms of realism, competence, and relevance) as well as the user's engagement (involvement and distance) and the user's satisfaction. We adapted the questionnaire proposed by Van Vugt and others to measure the behavior of the ECA along these dimensions (van Vugt et al., 2006). The perceived friendliness of the ECA was measured using the adjectives kind, warm, agreeable, and sympathetic of the IAS questionnaire (Wiggins, 1979).

As for the other two studies, we also measured the *a priori* attitude of participants towards virtual agents using the NARS questionnaire.

9.2.3 Hypotheses

Previous studies (Liu et al., 2008; Woolf et al., 2009; Levitan, 2013) have found that users' satisfaction about their interaction with an ECA is greater when the ECA adapts its behavior to the user's one. From these results, we could expect that the user would be more satisfied about the interaction when the ECA adapted its low-level signals according to their behaviors. We also assumed that the ECA adapting its lip corner puller (that is related to smiling) would be perceived as friendlier. Thus, our hypotheses were as follows:

H1Head: when the ECA adapted its head rotation, the users would be more satisfied with the interaction than the users interacting with the ECA in the Random condition.

H2aLips: when the ECA adapted its lip corner movement (AU12), the users would be more satisfied with the interaction than the users interacting with the ECA in the Random condition.

H2bLips: when the ECA adapted its lip corner movement (AU12), it would be evaluated as friendlier than the ECA in the Random condition.

H3Eyes: when the ECA adapted its eye rotation, the users would be more satisfied with the interaction than the users interacting with the ECA in the Random condition.

H4aAll: when the ECA adapted its head and eye rotations and lip corner movement, the users would be more satisfied with the interaction than the users interacting with the ECA in the Random condition.

H4bAll: when the ECA adapts its head and eye rotations and lip corner movement, it would be evaluated as friendlier than the ECA in the Random condition.

9.3 Analysis and Results

101 participants (55 females), almost equally distributed among the five conditions, took part in our experiment. 95% of participants were native French speakers. 32% of them were in the range of 18–25˚years old, 17% were in the range of 25–36, 21% were in the range of 36–45, 18% were in the range of 46–55, and 12% were over 55˚years old. For each dimension of the user's engagement questionnaire, as well as for that about the perceived friendliness of the ECA, Cronbach's αs were >0.8; we then computed the mean of the scores in order to have one score for each dimension. The mean and standard deviation of each measured dimension for each of the five conditions are shown in **Table 4**.

As our data were not normally distributed (the Shapiro test's $p < 0.5$), we used the unpaired Wilcoxon test (equivalent to *t*-test) to measure how participants' ratings differed between the Random condition and each of the other conditions.

In the Head condition, we could not find differences with the Random condition. We conclude that the hypothesis H1Head is rejected.

In the Lip Corners condition, compared to participants in the Random condition, participants were more involved ($W = 98.5$, p-*adj* <.05). We can also note that the ECA was evaluated as more positive on the relevance dimension ($W = 104.5$, p-*adj* <.05). We can conclude that the hypotheses H2aLips and H2bLips are not validated, but the adaptation of lip corner movement still has a positive effect on other dimensions related to the user's engagement.

In the Eyes condition, participants were satisfied with the ECA as they were with the ECA in the Random condition. Thus, the hypothesis H3Eyes is rejected.

In the All condition, the ECA was evaluated as friendlier ($W = 104.5$, p-$adj < .05$) than the ECA in the Random condition. So, H4aAll is supported, while H4bAll is rejected.

Results of the NARS questionnaire indicated that 40, 30, and 30% of participants, respectively, had a positive, neutral, and negative attitude toward virtual agents. An ANOVA test was performed to study the influence of participants' *a priori* toward virtual agents on their engagement in the interaction. Participants' prior attitude toward ECAs had a main effect on participants' distance ($F(1, 93) = 5.13, p < .05$)). Results of pairwise comparisons with Bonferroni adjustment highlighted that participants with a prior negative attitude were less engaged (more distant (p-$adj < .05$) and less involved (p-$adj < .05$)) than those with a prior positive attitude.

9.4 Discussion

The results of this study showed that participants' engagement and perception of the ECA's friendliness were positively impacted when the ECA adapted its low-level signals.

These results were significant only when the ECA adapted its lip corner movement (AU12) to the user's behavior (mainly their smile), that is, in the Lip Corners and All conditions. In the case of head and eye rotation adaptation, we found a trend on some dimensions but no significant differences compared to the Random condition. These results could be caused by the adopted evaluation setting where the ECA and the user faced each other. During the interaction, most participants gazed at the ECA without doing any postural shift or even changing their gaze and head direction. They were mainly still and staring at the ECA. The adaptive behaviors, that is, head and eye rotation of the ECA computed from the user's behaviors, remained constant throughout the interaction. They reflected participants' behaviors (that were not moving much). Thus, in the Head and Eyes adapting conditions, the ECA showed much less expressiveness and may have appeared much less lively, which may have impacted participants' engagement in the interaction.

10 GENERAL DISCUSSION

In our studies, we applied the interaction adaptation theory (see **Section 2**) on the ECA. That is, our adapting ECA had the requirement R that it needed to adapt in order to have a successful interaction. Its desire D was to maximize the user's experience by eliciting a specific impression toward the user or maintaining the user's engagement. Finally, its expectations (Es) were that the user's experience would be better when interacting with an adaptive ECA. All these factors rely on the general hypothesis that the user expects to interact with a social entity. According to this hypothesis, the ECA should adapt its behavior like humans do (Appel et al., 2012).

We have looked at different adaptation mechanisms through three studies, each focusing on a specific type of adaptation. In our studies, we found that these mechanisms impacted the user's experience of the interaction and their perception of the ECA. Moreover, in all three studies, interacting with an adaptive ECA vs. a nonadaptive ECA tended to be more positively perceived. More precisely, manipulating the agent's behaviors (Study 1) had an impact on the user's perception of the ECA while low-level adaptation (Study 3) positively influenced the user's experience of the interaction. Regarding managing conversational strategies (Study 2), the ECA was perceived as warmer when it managed those that increased the user's engagement vs. when it did not change them all along the interaction.

These results suggest that the IAT framework allows for enhancing human–agent interaction. Indeed, the adaptive ECA shows some improvement in the quality of the interaction and the perception of the ECA in terms of social attitudes.

However, not all our hypotheses were verified. This could be related to the fact that we based our framework on the general hypothesis that the user expects to interact with a social entity. The ECA did not take into account the fact that the user also had their specific requirements, desires, and expectations, along with the expectancy to interact with a social agent. Yet, the ECA did not check if the user still considered it a social entity during the interaction. It based its behaviors only on the human's detected engagement and impressions. Moreover, the modules to detect engagement or impressions work in a given time window, but they do not consider their evolution through time. For example, the engagement module computes that participants are engaged if they look straight at the ECA without reporting any information stating that the participants stare fixedly at the ECA. The fact that participants do not change their gaze direction toward the ECA could be interpreted as participants not viewing the ECA as a social entity with humanlike qualities (Appel et al., 2012).

Expectancy violation theory (Burgoon, 1993) could help to better understand this gap. This theory explains how confirmations and violations of people's expectancies affect communication outcomes such as attraction, liking, credibility persuasion, and learning. In particular, positive violations are predicted to produce better outcomes than positive confirmations, and negative violations are predicted to produce worse outcomes than negative confirmations. Expectancy violation theory has already been demonstrated to affect human–human interaction (Burgoon, 1993) and when people are in front of an ECA (Burgoon et al., 2016; Biancardi et al., 2017b) or a robot (Weber et al., 2018). In our work, we took into account the role of expectancies as part of IAT. Our results suggest that expectancies could play a more important role than the one we attributed to them and that they should be better modeled when developing human–agent adaptation. Future works in this context should combine expectancy violation theory with IAT. In this way, the ECA should be able to detect the user's expectancies in terms of beliefs and desires. It should also be able to check if those expectancies about the interaction correspond to the expected ones and then react accordingly. For example, in our studies, we found some effects of people's *a priori* about virtual agents: people who got higher scores in the NARS questionnaire generally perceived the ECA as warmer than people who got lower scores in the NARS questionnaire. This effect could have been mitigated if the agent could detect the user's *a priori*.

Even with these limits, the results of our studies show that an adaptive model for a virtual agent inspired from IAT partially managed to produce an impact on the user's experience of the interaction and on their perception of the ECA. This could be useful for personalizing systems for different applications such as education, healthcare, or

entertainment, where there is a need of adaptation according to users' type and behaviors and/or interaction contexts.

The different adaptation models we developed also confirm the potential of automatic behavior analysis for the estimation of different users' characteristics. These methods can be used to better understand the user's profile and can also be applied to human–computer interaction in general to inform adaptation models in real time.

Moreover, the use of adaptation mechanisms inspired from IAT could help mitigate the negative effect of some interaction problems that are more difficult to solve, due to, for example, technological limits of the system. Indeed, adaptation acts to enhance the agent's perception and the perceived interaction quality. Improving adaptation mechanisms may help to counterbalance technological shortcomings. It may also improve the acceptability of innovative technologies that are likely to be part of our daily lives, in the context of work, health, leisure, etc.

11 CONCLUSION AND FUTURE WORK

In this study, we investigated adaptation in human–agent interaction. In particular, we reported our work about three models focusing on different levels of the agent's adaptation (the behavioral, conversational, and signal levels), by framing them in the same theoretical framework (Burgoon et al., 2007). In all the adaptation mechanisms implemented in the models, the user's behavior is taken into account by the ECA during the interaction in real time. Evaluation studies showed a tendency toward a positive impact of the adaptive ECA on the user's experience and perception of the ECA, encouraging us to continue to investigate in this direction.

One limitation of our models is their reliance on the interaction scenario. Indeed, to obtain good performances of adaptation models using reinforcement learning algorithms, a scenario including an adequate number of steps is required. In our case, the agent ended up selecting a specific combination of behaviors only during the later part of the interaction. A longer interaction with more steps would allow an adaptive agent using reinforcement learning algorithms to better learn. Another possibility would be to have participants interacting more than once with the virtual agent. This latter would require adding a memory adaptation module (Ahmad et al., 2017). This would also allow for checking whether the same user prefers the same behavior and/or conversational strategies from the agent over several interactions. Similarly, regarding adaptation models reflecting the user's behavior, the less the user moves during the interaction, the less the agent's expressivity level. The interaction scenario should be designed in order to elicit the user's participation, including strategies to tickle users when they become too still and nonreactive. For example, one could use a scenario including a collaborative task where both the agent and the user would interact with different objects. In such a setting, although it would require us to extend our engagement detection module to include joint attention, we expect that the participants would also perform many more head movements that, in turn, could be useful for a better low-level adaptation of the agent.

In the future, our work could be improved and explored along further axes. We list three of them here. First, the three models presented in this article were implemented and evaluated independently from each other. It could be interesting to merge the three adaptation mechanisms in a broader model and investigate the impacts of the agent's adaptation along different levels at the same time. Second, in our studies, the agent adapted its behaviors to the user's ones without considering if the relationship between the behaviors of the dyad showed any specific interaction patterns. In particular, we have not made explicit if the agent's behavior should either match, reciprocate, complement, compensate, or mirror their human interlocutor's behavior (Burgoon et al., 2007). Also, we have not measured any similarities, synchronization, or imitation between the user's and the agent's behavior when we analyzed the data of our studies. Since adaptation may be signaled through a larger variety of behavior manifestations during an interaction, more adaptation mechanisms could be implemented. One last important direction for future work concerns the improvement of the interaction with the user. This would reduce possible secondary effects of uncontrolled variables, such as the user's expectancies, and allow for better studying of the effects of the agent's adaptation. We aim to improve the agent's conversational skills to ensure conversation repairs and interruptions and by letting the user choose the topic of conversation (e.g., from a set of possible ones) and drive the discussion. In addition to these improvements, the user's expectancies should also be better modeled by taking into account expectancy violation theory in addition to interaction adaptation theory.

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.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BB and SD: conceptualization, methodology, investigation, formal analysis, and writing—original draft preparation. CP: writing—review and editing, supervision, and funding acquisition.

FUNDING

The research topics addressed in this article have been investigated in the framework of the EU Horizon 2020 research and innovation program under Grant Agreement Number 769553 and the ANR project Impressions ANR-15-CE23-0023.

REFERENCES

Ahmad, M. I., Mubin, O., and Orlando, J. (2017). Adaptive Social Robot for Sustaining Social Engagement during Long-Term Children-Robot Interaction. *Int. J. Human-Computer Interaction* 33 (12), 943–962. doi:10.1080/10447318.2017.1300750

Altman, I., Vinsel, A., and Brown, B. B. (1981). Dialectic Conceptions in Social Psychology: An Application to Social Penetration and Privacy Regulation. *Adv. Exp. Soc. Psychol.* 14, 107–160. doi:10.1016/s0065-2601(08)60371-8

Andersen, P. A. (1985). "Nonverbal Immediacy in Interpersonal Communication," in *Multichannel Integrations of Nonverbal Behavior*. Editors A. A. Siegman and S. Feldstein (Hillsdale, NJ:Erlbaum: Psychology Press), 1–36.

Appel, J., von der Pütten, A., Krämer, N. C., and Gratch, J. (2012). Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction. *Adv. Human-Computer Interaction* 2012, 324694. doi:10.1155/2012/324694

Aragonés, J. I., Poggio, L., Sevillano, V., Pérez-López, R., and Sánchez-Bernardos, M.-L. (2015). Measuring warmth and competence at inter-group, interpersonal and individual levels/Medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. *Revista de Psicología Soc.* 30 (3), 407–438. doi:10.1080/02134748.2015.1065084

Argyle, M., and Dean, J. (1965). Eye-contact, Distance and Affiliation. *Sociometry* 28 (3), 289–304. doi:10.2307/2786027

Argyle, M. (1972). *Non-verbal Communication in Human Social Interaction*. New york: Cambridge University Press.

Bailenson, J. N., and Yee, N. (2005). Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Psychol. Sci.* 16, 814–819. doi:10.1111/j.1467-9280.2005.01619.x

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "Openface: an Open Source Facial Behavior Analysis Toolkit," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference, Lake Placid, NY, USA, 7-10 March 2016 (IEEE), 1–10. doi:10.1109/WACV.2016.7477553

Bergmann, K., Eyssel, F., and Kopp, S. (2012). "A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time," in International Conference on Intelligent Virtual Agents, Santa Cruz, CA, USA, September, 12-14 (Springer), 126–138. doi:10.1007/978-3-642-33197-8_13

Bernieri, F. J., Reznick, J. S., and Rosenthal, R. (1988). Synchrony, Pseudosynchrony, and Dissynchrony: Measuring the Entrainment Process in Mother-Infant Interactions. *J. Personal. Soc. Psychol.* 54 (2), 243–253. doi:10.1037/0022-3514.54.2.243

Bevacqua, E., Mancini, M., and Pelachaud, C. (2008). "A Listening Agent Exhibiting Variable Behavior," in International Conference on Intelligent Virtual Agents, Tokyo, Japan, September 1-3, 2008 (Springer), 262–269. doi:10.1007/978-3-540-85483-8_27

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017a). "Analyzing First Impressions of Warmth and Competence from Observable Nonverbal Cues in Expert-Novice Interactions," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, November 2017 (Glasgow: ACM), 341–349. doi:10.1145/3136755.3136779

Biancardi, B., Cafaro, A., and Pelachaud, C. (2017b). "Could a Virtual Agent Be Warm and Competent? Investigating User's Impressions of Agent's Non-verbal Behaviors," in Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, November 2017 (Glasgow: ACM), 22–24. doi:10.1145/3139491.3139498

Biancardi, B., Mancini, M., Lerner, P., and Pelachaud, C. (2019a). Managing an Agent's Self-Presentational Strategies during an Interaction. *Front. Robot. AI* 6, 93. doi:10.3389/frobt.2019.00093

Biancardi, B., Wang, C., Mancini, M., Cafaro, A., Chanel, G., and Pelachaud, C. (2019b). "A Computational Model for Managing Impressions of an Embodied Conversational Agent in Real-Time," in 2019 International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3-6 Sept. 2019 (IEEE). doi:10.1109/ACII.2019.8925495

Bickmore, T., Pfeifer, L., and Schulman, D. (2011). "Relational Agents Improve Engagement and Learning in Science Museum Visitors," in International Conference on Intelligent Virtual Agents, Reykjavik, Iceland, September 15-17, 2011 (Springer), 55–67. doi:10.1007/978-3-642-23974-8_7

Bickmore, T. W., Vardoulakis, L. M. P., and Schulman, D. (2013). Tinker: a Relational Agent Museum Guide. *Auton. Agent Multi-agent Syst.* 27 (2), 254–276. doi:10.1007/s10458-012-9216-7

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., et al. (2016). Application of Expectancy Violations Theory to Communication with and Judgments about Embodied Agents during a Decision-Making Task. *Int. J. Human-Computer Stud.* 91, 24–36. doi:10.1016/j.ijhcs.2016.02.002

Burgoon, J. K. (1993). Interpersonal Expectations, Expectancy Violations, and Emotional Communication. *J. Lang. Soc. Psychol.* 12 (1-2), 30–48. doi:10.1177/0261927x93121003

Burgoon, J. K., Stern, L. A., and Dillman, L. (2007). *Interpersonal Adaptation: Dyadic Interaction Patterns*. New york: Cambridge University Press.

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First Impressions in Human--Agent Virtual Encounters. *ACM Trans. Comput.-Hum. Interact.* 23 (4), 1–40. doi:10.1145/2940325

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., André, E., and Valstar, M. (2017). "The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, November 2017 (Glasgow: ACM), 350–359. doi:10.1145/3136755.3136780

Callejas, Z., Ravenet, B., Ochs, M., and Pelachaud, C. (2014). "A Computational Model of Social Attitudes for a Virtual Recruiter," in Proceedings of the 13th international conference on Autonomous Agents and Multi-Agent Systems, May 2014 (Paris: ACM), 93–100.

Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., Ricci, A., and Volpe, G. (2004). "Toward Real-Time Multimodal Processing: Eyesweb 4.0," in Proceedings of the Artificial Intelligence and the Simulation of Behavior (AISB), 2004 convention: motion. Emotion and cognition, Leeds, 22–26.

Cappella, J. N., and Greene, J. O. (1982). A Discrepancy-arousal Explanation of Mutual Influence in Expressive Behavior for Adult and Infant-adult Interaction1. *Commun. Monogr.* 49 (2), 89–114. doi:10.1080/03637758209376074

Cappella, J. N. (1991). Mutual Adaptation and Relativity of Measurement. *Studying interpersonal interaction* 1, 103–117. doi:10.1111/j.1468-2885.1991.tb00002.x

Cappella, J. N. (1981). Mutual Influence in Expressive Behavior: Adult-Adult and Infant-Adult Dyadic Interaction. *Psychol. Bull.* 89 (1), 101–132. doi:10.1037/0033-2909.89.1.101

Cassell, J., Bickmore, T., Vilhjálmsson, H., and Yan, H. (2000). "More Than Just a Pretty Face: Affordances of Embodiment," in International Conference of Intelligent Virtual Agents, January 2000 (Springer), 52–59.

Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009). "Detecting User Engagement with a Robot Companion Using Task and Social Interaction-Based Features," in Proceedings of the 2009 international conference on Multimodal interfaces, November 2009 (Cambridge, MA: ACM), 119–126.

Changchun Liu, C., Conn, K., Sarkar, N., and Stone, W. (2008). Online Affect Detection and Robot Behavior Adaptation for Intervention of Children with Autism. *IEEE Trans. Robot.* 24 (4), 883–896. doi:10.1109/tro.2008.2001362

Chartrand, T. L., and Bargh, J. A. (1999). The Chameleon Effect: The Perception-Behavior Link and Social Interaction. *J. Personal. Soc. Psychol.* 76 (6), 893–910. doi:10.1037/0022-3514.76.6.893

Condon, W. S., and Ogston, W. D. (1971). "Speech and Body Motion Synchrony of the Speaker-Hearer," in *The Perception of Language*. Editors D. L. Horton and J. J. J (Columbus, Ohio: Charles Merrill), 150–184.

Coninx, A., Baxter, P., Oleari, E., Bellini, S., Bierman, B., Henkemans, O. B., et al. (2016). Towards Long-Term Social Child-Robot Interaction: Using Multi-Activity Switching to Engage Young Users. *J. Human-Robot Interaction* 5 (1), 32–67.

Corrigan, L. J., Peters, C., Küster, D., and Castellano, G. (2016). "Engagement Perception and Generation for Social Robots and Virtual Agents,". *Toward Robotic Socially Believable Behaving Systems*. Editors A. Esposito and L. C. Jain (Springer), 29–51. doi:10.1007/978-3-319-31056-5_4

Cuddy, A. J. C., Fiske, S. T., and Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the Bias Map. *Adv. Exp. Soc. Psychol.* 40, 61–149. doi:10.1016/s0065-2601(07)00002-0

Dermouche, S., and Pelachaud, C. (2019). "Engagement Modeling in Dyadic Interaction," in 2019 International Conference on Multimodal Interaction, October 2019 (Suzhou, Jiangsu: ACM), 440–445.

Dindia, K. (1988). A Comparison of Several Statistical Tests of Reciprocity of Self-Disclosure. *Commun. Res.* 15 (6), 726–752. doi:10.1177/009365088015006004

Duchenne, B. (1990). *The Mechanism of Human Facial Expression or an Electro-Physiological Analysis of the Expression of the Emotions*. New York: Cambridge University Press. (Original work published 1862).

Ekman, P., Friesen, W., and Hager, J. (2002). *Facial Action Coding System (FACS). A Human Face*. Salt Lake City: Research Nexus.

Fischer-Lokou, J., Martin, A., Guéguen, N., and Lamy, L. (2011). Mimicry and Propagation of Prosocial Behavior in a Natural Setting. *Psychol. Rep.* 108 (2), 599–605. doi:10.2466/07.17.21.pr0.108.2.599-605

Fiske, S. T., Cuddy, A. J. C., and Glick, P. (2007). Universal Dimensions of Social Cognition: Warmth and Competence. *Trends Cognitive Sciences* 11 (2), 77–83. doi:10.1016/j.tics.2006.11.005

Gallois, C., Ogay, T., and Giles, H. (2005). "Communication Accommodation Theory: A Look Back and a Look Ahead," in *Theorizing about Intercultural Communication*. Editor W. B. Gudykunst (Thousand Oaks, CA: SAGE), 121–148.

Giles, H., Coupland, N., and Coupland, J. (1991). "Accommodation Theory: Communication, Context, and Consequence," in *Studies in Emotion and Social Interaction. Contexts of Accommodation: Developments in Applied Sociolinguistics*. Editors H, C. J. Giles and N. Coupland (Cambridge University Press), 1, 1–68. doi:10.1017/cbo9780511663673.001

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., and Breazeal, C. (2016). "Affective Personalization of a Social Robot Tutor for Children's Second Language Skills," in Thirtieth AAAI Conference on Artificial Intelligence, February 2016 (Phoenix, Arizona: ACM), 3951–3957. doi:10.5555/3016387.3016461

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *Am. sociological Rev.* 25 (2), 161–178. doi:10.2307/2092623

Gueguen, N., Jacob, C., and Martin, A. (2009). Mimicry in Social Interaction: Its Effect on Human Judgment and Behavior. *Eur. J. Soc. Sci.* 8 (2), 253–259.

Hale, J. L., and Burgoon, J. K. (1984). Models of Reactions to Changes in Nonverbal Immediacy. *J. Nonverbal Behav.* 8 (4), 287–314. doi:10.1007/bf00985984

Hemminahaus, J., and Kopp, S. (2017). "Towards Adaptive Social Behavior Generation for Assistive Robots Using Reinforcement Learning," in 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, Austria, 6-9 March 2017 (IEEE), 332–340.

Huang, L., Morency, L.-P., and Gratch, J. (2010). "Learning Backchannel Prediction Model from Parasocial Consensus Sampling: a Subjective Evaluation," in International Conference on Intelligent Virtual Agents, Philadelphia, PA, USA, September 20-22 (Springer), 159–172. doi:10.1007/978-3-642-15892-6_17

Huang, L., Morency, L.-P., and Gratch, J. (2011). "Virtual Rapport 2.0," in International Conference on Intelligent Virtual Agents, Reykjavik, Iceland, September 15-17, 2011 (Springer), 68–79. doi:10.1007/978-3-642-23974-8_8

Infante, D. A., Rancer, A. S., and Avtgis, T. A. (2010). *Contemporary Communication Theory*. IA: Kendall Hunt Dubuque.

Jones, E. E., and Pittman, T. S. (1982). Toward a General Theory of Strategic Self-Presentation. *Psychol. Perspect. self* 1 (1), 231–262.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental Dimensions of Social Judgment: Understanding the Relations between Judgments of Competence and Warmth. *J. Personal. Soc. Psychol.* 89 (6), 899–913. doi:10.1037/0022-3514.89.6.899

Katehakis, M. N., and Veinott, A. F., Jr (1987). The Multi-Armed Bandit Problem: Decomposition and Computation. *Mathematics OR* 12 (2), 262–268. doi:10.1287/moor.12.2.262

Kopp, S., Gesellensetter, L., Krämer, N. C., and Wachsmuth, I. (2005). "A Conversational Agent as Museum Guide - Design and Evaluation of a Real-World Application," in International Conference on Intelligent Virtual Agents, Kos, Greece, September 12-14 (Springer), 329–343. doi:10.1007/11550617_28

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). "Conditional Random fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, June 2001 (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc), 282–289.

Lakin, J. L., and Chartrand, T. L. (2003). Using Nonconscious Behavioral Mimicry to Create Affiliation and Rapport. *Psychol. Sci.* 14 (4), 334–339. doi:10.1111/1467-9280.14481

Levitan, R. (2013). "Entrainment in Spoken Dialogue Systems: Adopting, Predicting and Influencing User Behavior," in Proceedings of the 2013 NAACL HLT Student Research Workshop, Atlanta, Georgia, June 2013 (Atlanta, Georgia: Association for Computational Linguistics), 84–90.

Lisetti, C., Amini, R., Yasavur, U., and Rishe, N. (2013). I Can Help You Change! an Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Trans. Manage. Inf. Syst.* 4 (4), 1–28. doi:10.1145/2544103

Lubold, N., Walker, E., and Pon-Barry, H. (2016). "Effects of Voice-Adaptation and Social Dialogue on Perceptions of a Robotic Learning Companion," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7-10 March 2016 (IEEE), 255–262. doi:10.1109/HRI.2016.7451760

Maricchiolo, F., Gnisci, A., Bonaiuto, M., and Ficca, G. (2009). Effects of Different Types of Hand Gestures in Persuasive Speech on Receivers' Evaluations. *Lang. Cogn. Process.* 24 (2), 239–266. doi:10.1080/01690960802159929

Mills, C., Bosch, N., Krasich, K., and D'Mello, S. K. (2019). "Reducing Mind-Wandering during Vicarious Learning from an Intelligent Tutoring System," in International Conference on Artificial Intelligence in Education, Chicago, IL, USA, June 25-29, 2019 (Springer), 296–307. doi:10.1007/978-3-030-23204-7_25

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental Investigation into Influence of Negative Attitudes toward Robots on Human-Robot Interaction. *AI Soc.* 20 (2), 138–150. doi:10.1007/s00146-005-0012-7

Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in Virtual Agents and Robots. *ACM Trans. Interact. Intell. Syst.* 7 (3), 1–40. doi:10.1145/2912150

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., and Pelachaud, C. (2014). "Suggestions for Extending SAIBA with the VIB Platform," in Workshop on Architectures and Standards for IVAs, held at the '14th International Conference on Intelligent Virtual Agents (IVA 2014), Boston, USA, August 26 (Boston, USA: Bielefeld eCollections), 16–20. doi:10.2390/biecoll-wasiva2014-03

Peeters, G., and Czapinski, J. (1990). Positive-negative Asymmetry in Evaluations: The Distinction between Affective and Informational Negativity Effects. *Eur. Rev. Soc. Psychol.* 1 (1), 33–60. doi:10.1080/14792779108401856

Pelachaud, C. (2009). Modelling Multimodal Expression of Emotion in a Virtual Agent. *Phil. Trans. R. Soc. B* 364 (1535), 3539–3548. doi:10.1098/rstb.2009.0186

Pennebaker, J. W. (2011). The Secret Life of Pronouns. *New Scientist* 211 (2828), 42–45. doi:10.1016/s0262-4079(11)62167-2

Raffard, S., Salesse, R. N., Bortolon, C., Bardy, B. G., Henriques, J., Marin, L., et al. (2018). Using Mimicry of Body Movements by a Virtual Agent to Increase Synchronization Behavior and Rapport in Individuals with Schizophrenia. *Sci. Rep.* 8, 1–10. doi:10.1038/s41598-018-35813-6

Ritschel, H., Baur, T., and André, E. (2017). "Adapting a Robot's Linguistic Style Based on Socially-Aware Reinforcement Learning," in Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium, Lisbon, Portugal, 28 Aug.-1 Sept. 2017 (IEEE), 378–384. doi:10.1109/ROMAN.2017.8172330

Rizzo, A., Shilling, R., Forbell, E., Scherer, S., Gratch, J., and Morency, L.-P. (2016). "Autonomous Virtual Human Agents for Healthcare Information Support and Clinical Interviewing," in *Artif. intelligence Behav. Ment. Health Care*. Editor D. D. Luxton (San Diego: Academic Press), 53–79. doi:10.1016/b978-0-12-420248-1.00003-9

Rosenberg, S., Nelson, C., and Vivekananthan, P. S. (1968). A Multidimensional Approach to the Structure of Personality Impressions. *J. Personal. Soc. Psychol.* 9 (4), 283–294. doi:10.1037/h0026086

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B. W., de Sevin, E., Valstar, M. F., and Wöllmer, M. (2015). "Building Autonomous Sensitive Artificial Listeners," in 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21-24 Sept. 2015 (IEEE), 456–462. doi:10.1109/ACII.2015.7344610

Sidner, C. L., Bickmore, T., Nooraie, B., Rich, C., Ring, L., Shayganfar, M., et al. (2018). Creating New Technologies for Companionable Agents to Support Isolated Older Adults. *ACM Trans. Interact. Intell. Syst.* 8 (3), 1–27. doi:10.1145/3213050

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., et al. (2010). "Ada and grace: Toward Realistic and Engaging Virtual Museum Guides," in International Conference on Intelligent Virtual Agents, Philadelphia, PA, USA, September 20-22 (Springer), 286–300. doi:10.1007/978-3-642-15892-6_30

Toma, C. L. (2014). Towards Conceptual Convergence: An Examination of Interpersonal Adaptation. *Commun. Q.* 62 (2), 155–178. doi:10.1080/01463373.2014.890116

van Vugt, H. C., Hoorn, J. F., Konijn, E. A., and de Bie Dimitriadou, A. (2006). Affective Affordances: Improving Interface Character Engagement through Interaction. *Int. J. Human-Computer Stud.* 64 (9), 874–888. doi:10.1016/j.ijhcs.2006.04.008

van Waterschoot, J., Bruijnes, M., Flokstra, J., Reidsma, D., Davison, D., Theune, M., and Heylen, D. (2018). "Flipper 2.0," in International Conference on Intelligent Virtual Agents, November 2018 (Springer), 43–50. doi:10.1145/3267851.3267882

Verberne, F. M. F., Ham, J., Ponnada, A., and Midden, C. J. H. (2013). "Trusting Digital Chameleons: The Effect of Mimicry by a Virtual Social Agent on User Trust," in International Conference on Persuasive Technology, Sydney, NSW, Australia, April 3-5 (Sydney, NSW, Australia: ACM), 234–245. doi:10.1007/978-3-642-37157-8_28

Wang, C., Biancardi, B., Mancini, M., Cafaro, A., Pelachaud, C., Pun, T., and Chanel, G. (2019). "Impression Detection and Management Using an Embodied Conversational Agent," in International Conference on Human-Computer Interaction, Orlando, FL, USA, 26-31 July (Springer), 392–403. doi:10.1007/978-3-030-49062-1_18

Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., and André, E. (2018). "How to Shape the Humor of a Robot-Social Behavior Adaptation Based on Reinforcement Learning," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, October 2018 (Boulder, Colorado: ACM), 154–162. doi:10.1145/3242969.3242976

Wiering, M. A. (2005). "Qv (Lambda)-learning: A New On-Policy Reinforcement Learning Algrithm," in Proceedings of the 7th european workshop on reinforcement learning, 17–18.

Wiggins, J. S. (1979). A Psychological Taxonomy of Trait-Descriptive Terms: The Interpersonal Domain. *J. Personal. Soc. Psychol.* 37 (37), 395–412. doi:10.1037/0022-3514.37.3.395

Wojciszke, B., and Abele, A. E. (2008). The Primacy of Communion over agency and its Reversals in Evaluations. *Eur. J. Soc. Psychol.* 38 (7), 1139–1147. doi:10.1002/ejsp.549

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. (2009). Affect-aware Tutors: Recognising and Responding to Student Affect. *Int. J. Learn. Tech.* 4 (3-4), 129–164. doi:10.1504/ijlt.2009.028804

Yzerbyt, V., Provost, V., and Corneille, O. (2005). Not Competent but Warm... Really? Compensatory Stereotypes in the French-speaking World. *Group Process. Intergroup Relations* 8 (3), 291–308. doi:10.1177/1368430205053944

Zhang, Z., Bickmore, T. W., and Paasche-Orlow, M. K. (2017). Perceived Organizational Affiliation and its Effects on Patient Trust: Role Modeling with Embodied Conversational Agents. *Patient Educ. Couns.* 100 (9), 1730–1737. doi:10.1016/j.pec.2017.03.017

Zhao, R., Sinha, T., Black, A. W., and Cassell, J. (2016). "Socially-aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior," in International Conference on Intelligent Virtual Agents, Los Angeles, CA, USA, September 20-23, 2016 (Springer), 218–233. doi:10.1007/978-3-319-47665-0_20

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.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 08 September 2021
doi: 10.3389/fpsyg.2021.697093

The Symphony of Team Flow in Virtual Teams. Using Artificial Intelligence for Its Recognition and Promotion

Corinna Peifer[1], Anita Pollak[2], Olaf Flak[3], Adrian Pyszka[4], Muhammad Adeel Nisar[5], Muhammad Tausif Irshad[5], Marcin Grzegorzek[5], Bastian Kordyaka[1] and Barbara Kożusznik[2]*

[1] Department of Psychology, University of Lübeck, Lübeck, Germany, [2] Department of Social Science, Institute of Psychology, University of Silesia in Katowice, Katowice, Poland, [3] University of Silesia in Katowice, Katowice, Poland, [4] Department of Human Resource Management, College of Management, University of Economics in Katowice, Katowice, Poland, [5] Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

More and more teams are collaborating virtually across the globe, and the COVID-19 pandemic has further encouraged the dissemination of virtual teamwork. However, there are challenges for virtual teams – such as reduced informal communication – with implications for team effectiveness. Team flow is a concept with high potential for promoting team effectiveness, however its measurement and promotion are challenging. Traditional team flow measurements rely on self-report questionnaires that require interrupting the team process. Approaches in artificial intelligence, i.e., machine learning, offer methods to identify an algorithm based on behavioral and sensor data that is able to identify team flow and its dynamics over time without interrupting the process. Thus, in this article we present an approach to identify team flow in virtual teams, using machine learning methods. First of all, based on a literature review, we provide a model of team flow characteristics, composed of characteristics that are shared with individual flow and characteristics that are unique for team flow. It is argued that those characteristics that are unique for team flow are represented by the concept of collective communication. Based on that, we present physiological and behavioral correlates of team flow which are suitable – but not limited to – being assessed in virtual teams and which can be used as input data for a machine learning system to assess team flow in real time. Finally, we suggest interventions to support team flow that can be implemented in real time, in virtual environments and controlled by artificial intelligence. This article thus contributes to finding indicators and dynamics of team flow in virtual teams, to stimulate future research and to promote team effectiveness.

Keywords: team flow, team effectiveness, virtual teams, machine learning, collective communication

OPEN ACCESS

Edited by:
Ilaria Durosini, European Institute of Oncology (IEO), Italy

Reviewed by:
Rita Berger, University of Barcelona, Spain
Charles Walker, St. Bonaventure University, United States

****Correspondence:***
Corinna Peifer
corinna.peifer@uni-luebeck.de

Specialty section:
This article was submitted to Organizational Psychology, a section of the journal Frontiers in Psychology

Received: *18 April 2021*
Accepted: *26 July 2021*
Published: *08 September 2021*

Citation:
Peifer C, Pollak A, Flak O, Pyszka A, Nisar MA, Irshad MT, Grzegorzek M, Kordyaka B and Kożusznik B (2021) The Symphony of Team Flow in Virtual Teams. Using Artificial Intelligence for Its Recognition and Promotion. Front. Psychol. 12:697093. doi: 10.3389/fpsyg.2021.697093

INTRODUCTION

Advances in information and communication technology (ICT) provided the opportunity for virtual (team-) work and – due to globalization – more and more teams work together virtually over the globe (Jarvenpaa and Leidner, 1999; Raghuram et al., 2019). Organizations have adopted virtual teams for two main reasons. First, virtual teams are related to significant savings, such as reduced costs and time for traveling, and reduced meeting times. Second, virtual teams lead to higher

flexibility, enabling organizations to cope with modern challenges stemming from globalization, competition, changing organizational structures, and increasing service demands (Purvanova, 2014). In addition, the COVID-19 pandemic has boosted the implementation of virtual team work, with many employees working from home using virtual tools to collaborate with their teammates (Feitosa and Salas, 2020).

However, formal and informal interaction is different in virtual teams, including communication among team members and team leadership. For example, reduced informal interaction in virtual teams leads to difficulties to build trust among virtual team members. Trust is however crucial when team members decide to ask each other for help, mutually provide feedback, and address issues and conflicts (Bell and Kozlowski, 2002; Jarman, 2005; Purvanova, 2014). These factors have significant effects on team effectiveness, which per definition includes performance measures as well as team members' satisfaction with their working experience (Gilson et al., 2014; Pyszka, 2015a,b). Accordingly, coping with the particular challenges of virtual team work is essential for virtual teams.

A concept with a high potential for fostering team effectiveness is the concept of team flow (van den Hout et al., 2018). Team flow is a shared experience of flow, characterized by the pleasant feeling of absorption in an optimally challenging activity (Peifer and Engeser, 2021), and of optimal team-interaction during an interdependent task (van den Hout et al., 2018). Research on team flow is still scarce and particularly lacking for virtual teams. Furthermore, conditions of team flow will fluctuate during task completion and, thus, it is important to look for the dynamics of team-flow during the task. Team processes in general are dynamic phenomena but in current research we observe predominantly static treatment of team processes (Kozlowski and Chao, 2012).

However, assessing the dynamics of team flow is a challenge, as traditional measures of team flow are based on self-report questionnaires, which require an interruption of the team process. In order to study the dynamics of team flow during task completion, we thus need to identify a continuous, interruption-free team flow-indicator. Such an indicator can likely be found based on behavioral and sensor data. The evolutions in artificial intelligence and wearable sensor technology made it possible to collect physiological and sensor data and to predict emotional states. In this article, we will thus present an approach to measure team flow in virtual teams using machine learning methods. Based on a literature review, we will present physiological and behavioral characteristics of team flow. We will derive indicators which are suitable for machine learning in order to recognize them in real time. Finally, we will suggest interventions to foster team flow that can be implemented in real time, in virtual environments and controlled by artificial intelligence.

Team Effectiveness and Trust in the Context of Virtual Teams

A team can be defined as "a small number of people with complementary skills who are committed to a common purpose, set of performance goals, and approach for which they hold themselves mutually accountable" (Katzenbach and Smith, 1993, p. 112). In difference to working groups, the performance of teams exceeds the mere sum of individual performances (Katzenbach and Smith, 1993). Furthermore, teams should be understood as complex, multilevel systems that function over time, tasks, and contexts (Pyszka, 2015b). The analysis of existing team effectiveness models shows a large variety of approaches and factors influencing team effectiveness. Input-Process-Output (IPO) models make predictions about the conditions and processes that lead to increased team effectiveness. One of the most popular IPO models has been developed by Hackman (1983) and he subdivides effectiveness into the components: task performance, ability to cooperate in the future, creativity, and satisfaction of team members. In recent years, research investigated team effectiveness in face-to-face compared to virtual teams. It was found that teams using computer mediated communication systems (CMCS) communicate less effectively in many circumstances than teams meeting face-to-face (Warkentin et al., 1997). A review of the literature indicates that the conditions that impact the effectiveness of virtual teams are still ambiguous (Ebrahim et al., 2009; Hertel et al., 2005; Purvanova, 2014). According to Olson and Olson (2006), the effectiveness of virtual teams is exposed to many challenges, including: the nature of work, the common ground of the team members, the competitive/cooperative culture, the level of technology competence of the team members, and the level of technical infrastructure in which the work resides. The most commonly reported challenge in virtual team work is that virtual communication is not an adequate substitute for face-to-face communication (de Guinea et al., 2012) which might lead to a lack of trust in colleagues (Baskerville and Nandhakumar, 2007).

There are different facets of trust, which become visible in at least two different definitions of the phenomenon. Accordingly, trust can be defined as "one's expectations, assumptions, or beliefs about the likelihood that another's future actions will be beneficial, favorable, or at least not detrimental to one's interests" (Robinson, 1996, p. 576). Another definition understands trust as a party's "willingness to be vulnerable to the action of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712; see also Das and Teng, 1998; Man and Roijakkers, 2009).

Due to the lack of informal interaction, lack of knowledge about what others are doing, trust is difficult to build in virtual teams (Olson and Olson, 2006; Breuer et al., 2016). Trust determines whether team members ask each other for help, share feedback, and discuss issues and conflicts (Breuer et al., 2016; de Jong et al., 2016). Therefore, trust has a significant effect on team effectiveness (de Jong et al., 2016). This represents an entirely new paradigm of communication that is needed in *virtual* teams, that must be learned, with little means of social control, with new tools and techniques of social interaction which need to foster familiarity and proficiency (Warkentin et al., 1997).

Effects of Flow on Trust and Team Effectiveness in Virtual Teams

A concept with a high potential for fostering trust and team effectiveness is the concept of team flow

(van den Hout et al., 2018). Previous research already indicated that the concept of flow can be a meaningful antecedent of trust in virtual settings (Bilgihan et al., 2015), whereby the presence of flow increased the perception of trust. A potential explanation is that positive emotions resulting from the experience of flow contribute to building an atmosphere of benevolence in which team members feel good and rightly. Simultaneously, shared positive experiences foster trust in the team's achievement as well as reciprocal stimulation and inspiration. Maintaining such beneficial conditions in the work team over time conveys a sense of safety and stability, as well as dependability and trustworthiness – thereby fostering different facets of trust, i.e., positive expectations toward team members' future actions and low need to control others.

A broad empirical evidence exists for the links between flow and efficiency (for an overview see Peifer and Wolters, 2021). Those links have been confirmed for different efficiency variables, such as increased wellbeing (e.g., Peifer et al., 2020b), work satisfaction (Maeran and Cangiano, 2013), qualitative and quantitative performance (Peifer and Zipp, 2019), in- and extra-role performance (Demerouti, 2006), learning outcomes (Engeser and Rheinberg, 2008), service quality (Kuo and Ho, 2010), and creativity (Zubair and Kamal, 2015). Also, in teams performing complex planning tasks, team flow was found to be positively related to team performance (Heyne et al., 2011). Similarly, a study investigating student teams performing a project management task found that the flow of team members was associated with team performance (Aubé et al., 2014). Such positive associations of team flow with team performance were also found in a video game experiment (Keith et al., 2014) as well as in the work context (van den Hout et al., 2019). In a longitudinal study, students were asked to compose a piece of music and their flow experience during the process was positively related to creativity of the team product as assessed by an expert jury (MacDonald et al., 2006), which provides evidence also for long-term effects of flow on team effectivity.

Components of Team Flow

Team flow needs to be distinguished from individual flow. *Individual flow* is a pleasant experience of being fully absorbed in an optimally challenging task (Csikszentmihalyi, 1975, 1990). Its core characteristics are a high degree of absorption with the task, a perceived demand-skill balance, and enjoyment (Peifer and Engeser, 2021). Building upon the definition of individual flow, team flow has been defined as "a shared experience of flow derived from an optimized team dynamic during the execution of interdependent personal tasks" (van den Hout et al., 2018, p. 400). As a shared experience of flow, team flow shares the characteristics of individual flow, i.e., a high degree of absorption with the task, a perceived demand-skill balance, and enjoyment (Pels et al., 2018; van den Hout et al., 2018; Peifer and Engeser, 2021); but entails additional, team flow-specific characteristics that reflect the social nature of the phenomenon (Pels et al., 2018).

The literature on flow in social situations is yet quite scarce and within this literature, the approaches to the concept vary from individual flow in social contexts to interdependent flow in dyads or teams (Walker, 2021). Terms that can be found in literature are e.g., social flow (Walker, 2010), collective flow (Quinn, 2003, 2005; Šimleša, 2018), group flow (Sawyer, 2003, 2006, 2008), and team flow (van den Hout et al., 2018), to name the most common ones. In the following, when talking about *team flow,* we will refer to those social flow phenomena, which are described as *a shared social experience during a group's interdependent interaction.*

Sawyer (2003) was one of the firsts who proposed a concept of group flow. He emphasized the interdependence of the group as a particular characteristic of group flow as compared to individual flow. With his emphasis on interdependence, Sawyer's understanding of group flow aligns with our understanding of team flow. Sawyer proposed a clear differentiation between group flow and individual flow, claiming that the group can be in flow when the members are not experiencing individual flow and that members can be in individual flow while the group is not in flow (Sawyer, 2008). According to this claim, characteristics of group flow may in part be different from characteristics of individual flow. Based on a qualitative approach (although the details of this study were not published), Sawyer (2008) discussed 10 conditions of group flow: (1) the group's goal (clear vs. open depending on the task), (2) close listening to one another, (3) complete concentration to the group task, (4) being in control of their actions and environment, (5) the ability of group members to merge their egos with the group mind, (6) equal participation of group members, (7) familiarity with group members performance styles, a shared understanding of the group's goals and conventions, and shared tacit knowledge, (8) constant and spontaneous communication, (9) moving the process forward, and (10) the potential for failure.

In his approach, Quinn (2003, 2005) defines "collective flow" as the experience of "moving together toward shared or complementary goals, adjusting in real time to each other's expectations, needs, and contributions, and learning how others work and how to interact effectively along the way" (p. 637). With this definition, he points to the dynamic nature of team flow in team processes, in which team members need to react to each other. Similar to Sawyer (2006), Quinn differentiates between individual flow and team flow (what he calls collective flow) and proposes additional distinct conditions for team flow: (1) the coordination of activities, (2) a collective goal that structures the joint activity, and (3) comparable skill levels. And also, Walker (2010, 2021) differentiates individual flow and team flow (what he calls "interactive social flow"), and as further conditions of team flow he suggests: (1) agreement on goals, procedures, roles, and patterns of interpersonal relations and (2) uniformly high competency of team members (Walker, 2010). In a similar vein, van den Hout et al. (2018) propose that "team flow has similar conditions as individual flow, but teams are subject to additional considerations, specifically team communication, information sharing, and team member perceptions of teammate performance and effort" (p. 400). Based on this, and on previous literature on social flow, van den Hout et al. (2018) proposed a team flow model, with the following antecedents: (1) collective ambition, (2) common goal, (3) aligned personal goals, (4) high skill integration, (5) open communication (6) safety, and

(7) mutual commitment. In an empirical study using a cross-sectional approach, van den Hout et al. (2019) found evidence for their proposed team flow conditions. In their scoping review on group flow, Pels et al. (2018) also distinguished between individual aspects of group flow and collective aspects of group flow. Individual aspects identified in the literature (compare Pels et al., 2018, table 1, p. 6ff) were the individual experience of flow, as well as the flow characteristics enjoyment, demand-skill balance and absorption (including feeling one with the group). These individual aspects are in line with the core characteristics of flow according to Peifer and Engeser (2021). In their summary of empirical findings on group flow, they further list "aspects of competence (e.g., knowing others' skills; Kaye and Bryce, 2012), interaction (e.g., effective communication; Kaye, 2016), and of positive relationships (e.g., trust within the group; Armstrong, 2008)" as antecedents of group flow. Other identified aspects within the definitions provided in the literature (compare Pels et al., 2018, table 1, p. 6ff) relate to the aspect of common goals such as "purposeful communication" (Duff et al., 2014), "concurrent engagement in a shared goal-oriented activity" (Hart and Di Blasi, 2015), or "common focus" (Kaye and Bryce, 2014). Also, reported aspects of group flow according to Pels et al. (2018, table 1, p. 6ff) are that of interactional synchrony (Zumeta et al., 2016) and social contagion (Bakker et al., 2011; Aubé et al., 2014).

Despite the variety of the proposed terms, characteristics and conditions of team flow, the characteristics and conditions of team flow show commonalities on a level of content. As can be seen in **Table 1**, the conditions of team flow as described by the just referenced authors can be summarized into four major categories: (1) communication and feedback, (2) goal commitment, (3) equal participation, and (4) trust.

Importantly though, we need to distinguish conditions from characteristics. Also for individual flow, there has been a long discussion about which of the flow characteristics are conditions, core components or outcomes (see e.g., Landhäußer and Keller, 2012). For some core components, such as the challenge-skill balance (or: demand-skill balance), some authors argue it is a condition, others count it as a component (Landhäußer and Keller, 2012). In the meantime, there is some agreement in the literature that the objective presence of a demand-skill balance is a condition of flow, while the perceived demand-skill balance is defined as a component (Peifer and Engeser, 2021). As the

TABLE 1 | Meta-categories of team flow-specific characteristics.

Meta-category	Sawyer (2003, 2008)	Quinn (2003, 2005)	Walker (2010)	van den Hout et al. (2018)	Pels et al. (2018)
Communication and feedback	Close listening to one another, constant and spontaneous communication	The coordination of activities	Agreement on goals, procedures, roles, and patterns of interpersonal relations	Open communication	Interaction, e.g., effective communication; fluent, positive interactions within the group (p. 18)
Shared goal commitment	A shared understanding of the group's goal, complete concentration to the group task, the ability of group members to submerge their egos to the group mind, shared understanding of the group's goals and conventions, and shared tacit knowledge	A collective goal that structures the joint activity	Agreement on goals, procedures, roles, and patterns of interpersonal relations	Collective ambition, common goal, aligned personal goals, and mutual commitment	Common goals, e.g., purposeful communication, concurrent engagement in a shared goal-oriented activity, or common focus
Equal participation	Equal participation of group members	Comparable skill levels	Uniformly high competency of team members	High skill integration	Interactional synchrony; contagion effect; a shared state of balance; a high collective competence
Trust in each other's knowledge, skills, and attitudes	Being in control of their actions and environment, familiarity with group members performance styles, a shared understanding of the group's goals and conventions, and shared tacit knowledge	Comparable skill levels	Uniformly high competency of team members	A shared belief that the team is safe; mutual trust as characterized by: (a) a willingness to be vulnerable, (b) mutual respect, (c) confidence in the working environment, and (d) team potency/efficacy (p. 410)	Positive relationships, trust within the group; aspects of competence, knowing each other's skills

Some of the components are included in more than one category.

just described team flow conditions are unique for team flow as compared to individual flow, we argue that the perception of their presence can be regarded as a component of team flow.

Taken together, team flow is composed of those characteristics that are shared with individual flow combined with characteristics that are unique for team flow. The resulting components are listed in **Table 2**.

While the literature on flow in social contexts (and on team flow specifically) is scarce, literature on team flow in virtual contexts barely exists, although virtual teams are increasingly important in today's workplaces. Compared to real-world settings, we argue that achieving the outlined team flow characteristics: (1) communication and feedback, (2) shared goal commitment, (3) equal participation, and (4) trust are particularly challenging in virtual contexts, due to the lack of informal communication. Accordingly, reaching team flow in virtual teams should be more difficult than in face-to-face contexts. Even more so it is necessary to identify indicators that can measure team flow in virtual teams in order to find and evaluate approaches to promote team flow in virtual teams. To find such indicators, we should measure the presence of the characteristics of team flow in real time during the team process. A concept, which could help finding such indicators as it is largely overlapping with the specific team flow conditions, is the concept of collective communication.

Collective Communication

According to Watzlawick and Beavin (1967), all behavior in the presence of another person is communicative (with presence going beyond physical presence). Thereby communication is more than verbal productions but includes all behavior in the social context (Watzlawick and Beavin, 1967). Communication is further based on individual characteristics like openness, and it can for example be direct and indirect, verbal and non-verbal, and as such it can be measured by different indicators like communication style or listening ability, and many others (Hargie and Tourish, 2000).

The term "*collective communication*" is a particular group-related communication style, which refers to a group's (or team's) behavior and does not necessarily correlate with individual communication (Kozusznik et al., 2018), although it may vary from the sum of individual factors (Woolley et al., 2010). Collective communication can be described as the connections among people, feedbacks, and interrelations (Weick and Roberts, 1993). It refers to the bundle of messages from all group members given at the same time or otherwise in the form of feedback. Each message provides direct or indirect feedback to the team members. Non-verbal social feedback can be derived from smiles, attention, tone of voice, or other social cues. They usually represent spontaneous reactions without intentions of teaching or otherwise influencing. Besides, it can be found that it does not induce additional cognitive loads (Knox, 2012). Collective communication appears when group participants have equal chances to participate and communicate in the discussion (Woolley et al., 2010). It includes specific forms of communication that guide and prioritize activities within the team while maintaining all its members' equality and well-being. This ensures a spontaneous and expressive response in a safe and comfortable environment for the individual, allowing for convenient speech without fear of losing one's meaning (Kożusznik, 2005).

It was found that collective communication strengthens the team members' process control and improves knowing each other and mutual understanding (Riedl and Woolley, 2017). Operationally, it allows to reduce the redundancy of statements, thoughts, and actions, eliminates delays, facilitates the use of participants' knowledge, and provides people crucial information and opportunities to perform. Specific indicators of collective communication are the motivational drivers of participation in the communication. Existing research confirms that equal communication gives rise to an equal rhythm of communication that can predict group performance on a wide variety of tasks (Woolley et al., 2010). Interrupting others is correlated with domineering (Kożusznik, 2005; Woolley et al., 2010). When one team member interrupts the speaker, it causes a decrease in effectiveness and impairs the speaker's well-being (Bouskila-Yam and Kluger, 2011). The temporal patterning of activities is an important aspect of team effectiveness (McGrath, 1984). Alternating interaction is an orderly process of verbal and non-verbal activities which help regulate the flow of conversation, enable turn-taking, and provide feedback. Additionally, rather than a randomly distributed communication pattern, there tend to be periods of high activity (bursts) of the group followed by periods of little activity that enhance team flow (Riedl and Woolley, 2017). Questioning is also related to the effectiveness of the group performance (Bouskila-Yam and Kluger, 2011), as it leads to increased mutual understanding, improved team coordination and more clarity in the work process.

COMPLEMENTING A TEAM FLOW MEASURE BY MEANS OF COLLECTIVE COMMUNICATION

When comparing the characteristics of team flow with the concept of collective communication, their strong overlap becomes evident, as shown in **Table 3**.

Accordingly, the concept of collective communication represents the team flow characteristics and it can be used to operationalize team flow indicators. Indicators of collective communication have already been identified, which can now be used to complement a measure of team flow using machine learning.

Collective communication correlates with a set of behavioral markers: equal distribution of conversational turn-taking, the number of speaking turns, silence, voice volume, number of interruptions, facilitating listening, space offering, "I" and "We" and burstiness (Kożusznik, 2005; Kluger and Nir, 2009; Woolley et al., 2010). Also, measures like time of speaking can measure it, as well as the number of questions, speed of talking, and others (see **Table 5**).

TABLE 2 | Components of team flow.

Component	Shared between individual and team flow	Unique for team flow
Absorption	X	
Perceived demand-skill balance	X	
Enjoyment	X	
Communication and feedback		X
Shared goal commitment		X
Equal participation		X
Trust in each other's knowledge, skills, and attitudes		X

TABLE 3 | Overlaps between collective communication and team flow.

Team flow characteristic	Characteristic of collective communication
Communication and feedback	Stems from the connections among people, feedback, and interrelations; ensures a spontaneous and expressive response[1]
Goal commitment	Includes specific forms of communication that guide and prioritize activities within the team[2]
Equal participation	Appears when group participants have equal chances to participate and communicate in the discussion[3]
Trust in each other's knowledge, skills, and attitudes	Takes place in a safe and comfortable environment for the individual, allowing for convenient speech without fear of losing one's meaning[4]

[1]*Weick and Roberts, 1993*; [2]*McGrath, 1984*; *van Dyne et al., 2003*; *Bies, 2009*; *Ross, 2014*; [3]*Weick and Roberts, 1993*; *Kożusznik, 2005*; *Woolley et al., 2010*; [4]*Hietanen et al., 1998*; *Bouskila-Yam and Kluger, 2011*.

Team Flow-Indicators Suitable for Machine Learning

As described above, team flow is composed of those characteristics that are shared with individual flow combined with characteristics that are unique for team flow (compare **Table 2**). Accordingly, a team flow measure should be operationalized based on all these characteristics.

For individual flow – i.e., for those team flow characteristics that are shared with individual flow – there are already elaborated concepts and studies on its physiological correlates (for an overview see Tozman and Peifer, 2016; Peifer and Tan, 2021) and their potential use for machine learning (Peifer et al., 2020a; Rissler et al., 2020). Studies show that individual flow experience is for example associated with heart rate variability (Peifer et al., 2014), electrodermal activity (de Manzano et al., 2010), respiration (de Manzano et al., 2010), blinking rate (Rau et al., 2017; Peifer et al., 2019a), or facial muscle activation (Kivikangas, 2006; de Manzano et al., 2010; Nacke and Lindley, 2010). Those indicators can be sorted according to the components of flow, i.e., if they relate to absorption, perceived demand-skill balance and/or enjoyment. In **Table 4** we propose physiological measures that can be used to assess individual flow in real time and which are suitable for machine learning.

In order to complement the measurement of the just described team flow components, we need to include also indicators for those characteristics, that are unique for team flow, i.e., (1) communication and feedback, (2) shared goal commitment, (3) equal participation, and (4) trust. As discussed, this can be reached using the concept of collective communication, as behavioral measures of collective communication already exist (**Table 5**). An advantage of their measurement in virtual teams is, that they can be assessed using the video camera and the audio signal. Such indicators of collective communication are: (a) equal communication (b) number of speaking turns; (c) interruptions; (d) facilitating listening; (e) number of the use of We and I, (f) burstiness; (g) vocal expression/melody of team voice; (h) silence; and (i) space offering. The detailed definitions and measurement of those proposed indicators is described below and also presented in **Table 5**.

For measuring *equal communication* each subject can be assessed to gain information about "equal" vs. "unequal," i.e., each group participant will be compared to the overall discussion time. Measurement of the *number of speaking turns* requires constant monitoring and counting changes in the course of discussion. Similarly, *interruptions* can be calculated during permanent monitoring with counting the number of times when a person interrupts another person and starts his/her speech. Collecting samples of *facilitating listening* can be achieved by recognizing the presence of particular actions such as questioning and focusing the current speaker. The *use of I/We* can be assessed by monitoring and counting the numbers of "We" and "I" in whole and separate parts of the communication process. For *burstiness* (or: liveliness), individual and continuous estimation will be used, and the measurement of each participant will be compared with results of other team members. *Vocal expression/the melody of the voice* can be measured via differences in average behavior, e.g., in speed of talking, voice pitch, and volume of voice. *Silence* requires continuous estimation and determination of its duration to compare the outcomes of different individuals.

Using Machine Learning to Develop an Application-Based Algorithm for Team Flow Analysis

Machine learning approaches have been widely practiced in recent times by using data from video, motion, and physiological sensors for the recognition of physical and cognitive activities in the field of medical data science (Irshad et al., 2020; Nisar et al., 2020). Therefore, based on the indicators described in **Tables 4**, **5**, it is possible to develop an application for team flow analysis in the form of an end-to-end machine learning-based algorithm that takes input from multiple sensors including wearable devices, cameras and microphones, and predict the cognitive states of the participants of virtual teams by analyzing not only their own physiological data

TABLE 4 | Physiological measures of flow experience applicable for machine learning as part of a team flow algorithm (compare Peifer and Engeser, 2021; Peifer and Tan, 2021).

(Team-) flow indicator	Definition	Physiological measures	IT tools and instruments
Absorption	An immersive feeling of effortless concentration on the task at hand, characterized by the centering of attention, while irrelevant information, including self-referential thoughts, are shielded from attention. Accordingly, physiological correlates of attention, mental effort, and self-referential thoughts are candidates for machine learning indicators of absorption.	Frequency of spontaneous eye blinks per minute (Peifer et al., 2019b)	EMG, eye tracker, or smart glasses
		Eye movements (Foy and Chapman, 2018)	EMG, eye tracker, or smart glasses
		Alpha and theta activity in the brain (Katahira et al., 2018)	EEG
		Heart rate variability (Thayer et al., 2009; Keller et al., 2011; Peifer et al., 2014)	ECG; wristband and smartwatch
		Head movement (Mittal et al., 2016)	Gyroscope
		Body movements (Tang and Zeng, 2009)	Accelerometer
Perceived demand-skill balance	The perception that the demands of the task are in balance with the skills and resources of the individual, which goes along with neither boredom, nor overload, but just the right degree of activation. Accordingly, physiological correlates of arousal and mental effort are candidates for machine learning indicators of a perceived demand-skill balance.	Heart rate (Azarbarzin et al., 2014)	ECG/wristband/ smartwatch
		Heart rate variability (Keller et al., 2011)	ECG; wristband and smartwatch
		Skin temperature (Or and Duffy, 2007; Marinescu et al., 2018)	Wristband
		Electrodermal activity (Boucsein, 1992; Dawson et al., 2007)	EDA sensors
		Respiration (de Manzano et al., 2010; Lean and Shan, 2012)	Chestbelt
		Voice pitch (Johannes et al., 2007)	Microphone
Enjoyment	The perception of an inherent pleasure and satisfaction during the task, which is associated with positive affect and intrinsic motivation. Accordingly, physiological correlates of positive affect, liking, and intrinsic motivation are candidates for machine learning indicators of perceived enjoyment.	Facial muscle activation (Kivikangas, 2006; Nacke and Lindley, 2010)	EMG and camera
		Pupil width (Bradley et al., 2017)	EMG, eye tracker, and smart glasses

but also their interaction and communication with other team members. In order to build such an application, it is necessary to answer several questions like how to handle the heteroscedasticity of different input signals (i.e., the variability of variance of errors of the input data), what will be the useful features (Amjad et al., 2021), which feature extraction and classifications techniques will be used? The artificial neural networks handle all these issues and are able to learn and model non-linear and complex relationships between input and output (Li et al., 2018, 2020). Therefore, end-to-end machine learning algorithms based on artificial neural networks can learn the whole path between the raw sensory data and the final outcomes, for example, the cognitive state and interactions of the participants of virtual teams. However, once the models are trained, it is hard to semantically understand the very complex processing between the input and the output. So, a deep neural network behaves like a black box that does not explain the reasons for particular outcomes. To explain and interpret the transformation between the input data vector and the output cognitive states, we must investigate the application of multi-task deep neural networks sharing some hidden layers and training others specifically for certain environmental and social constellations (e.g., team roles, time working as a team), certain tasks or task characteristics, and certain groups of people similar to each other in terms of their individual differences (e.g., personality, experience, age, gender, etc.). If we manage to explain the predictive decisions of our machine learning algorithms, we will generate new scientific findings in the area of cognitive state analysis. In other words, we will not only be able to implement our deep learning models as a black box, but we will also be able to describe the distinctive features found in the sensor data which are specific for team flow.

Another potential challenge is the situation that people could be too diverse in terms of individual differences or team roles (Driskell et al., 2017) – for one single machine learning approach analyzing their cognitive state and the interaction with their colleagues. It might come out that there are some different types (clusters) of people in the virtual teams having similar properties so that one machine learning configuration works better for one cluster, another for another. It would not be efficient to develop completely isolated machine learning systems for all clusters separately from each other. The machine learning system incorporating the parametric models for different clusters can be helpful to avoid the completely independent systems for each cluster type. When assigning concrete values to the parameters, the generic system can be converted into a concrete system for a certain cluster or team roles. Estimating these

TABLE 5 | Definitions and measurement of collective communication enhancing team flow.

Collective communication indicator	Definition	Measurement	IT tools and instruments
Equal communication	A reciprocal process formed by absence of domination and equality of participation in discussion. It indicates that each participant of the discussion is entitled to the same or equal period to speak. Furthermore, individuals flexibly exchange messages as the sender and recipient due to "heedful interrelating" (Weick and Roberts, 1993). It stems from understanding the consequences of the individuals' relations and the flexibility of their behavior patterns for the team's effectiveness (Woolley et al., 2010)	(1) The recorded time of individual speaking; (2) The proportion of being an active participant in the conversation.	Microphone
Number of speaking turns	Situational characteristics or events that influence the occurrence of behavioral reactions form the rhythm of conversation. It serves a diversity of exchange information (Woolley et al., 2010).	(1) The number of times the individual was speaking during the discussion.	Microphone
Interruptions	The action of interfering by asking or giving comments, or the process of being interfered by an individual or others. It alters the process of communication and causes changes in motivational-affective state or even in human behavior (Bouskila-Yam and Kluger, 2011).	(1) The number of times a person interrupts another person and starts his/her speech.	Microphone
Facilitating listening	The individual factor of personal style of communicating is relevant to making a conscious effort to consider another person's position, especially by asking questions or observing his/her behavioral reactions (Bouskila-Yam and Kluger, 2011).	(1) The number of questions asked during the conversation; (2) The number of looking into eyes of speaking person; (3) The length of time of looking into the eyes of the speaking person.	Microphone/smart glasses/eye tracker
Using I/We	The use of a specific pronoun ("We" vs. "I") results from the currently experienced state, conceptualized as a positive fulfilling which relates to action or reaction to the situation. Choosing "We" vs. "I" helps emphasize an individual or group perspective (Torrente et al., 2013).	(1) The number of "We" used in the communication process; (2) The number of "I" used in the communication process.	Microphone
Burstiness	The temporal patterning of activities or synchronous interaction is an orderly process of verbal and non-verbal activities. It helps monitor and regulate the flow of conversation, enable turn-taking, and provide feedback (Riedl and Woolley, 2017).	The number of "bursty" signals such as: (1) The number of moments a person laughs; (2) The number of moments a person cries; (3) The number of times in which voices are overlapping (4) The number of times in which the loudness of the voices changes.	Microphone
Vocal expression/melody of voice	A form of voice cues helps communicate emotions and infer other people's emotions in everyday life (Wallbott and Scherer, 1986; Zebrowitz, 1990; Sundberg, 1998).	Making differences to the average levels in: (1) Strength of/volume of voice; (2) Speed of talking; (3) Voice pitch (low and high tones of voice).	Microphone
Silence	A multidimensional construct which is characterized as absence of voice or speaking up. It can be hidden or disused in the act of voice (silence is more than the absence of noise). It has two functions: one positive that can improve problem solving or learning and one dysfunctional that undermines the interests of organizations and influences relationships (van Dyne et al., 2003; Bies, 2009; Ross, 2014).	(1) The number of moments of silence in the middle of a discussion; (2) The length of silence episodes; (3) The ratio of the length of episodes with verbal communications to the length of silence episodes (no verbal expression).	Microphone
Space offering	Ability, willingness, and skills of the team member to support other members to take active participation in the team discussion (Kożusznik, 2005).	(1) The number of questions (2) The length of individual statements (3) The length of silence after the statements	Microphone

meta-parameters can be realized by a supervised machine learning approach.

To train the algorithm, subjective measures of team flow will be needed, potentially complemented by observer ratings. Even after an algorithm has been identified, such subjective measures should be used at least intermittently to complement objective data and to validate and improve the algorithm.

INTERVENTIONS TO FOSTER FLOW IN VIRTUAL TEAMS

A machine learning system that identifies team flow can help not only to measure, but also to promote team flow and its dynamics over time. This means, the machine learning system can be used as a decision support system, that can identify team processes

as fostering or hindering for team flow and provide feedback if team processes deviate significantly from their optimal level. The level of support by the machine learning system can vary along a continuum of low to high support starting with information and feedback only, up to the proposal of suitable interventions (Peifer et al., 2020a). Also, the decision authority may range from full authority by the team members/team leader up to full authority by the system (Parasuraman et al., 2000), e.g., the system can mute speakers or send automated instructions on interventions or on how to proceed.

What Could Real-Time Interventions Look Like?

The specific intervention depends on which team flow indicator deviates from its optimal level. In the case that indicators reflecting *absorption* deviate, the machine learning system could propose a pause or a meditation in order to regain energy and focus (Peifer and Tan, 2021). A deviation of the *perceived demand-skill balance* can be improved by re-defining individual goals, or social support. If *enjoyment* is lacking, the machine learning system could propose team interventions that foster positive emotions, such as providing compliments to each other, or the use of humor. A huge selection of potential interventions can be found in the field of positive psychology (Meyers et al., 2013).

If indicators reflecting *communication and feedback* deviate, e.g., the number of questions is small, or there is a high degree of silence, the machine learning system invites the discussant to ask questions. In instances of escaping (e.g., the listener is busy with her phone or computer), the system could remind the respective team member to join the team process.

If *shared goal commitment* is not given, e.g., as reflected by the use of "I" dominating the use of "We," the machine learning system can invite participant(s) to use "We" instead of "I." Using "We" strengthens the collective communication which resides in the connections between the units and the flexibility of their patterns of behavior (Weick and Roberts, 1993). The pronoun "I" represents an individual approach to measuring work engagement, the pronoun "We" represents the collective engagement conceptualized as a positive, fulfilling, work-related state shared with vigor and dedication (Schaufeli et al., 2002; Torrente et al., 2013).

If *equal participation* is not given – e.g., one person dominates and consumes too much time in the conversation – the system could inform this person (or the manager) and ask to shorten the time of speaking. When the system identifies that the number of speaking turns is small, it invites participants to make another round of conversation. In case of too many interruptions and talking into each other statement's, the system asks to give some time and possibilities to others to let them talk.

If indicators of *trust* deviate, e.g., as reflected by the number of "bursty" signals (lough, cry, overlapping of the voices, and loudness of the voices), machine learning system could inform participants to stop or reduce these signals. Burstiness is related to levels of interpersonal synchrony or temporal coordination and influences trust (KoŻusznik and Polak, 2016) and effectiveness (McGrath, 1984). Also, e.g., if the machine learning system detects that the voice is too loud it can inform the participants and invite them to decrease the volume. If the machine learning system detects too low or too fast speed of talking it can ask the participant to regulate the speech. Similarly, the machine learning system could analyse the voice pitch and monotony and invite the participants to make the voice lower or more vivid. There is a correlation between group vocal expression and trust and well-being among its members (Hietanen et al., 1998). Also larger scale interventions to improve trust can be imagined (such as team building interventions), if a team continues to show signs of low trust among each other.

DISCUSSION

In this article we presented an approach to measuring team flow in virtual teams using machine learning methods. To provide a basis for the development of suitable data for the machine learning algorithm, we have disentangled the characteristics of team flow. We suggested that team flow is composed of characteristics that are shared with individual flow – i.e., absorption, perceived demand-skill balance and enjoyment – , and characteristics that are unique to team flow – i.e., communication and feedback, shared goal commitment, equal participation, and trust. Based on these characteristics and existing research on flow and collective communication, we have identified physiological and behavioral indicators that are suitable as machine learning input data. Furthermore, we have outlined how these data can be used in machine learning to develop an algorithm that assesses team flow in real time. Also, we have identified potential challenges in this endeavor. Finally, we have suggested how the real-time measurement of flow can result in interventions to improve team flow during the team process. In the following, we are discussing the underlying theoretical approach, as well as implications of our approach for research and practice.

Underlying Theoretical Approach

To clarify the uniqueness of team flow, we chose a relational approach. The theory of relational models describes four fundamental forms of social relationships: communal sharing, authority ranking, equality matching, and cost benefit analysis (market pricing) (Fiske, 1991; Fiske and Haslam, 2005). Our assumptions are best supported by equality matching, which builds the basis for turn-taking, equal rights, even sharing, voting, and balanced reciprocity, as well as enabling people to return the same kind of thing they received (Haslam, 2004). Collective communication appears when group participants have equal chances to participate and communicate in the discussion (Woolley et al., 2010). It includes specific forms of communication that guide and prioritize activities within the team while maintaining all its members' equality and well-being. This ensures a spontaneous and expressive response in a safe and comfortable environment for the individual, allowing for convenient speech without fear of losing one's meaning (Kożusznik, 2005). The relational approach also allows us to capture the dynamics of team flow and, thus, to

determine optimal conditions of the collective communication that enhances team flow in a virtual team.

However, also a motivational approach is relevant in the context of team flow, as the enjoyment of the task at hand, including the resulting intrinsic motivation, is a shared characteristic of flow and team flow. This also applies to the characteristics "absorption" and "perceived demand-skill balance," which can be attributed to a cognitive approach. Accordingly, relational, motivational as well as cognitive indicators of team flow have been proposed as part of the suggested team flow measure and also as parts of potential interventions.

Implications for Future Research

We proposed a machine learning system that employs multimodal sensory data to measure team flow in virtual teams. The heteroscedastic input data provided to the machine learning system not only cover physiological data of the members of a virtual team but also consider the vital key aspects of their communication with the co-members. The main objective of the machine learning system is indeed to detect the team flow using all sensory inputs, however, it will also be interesting in the future to see the impact of each kind of input signals in recognizing the team flow. The effectiveness of the machine learning system confides in detecting team flow with good accuracy as well as identifying the distinctive features of the input data that are specific for team flow. Explainable artificial intelligence administers the techniques which highlight the meaningful distinctive features in achieving the desired outcome of the system. A major advantage of an AI-based analysis of team flow is that team members do not need to be interrupted to measure team flow, which allows to assess the process with its fluctuations over time. Also, the automated analysis allows a more objective investigation of the context that can add to traditional self-report measures. In combination, self-report and machine learning data will allow to find a larger, more holistic model of team flow. Accordingly, the machine learning system can serve the function of a real laboratory and help to better understand the concept of team flow, the fluctuations of team flow over time and conditions that promote or hinder team flow.

A still unanswered question in flow research is the relationship between individual flow and team flow (van den Hout et al., 2018; Walker, 2021). Currently, there is no final agreement regarding the relationship of both flow forms of flow and their dynamics over time. The suggested machine learning system provides the opportunity to gain insights into their interplay by holistically relating flow and team flow indicators and sensor data. This will contribute substantial new information to the debate regarding the interplay between individual and social flow.

Also, the effects of different context factors can be studied in more detail. For example, it is well-documented that tasks have differential effects on the team process in terms of losses and gains, and, as a consequence, on team performance. Some tasks were for example found to facilitate social loafing, beyond them additive tasks (Kerr and Bruun, 1981), and easy tasks, while difficult tasks lead to increased performance (Jackson and Williams, 1985). Furthermore, group think was rather found in judgment tasks and less so in intellective tasks (McGrath, 1984). Using machine learning methods to detect team flow and team flow dynamics in different types of tasks could provide deeper information about which tasks are particularly flow-promoting or flow-hindering, about the mechanisms that are responsible, and also which characteristics of the flow experience are particularly affected (e.g., equal participation, goal commitment, perceived demand-skill balance, etc.). This helps us to achieve a better understanding of how to design tasks to achieve team flow and increase team efficiency. Also, this knowledge can be used to propose changes in the type of task in order to stimulate team flow in ongoing team processes. Future studies should thus systematically investigate task types and task characteristics by controlling for the type of task and/or systematically varying different tasks and task characteristics.

Another important domain that affects team processes relates to formal and informal social roles and relationships within the team (Belbin, 1981; Driskell et al., 2017). Teams mostly consist of a team manager and team members, with certain individual characteristics as well as formal and informal roles (Belbin, 1981). Team members including the team manager depend on each other, activated by managerial actions as a constellation of specific objectives, resources and processes (Sohmen, 2013). Also in virtual teams, challenges relate to difficulties in team leadership and the coordination of the team members' activities (Pinjani and Palvia, 2013). Accordingly, the investigation of team constellations, in terms of team leadership style (e.g., teams managed by a leader vs. self-managed teams) and in terms of team roles and individual characteristics, are further relevant questions that should be systematically addressed in future research on team flow. Furthermore, there is likely an interplay between individual characteristics, formal and informal team roles, task characteristics, and the time working together as a team which is worth investigating. By means of the machine learning system, it will be possible to answer more complex questions about why some teams are more effective than others in the future. Corresponding findings can inform more advanced versions of the machine learning system using tools from user-centered design to differentiate between different groups of users within a team based on task characteristics, the team constellation, and the team members' individual characteristics (Pyszka, 2015b).

Implications for Practice

A machine learning system measuring team flow can be used in practice to identify team processes that are promoting or hindering team flow and to derive suitable interventions during the team process. This is even more relevant for practice, as team flow is highly related to team effectiveness – including team performance and team satisfaction (Peifer and Wolters, 2021). Such a machine learning system could complement existing online management tools (e.g., TransistorsHead.com) that are already used to record team members' actions in virtual environments (Flak, 2013, 2019; Flak et al., 2017) and could be incorporated in a more holistic, artificial team management tool (Flak, 2020). It was found that declarations of team management processes based on memory are highly imprecise and subjective

as compared to the objective parameters recorded by online management tools (Flak and Pyszka, 2013). Accordingly, the implementation of artificial team management tools has the potential to provide more objective feedback, more objective decision criteria and more suitable interventions to the team. Moreover, improving skills related to collective communication through the implementation of artificial team management could contribute to enhance relational links and information exchange in teams, as well as buffer the impact of personality and team role diversity.

The development and implementation of a machine learning system comes with substantial set-up costs. However, we follow Pyszka's (2015a) argumentation and understand effectiveness in an evolutionary manner assuming that a change from economic efficiency assessment today toward the evaluation of the potential of solutions will enable even higher levels of effectiveness in the long-term.

Accordingly, the implementation of a machine learning system promises added value for organizations: the machine learning system can lead to higher levels of employee satisfaction, having a positive influence on productivity of an organization (Harter et al., 2002). The machine learning system could be developed even further and integrate additional sources of data (such as characteristics and preferences of employees) toward a holistic system of organizational team management. Due to its innovativeness and low dissemination the implementation of such a machine learning system, it promises competitive advantages over other competitors based on the opportunity for better teamwork, which can improve the process efficiency of an organization in an innovative and unique manner leading to advantageous market positions in the future (Dean, 2014).

CONCLUSION

This article proposed an approach to measure and ultimately promote dynamic, not static, team functioning in virtual teams using machine learning methods. For this, the concept of team flow is a promising target state with high significance for team effectiveness. The concept of collective communication can provide suitable indicators of team flow specific characteristics, which can be used to complement a machine learning algorithm. Such an algorithm can then be used to not only identify, but also promote team flow, by providing feedback to the users and proposing interventions as part of an automated team management system.

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

CP, BKoż, APo, OF, APy, MN, MI, and MG conceived of the presented idea. All authors developed the theory and wrote parts of the manuscript, based on their respective expertise in the relevant concepts, i.e., (team) flow (CP and BKor), communication (BKoż and APo), management (OF and APy), and machine learning (MG, MN, and MI). CP supervised the writing process. All authors discussed the article and contributed to the final manuscript.

REFERENCES

Amjad, F., Khan, M. H., Nisar, M. A., Farid, M. S., and Grzegorzek, M. (2021). A Comparative Study of Feature Selection Approaches for Human Activity Recognition Using Multimodal Sensory Data. *Sensors* 21:2368. doi: 10.3390/s21072368

Armstrong, A. C. (2008). The fragility of group flow: The experiences of two small groups in a middle school mathematics classroom. *J. Mathemat. Behav.* 27, 101–115. doi: 10.1016/j.jmathb.2008.08.001

Aubé, C., Brunelle, E., and Rousseau, V. (2014). Flow experience and team performance: The role of team goal commitment and information exchange. *Motiv. Emot.* 38, 120–130. doi: 10.1007/s11031-013-9365-2

Azarbarzin, A., Ostrowski, M., Hanly, P., and Younes, M. (2014). Relationship between arousal intensity and heart rate response to arousal. *Sleep* 37, 645–653. doi: 10.5665/sleep.3560

Bakker, A. B., Oerlemans, W., Demerouti, E., Slot, B. B., and Ali, D. K. (2011). Flow and performance: A study among talented Dutch soccer players. *Psychol. Sport Exerc.* 12, 442–450. doi: 10.1016/j.psychsport.2011.02.003

Baskerville, R., and Nandhakumar, J., (2007). Activating and perpetuating virtual teams: Now that we're mobile, where do we go? *IEEE Transact. Prof. Comm.* 50, 17–34. doi: 10.1109/TPC.2006.890849

Bell, B. S., and Kozlowski, S. W. J. (2002). A typology of virtual teams: Implications for effective leadership. *Group Org. Manag.* 27, 14–49. doi: 10.1177/1059601102027001003

Belbin, R. M. (1981). *Management Teams: Why They Succeed or Fail*. Oxford: Butterworth-Heinemann.

Bies, R. J. (2009). "Sounds of silence: Identifying new motives and behaviors," in *Voice and Silence in Organizations*, eds J. Greenberg and M. S. Edwards (Bingley: Emerald), 157–171.

Bilgihan, A., Nusair, K., Okumus, F., and Cobanoglu, C. (2015). Applying flow theory to booking experiences: An integrated model in an online service context. *Inform. Manag.* 52, 668–678. doi: 10.1016/j.im.2015.05.005

Boucsein, W. (1992). *Electrodermal Activity. The Springer Series in Behavioral Psychophysiology and Medicine Ser*. New York, NY: Springer.

Bouskila-Yam, O., and Kluger, A. N. (2011). Strength-based performance appraisal and goal setting. *Hum. Resour. Manag. Rev.* 21, 137–147. doi: 10.1016/j.hrmr.2010.09.001

Bradley, M. M., Sapigao, R. G., and Lang, P. J. (2017). Sympathetic ANS modulation of pupil diameter in emotional scene perception: Effects of hedonic content, brightness, and contrast. *Psychophysiology* 54, 1419–1435. doi: 10.1111/psyp.12890

Breuer, C., Hüffmeier, J., and Hertel, G. (2016). Does trust matter more in virtual teams? A meta-analysis of trust and team effectiveness considering virtuality and documentation as moderators. *J. Appl. Psychol.* 101, 1151–1177. doi: 10.1037/apl0000113

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety*. San Francisco: Jossey-Bass Publishers, 150–162.

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

Das, T. K., and Teng, B. -S. (1998). Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances. *Acad. Manag. Rev.* 23, 491–512. doi: 10.5465/amr.1998.926623

Dawson, M. E., Schell, A. M., Filion, D. L., and Berntson, G. G. (2007). "The Electrodermal System," in *Handbook of Psychophysiology*, eds G. G. Berntson, J. T. Cacioppo, and L. G. Tassinary (Cambridge, MA: Cambridge University Press), 157–181. doi: 10.1017/CBO9780511546396.007

de Guinea, A. O., Webster, J., and Staples, S. D. (2012). A meta-analysis of the consequences of virtualness on team functioning. *Inform. Manag.* 49, 301–308. doi: 10.1016/j.im.2012.08.003

de Jong, B. A., Dirks, K. T., and Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *J. Appl. Psychol.* 101, 1134–1150. doi: 10.1037/apl0000110

de Manzano, Ö, Theorell, T., Harmat, L., and Ullén, F. (2010). The Psychophysiology of Flow During Piano Playing. *Emotion* 10, 301–311. doi: 10.1037/a0018432

Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. Hoboken, NJ: John Wiley & Sons, Ltd.

Demerouti, E. (2006). Job characteristics, flow, and performance: The moderating role of conscientiousness. *J. Occupat. Health Psychol.* 11, 266–280. doi: 10.1037/1076-8998.11.3.266

Driskell, T., Driskell, J. E., Burke, C. S., and Salas, E. (2017). Team Roles: A Review and Integration. *Small Group Res.* 48, 482–511. doi: 10.1177/1046496417711529

Duff, S. N., Del Giudice, K., Johnston, M., Flint, J., and Kudrick, B. (2014). A Systems Approach to Diagnosing and Measuring Teamwork in Complex Sociotechnical Organizations. *Proceedings of the Human Factors and Ergonomics Society* 58, 573–577. doi: 10.1177/1541931214581121

Ebrahim, N. A., Ahmed, S., and Taha, Z. (2009). Virtual teams: a literature review. *Australian J. Basic Appl. Sci.* 3, 2653–2669.

Engeser, S., and Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motiv. Emot.* 32, 158–172. doi: 10.1007/s11031-008-9102-4

Feitosa, J., and Salas, E. (2020). Today's virtual teams: adapting lessons learned to the pandemic context. *Organizational Dynamics* 2020, 100777. doi: 10.1016/j.orgdyn.2020.100777

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations*. New York, NY: Free Press.

Fiske, A. P., and Haslam, N. (2005). "The four basic social bonds: Structures for coordinating interaction," in *Interpersonal Cognition*, ed. M. Baldwin (New York, NY: Guilford Press), 267–298.

Flak, O. (2013). "Results of Observations of Managers Based on the System of Organizational Terms", in *Business and Non-profit Organizations Fading Increased Competition and Growing Customers' Demands*, eds A. Nalepka & A. Ujwary-Gil (Nowy Sącz: Wyższa Szkoła Biznesu), 89–102.

Flak, O. (2019). System of Organizational Terms as a Theoretical Foundation Of Cultural Identity Research Using an Online Research Tool for Teaching Reflective Practice. *Internat. J. Arts Sci.* 12, 243–256.

Flak, O. (2020). *System of organizational terms as a methodological concept in replacing human managers with robots: Lecture Notes in Networks and Systems. Lecture Notes in Networks and Systems*. New York, NY: Springer, 471–500.

Flak, O., and Pyszka, A. (2013). Differences in perception of the participants in the management process and its real trajectory. *J. Entrep. Manag. Innov.* 9, 53–72. doi: 10.7341/2013943

Flak, O., Yang, C., and Grzegorzek, M. (2017). "Action Sequence Matching of Team Managers," in *Proceedings of the 5th International Conference on Pattern Rcognition Applications and Methods ICPRAM*, eds M. de Marsico, G. S. Di Baja, and A. Fred (Vienna).

Foy, H. J., and Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Appl. Erg.* 73, 90–99. doi: 10.1016/j.apergo.2018.06.006

Gilson, L. L., Maynard, T. M., Young, N. C. J., Vartiainen, M., and Hakonen, M. (2014). Virtual teams research: 10 years, 10 themes, and 10 opportunities. *J. Manag.* 41, 1313–1337. doi: 10.1177/0149206314559946

Hackman, J. R. (1983). *A normative model of work team effectiveness*. New Haven: Yale School of Organization and Management.

Hargie, O., and Tourish, D. (eds) (2000). *Handbook of communication audits for organisations* (1. publ). Milton Park: Routledge.

Hart, E., and Di Blasi, Z. (2015). Combined flow in musical jam sessions: A pilot qualitative study. *Psychol. Music* 43, 275–290. doi: 10.1177/0305735613502374

Harter, J. K., Schmidt, F. L., and Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *J. Appl. Psychol.* 87, 268–279. doi: 10.1037/0021-9010.87.2.268

Haslam, N. (2004). *Relational models theory: a contamporary overview*. Mahwah: Erlbaum.

Hertel, G., Geister, S., and Konradt, U. (2005). Managing virtual teams: A review of current empirical research. *Hum. Resour. Manag. Rev.* 15, 69–95. doi: 10.1016/j.hrmr.2005.01.002

Heyne, K., Pavlas, D., and Salas, E. (2011). An investigation on the effects of flow state on team process and outcomes. *Proc. Hum. Fact Erg. Soc.* 2011, 475–479. doi: 10.1177/1071181311551098

Hietanen, J. K., Surakka, V., and Linnankoski, I. (1998). Facial electromyographic responses to vocal affect expressions. *Psychophysiology* 35, 530–536. doi: 10.1017/S0048577298970445

Irshad, M. T., Nisar, M. A., Gouverneur, P., Rapp, M., and Grzegorzek, M. (2020). Ai Approaches Towards Prechtl's Assessment of General Movements: A Systematic Literature Review. *Sensors* 20:5321. doi: 10.3390/s20185321

Jackson, J. M., and Williams, K. D. (1985). Social loafing on difficult tasks: Working collectively can improve performance. *J. Personal. Soc. Psychol.* 49, 937–942. doi: 10.1037/0022-3514.49.4.937

Jarman, R. (2005). When success isn't everything – Case studies of two virtual teams. *Group Decis* 14, 333–354. doi: 10.1007/s10726-005-0318-3

Jarvenpaa, S. L., and Leidner, D. E. (1999). Communication and Trust in Global Virtual Teams. *Org. Sci.* 10, 693–815. doi: 10.1287/orsc.10.6.791

Johannes, B., Wittels, P., Enne, R., Eisinger, G., Castro, C. A., Thomas, J. L., et al. (2007). Non-linear function model of voice pitch dependency on physical and mental load. *Eur. J. Appl. Physiol.* 101, 267–276. doi: 10.1007/s00421-007-0496-6

Katahira, K., Yamazaki, Y., Yamaoka, C., Ozaki, H., Nakagawa, S., and Nagata, N. (2018). Eeg Correlates of the Flow State: A Combination of Increased Frontal Theta and Moderate Frontocentral Alpha Rhythm in the Mental Arithmetic Task. *Front. Psychol.* 9:300. doi: 10.3389/fpsyg.2018.00300

Katzenbach, J. R., and Smith, D. K. (1993). The rules for managing cross−functional reengineering teams. *Plan. Rev.* 21, 12–13. doi: 10.1108/eb054404

Kaye, L. K. (2016). Exploring flow experiences in cooperative digital gaming contexts. *Computers in Human Behavior* 55, 286–291. doi: 10.1016/j.chb.2015.09.023

Kaye, L. K., and Bryce, J. (2012). Putting The "Fun Factor" Into Gaming: The Influence of Social Contexts on Experiences of Playing Videogames. *Internat. J. Intern. Sci.* 7, 23–36.

Kaye, L. K., and Bryce, J. (2014). Go with the flow: The experience and affective outcomes of solo versus social gameplay. *J. Gam. Virt. Worlds* 6, 49–60. doi: 10.1386/jgvw.6.1.49_1

Keith, M., Anderson, G., Dean, D. L., and Gaskin, J. E. (2014). "The effects of team flow on performance: A video game experiment," in *Thirteenth Annual Workshop on HCI Research in MIS*, (Auckland).

Keller, J., Bless, H., Blomann, F., and Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *J. Exp. Soc. Psychol.* 47, 849–852. doi: 10.1016/j.jesp.2011.02.004

Kerr, N. L., and Bruun, S. E. (1981). Ringelmann Revisited. *Personal. Soc. Psychol. Bull.* 7, 224–231. doi: 10.1177/014616728172007

Kivikangas, J. M. (2006). *Psychophysiology of flow experience: An explorative study*. Master's thesis, Helsinki: University of Helsinki.

Kluger, A. N., and Nir, D. (2009). The feedforward interview. *Hum. Resour. Manag. Rev.* 20, 235–246. doi: 10.1016/j.hrmr.2009.08.002

Knox, W. B. (2012). *Learning from human-generated reward*. Thesis, TX: The University of Texas at Austin.

Kozlowski, S., and Chao, G. T. (2012). The Dynamics of emergency: Cognition and cohesion in work teams. *Manag. Dec. Econ.* 33, 335–354.

Kożusznik, B. (2005). *Wpływ Społecznyw Organizacji*. Katowice: Polskie Wydawnictwo Ekonomiczne. doi: 10.1002/mde.2552

Kozusznik, B., Paliga, M., Smorczewska, B., Grabowski, D., and Kozusznik, M. W. (2018). Development and validation of the team influence relations scale (TIReS): Beyond the measurement of individual influence in teams. *Baltic J. Manag.* 13, 84–103. doi: 10.1108/BJM-01-2017-0023

KoŻusznik, B., and Polak, J. (2016). "Regulation of influence: An ethical perspective on how to stimulate cooperation and trust in innovative social dialogue," in *Building trust and constructive conflict management in organizations*, eds P. Elgoibar, M. Euwema, and L. Munduate (New York, NY: Springer), 169–184. doi: 10.1007/978-3-319-31475-4_10

Kuo, T. H., and Ho, L. (2010). Individual difference and job performance: The relationships among personal factors, job characteristics, flow experience, and service quality. *Soc. Behav. Personal.* 38, 531–552. doi: 10.2224/sbp.2010.38.4.531

Landhäußer, A., and Keller, J. (2012). Flow and its affective, cognitive, and performance-related consequences. *Adv. Flow Res.* 9781461423, 65–85. doi: 10.1007/978-1-4614-2359-1_4

Lean, Y., and Shan, F. (2012). Brief review on physiological and biochemical evaluations of human mental workload. *Hum. Fact. Erg. Manufact. Serv. Industr.* 22, 177–187. doi: 10.1002/hfm.20269

Li, F., Shirahama, K., Nisar, M. A., Huang, X., and Grzegorzek, M. (2020). Deep Transfer Learning for Time Series Data Based on Sensor Modality Classification. *Sensors* 20:4271. doi: 10.3390/s20154271

Li, F., Shirahama, K., Nisar, M. A., Köping, L., and Grzegorzek, M. (2018). Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. *Sensors* 18:679. doi: 10.3390/s18020679

MacDonald, R., Byrne, C., and Carlton, L. (2006). Creativity and flow in musical composition: an empirical investigation. *Psychol. Music* 34, 292–306. doi: 10.1177/0305735606064838

Maeran, R., and Cangiano, F. (2013). Flow experience and job characteristics: Analyzing the role of flow in job satisfaction. *TPM* 20, 13–26. doi: 10.4473/TPM20.1.2

Man, A. -P., and Roijakkers, N. (2009). Alliance Governance: Balancing Control and Trust in Dealing with Risk. *Long Range Plan.* 42, 75–95. doi: 10.1016/j.lrp.2008.10.006

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sánchez López, T., McDowell, M., and Morvan, H. P. (2018). Physiological Parameter Response to Variation of Mental Workload. *Hum. Fact.* 60, 31–56. doi: 10.1177/0018720817733101

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McGrath, J. E. (1984). *Groups: Interaction and Performance*, Vol. 29. Hoboken, NJ: Prentice-Hall, Inc.

Meyers, M. C., van Woerkom, M., and Bakker, A. B. (2013). The added value of the positive: A literature review of positive psychology interventions in organizations. *Eur. J. Work Org. Psychol.* 22, 618–632. doi: 10.1080/1359432X.2012.694689

Mittal, A., Kumar, K., Dhamija, S., and Kaur, M. (2016). "Head movement-based driver drowsiness detection: A review of state-of-art techniques," in *Proceedings of 2nd IEEE International Conference on Engineering and Technology (ICETECH-2016): 17th & 18th March, 2016* (Venue: Rathinam Technical Campus), 903–908. doi: 10.1109/ICETECH.2016.7569378

Nacke, L. E., and Lindley, C. A. (2010). Affective Ludology, Flow and Immersion in a First- Person Shooter: measurement of player experience. *J. Can. Game Stud. Assoc.* 3, 1–21.

Nisar, M. A., Shirahama, K., Li, F., Huang, X., and Grzegorzek, M. (2020). Rank Pooling Approach for Wearable Sensor-Based ADLs Recognition. *Sensors* 20:3463. doi: 10.3390/s20123463

Olson, J. S., and Olson, G. M. (2006). Bridging distance: empirical studies of distributed teams. *Hum. Comp. Interact. Manag. Inform. Syst.* 2, 27–30.

Or, C. K., and Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupat. Erg.* 7, 83–94. doi: 10.3233/OER-2007-7202

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transact. Syst. Man Cybern. Part Syst. Hum.* 30, 286–297. doi: 10.1109/3468.844354

Peifer, C., Butalova, N., and Antoni, C. H. (2019a). *Dopaminergic activity or visual attention? Spontaneous eye blink rate as an indirect measure of flow experience*. Bochum: Ruhr-Universität Bochum.

Peifer, C., Butalova, N., and Antoni, C. H. (2019b). *Dopaminergic activity or visual attention? Spontaneous eye blink rate as an indirect measure of flow experience*. Melbourne: 6th World Congress on Positive Psychology.

Peifer, C., and Engeser, S. (2021). "Theoretical Integration and Future Lines of Flow Research," in *Advances in Flow Research*, 2nd Edn, eds C. Peifer and S. Engeser (New York, NY: Springer), 417–439. doi: 10.1007/978-3-030-53468-4_16

Peifer, C., Kluge, A., Rummel, N., and Kolossa, D. (2020a). "Fostering Flow Experience in HCI to Enhance and Allocate Human Energy," in *Engineering Psychology and Cognitive Ergonomics. Mental Workload, Human Physiology, and Human Energy. HCII 2020. Lecture Notes in Computer Science*, eds D. Harris and W. C. Li (New York, NY: Springer), 204–220. doi: 10.1007/978-3-030-49044-7_18

Peifer, C., Schulz, A., Schächinger, H., Baumann, N., and Antoni, C. H. (2014). The relation of flow-experience and physiological arousal under stress - Can u shape it? *J. Exp. Soc. Psychol.* 53, 62–69. doi: 10.1016/j.jesp.2014.01.009

Peifer, C., Syrek, C., Ostwald, V., Schuh, E., and Antoni, C. H. (2020b). Thieves of Flow: How Unfinished Tasks at Work are Related to Flow Experience and Wellbeing. *J. Hap. Stud.* 21, 1641–1660. doi: 10.1007/s10902-019-00149-z

Peifer, C., and Tan, J. (2021). "Psychophysiological correlates of flow experience," in *Advances in Flow Research*, 2nd Edn, eds C. Peifer and S. Engeser (New York, NY: Springer).

Peifer, C., and Wolters, G. (2021). "11: Flow experience in the context of work," in *Advances in Flow Research*, 2nd Edn, eds C. Peifer and S. Engeser (New York, NY: Springer), doi: 10.1007/978-3-030-53468-4_11

Peifer, C., and Zipp, G. (2019). All at once? The effects of multitasking behavior on flow and subjective performance. *Eur. J. Work Org. Psychol.* 28, 682–690. doi: 10.1080/1359432X.2019.1647168

Pels, F., Kleinert, J., and Mennigen, F. (2018). Group flow: a scoping review of definitions, theoretical approaches, measures and findings. *PLoS One* 13:e0210117. doi: 10.1371/journal.pone.0210117

Pinjani, P., and Palvia, P. (2013). Trust and knowledge sharing in diverse global virtual teams. *Inform. Manag.* 50, 144–153. doi: 10.1016/j.im.2012.10.002

Purvanova, R. K. (2014). Face-to-face versus virtual teams: What have we really learned? *The Psycholog. Manag. J.* 17, 2–29. doi: 10.1037/mgr0000009

Pyszka, A. (2015a). Istota efektywności. Definicje i wymiary [Defining Effectiveness and it's Dimensions]. *Studia Ekonomiczne* 230, 13–25.

Pyszka, A. (2015b). Modele i determinanty efektywności zespołu [Models and Determinants of Teamwork Effectiveness]. *Studia Ekonomiczne* 230, 36–54.

Quinn, R. W. (2003). *Nuclear weapons and daily deadlines: The energy and tension of flow in knowledge work*. Thesis, MI: University of Michigan.

Quinn, R. W. (2005). Flow in knowledge work: High performance experience in the design of national security technology. *Administr. Sci. Q.* 50, 610–641. doi: 10.2189/asqu.50.4.610

Raghuram, S., Hill, S. N., Gibbs, J. L., and Maruping, L. M. (2019). Virtual Work: Bridging Research Clusters. *Acad. Manag. Ann.* 13:1. doi: 10.5465/annals.2017.0020

Rau, P. L. P., Tseng, Y. C., Dong, X., Jiang, C., and Chen, C. (2017). The Effect of Personality on Online Game Flow Experience and the Eye Blink Rate as an Objective Indicator. *Adv. Hum. Comput. Interact.* 2017, 1–8. doi: 10.1155/2017/4675401

Riedl, C., and Woolley, A. W. (2017). Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-cased problem solving performance. *Acad. Manag. Discov.* 3:4. doi: 10.5465/amd.2015.0097

Rissler, R., Nadj, M., Li, M. X., Loewe, N., Knierim, M. T., and Maedche, A. (2020). To Be or Not to Be in Flow at Work: Physiological Classification of Flow using Machine Learning. *IEEE Transact. Affect. Comput.* 1:3045269. doi: 10.1109/taffc.2020.3045269

Robinson, S. L. (1996). Trust and breach of the psychological contract. *Adm. Sci. Q.* 41, 574–599. doi: 10.2307/2393868

Ross, A. (2014). The social work voice: How could unions strengthen practice? *Aotearoa N. Zealand Soc. Work* 26:4. doi: 10.11157/anzswj-vol26iss4id21

Sawyer, K. R. (2003). *Group creativity: Music, theater, collaboration*. Hove: Psychology Press.

Sawyer, K. R. (2006). Group creativity: musical performance and collaboration. *Psychol. Music* 34, 148–165. doi: 10.1177/0305735606061850

Sawyer, K. R. (2008). *Group genius: The creative power of collaboration*. New York, NY: BasicBooks.

Schaufeli, W. B., Salanova, M., and González-romá, V. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *J. Happ. Stud.* 3, 71–92. doi: 10.1023/A:1015630930326

Šimleša, M. (2018). *Collectif flow: sociocognitive model of optimal collaboration*. Doctoral thesis, Paris: Université Paris Descartes.

Sohmen, V. S. (2013). Leadership and teamwork: Two sides of the same coin. *J. Informat. Technol. Econom. Dev.* 4, 1–18.

Sundberg, J. (1998). Expressivity in singing. A review of some recent investigations. *Logoped. Phoniatr. Vocol.* 23, 121–127. doi: 10.1080/140154398434130

Tang, Y., and Zeng, Y. (2009). Quantifying designer's mental stress during the conceptual design process using kinesics study. *Internat. Conf. Eng. Design* 2009:09.

Thayer, J. F., Hansen, A. L., Saus-Rose, E., and Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation, and health. *Ann. Behav. Med. Soc. Behav. Med.* 37, 141–153. doi: 10.1007/s12160-009-9101-z

Torrente, P., Salanova, M., Llorens, S., and Schaufeli, W. B. (2013). "From "I" to "We": the factorial validity of a team work engagement scale," in *Occupational Health Psychology: From Burnout to Well-Being*, eds S. P. Goncalves and J. G. Neves (Rosemead, CA: Scientific & Academic Publishing Co).

Tozman, T., and Peifer, C. (2016). "Experimental paradigms to investigate flow experience and its psychophysiology: Inspired from stress theory and research," in *Flow Experience*, eds L. Harmat, F. Andersen, and G. Sadlo (New York, NY: Springer International Publishing), 329–350. doi: 10.1007/978-3-319-28634-1_20

van den Hout, J. J., Davis, O. C., and Weggeman, M. C. (2018). The Conceptualization of Team Flow. *J. Psychol. Interdiscip. Appl.* 152, 388–423. doi: 10.1080/00223980.2018.1449729

van den Hout, J. J., Gevers, J. M., Davis, O. C., and Weggeman, M. C. (2019). Developing and Testing the Team Flow Monitor (TFM). *Cog. Psychol.* 6:1. doi: 10.1080/23311908.2019.1643962

van Dyne, L., Ang, S., and Botero, I. C. (2003). Conceptualizing employee silence and employee voice as multidimensional constructs. *J. Manag. Stud.* 40, 1359–1392. doi: 10.1111/1467-6486.00384

Walker, C. J. (2010). Experiencing flow: is doing it together better than doing it alone? *J Posit. Psychol.* 5, 3–11. doi: 10.1080/17439760903271116

Walker, C. J. (2021). "Social flow," in *Advances in Flow Research*, 2nd Edn, eds C. Peifer and S. Engeser (New York, NY: Springer), 263–286. doi: 10.1007/978-3-030-53468-4_10

Wallbott, H. G., and Scherer, K. R. (1986). How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Soc. Sci. Inform.* 25, 763–795. doi: 10.1177/053901886025004001

Warkentin, M. E., Sayeed, L., and Hightower, R. (1997). Virtual teams versus face—to—face teams: An exploratory study of a web—based conference system. *Decis. Sci.* 28, 975–996. doi: 10.1111/j.1540-5915.1997.tb01338.x

Watzlawick, P., and Beavin, J. (1967). Some Formal Aspects of Communication. *Am. Behav. Sci.* 10, 4–8. doi: 10.1177/000276420100802

Weick, K. E., and Roberts, K. H. (1993). Collective mind in organizations: Heedful interrelating on flight decks. *Administr. Sci. Q.* 38, 357–381. doi: 10.2307/2393372

Woolley, A., Chabris, C., Pentland, A., Hashmi, N., and Malone, T. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 686–688. doi: 10.1126/science.1193147

Zebrowitz, L. A. (1990). *Mapping social psychology series. Social perception*. Boston, MA: Thomson Brooks/Cole Publishing Co.

Zubair, A., and Kamal, A. (2015). Work Related Flow, Psychological Capital, and Creativity Among Employees of Software Houses. *Psychol. Stud.* 60, 321–331. doi: 10.1007/s12646-015-0330-x

Zumeta, L., Basabe, N., Wlodarczyk, A., Bobowik, M., and Paez, D. (2016). Flujo Compartido y Reuniones Colectivas Positivas. *Anales De Psicología* 32:717. doi: 10.6018/analesps.32.3.261651

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.

frontiers in Psychology

PERSPECTIVE
published: 27 September 2021
doi: 10.3389/fpsyg.2021.720476

"Like I'm Talking to a Real Person": Exploring the Meaning of Transference for the Use and Design of AI-Based Applications in Psychotherapy

Michael Holohan and Amelia Fiske*

Institute of History and Ethics in Medicine, School of Medicine, Technical University of Munich, Munich, Germany

AI-enabled virtual and robot therapy is increasingly being integrated into psychotherapeutic practice, supporting a host of emotional, cognitive, and social processes in the therapeutic encounter. Given the speed of research and development trajectories of AI-enabled applications in psychotherapy and the practice of mental healthcare, it is likely that therapeutic chatbots, avatars, and socially assistive devices will soon translate into clinical applications much more broadly. While AI applications offer many potential opportunities for psychotherapy, they also raise important ethical, social, and clinical questions that have not yet been adequately considered for clinical practice. In this article, we begin to address one of these considerations: the role of transference in the psychotherapeutic relationship. Drawing on Karen Barad's conceptual approach to theorizing human–non-human relations, we show that the concept of transference is necessarily reconfigured within AI-human psychotherapeutic encounters. This has implications for understanding how AI-driven technologies introduce changes in the field of traditional psychotherapy and other forms of mental healthcare and how this may change clinical psychotherapeutic practice and AI development alike. As more AI-enabled apps and platforms for psychotherapy are developed, it becomes necessary to re-think AI-human interaction as more nuanced and richer than a simple exchange of information between human and nonhuman actors alone.

Keywords: artificial intelligence, psychotherapy, mental healthcare, chatbots, transference, embedded ethics, science and technology studies, agential realism

OPEN ACCESS

Edited by:
Stefano Triberti,
University of Milan, Italy

Reviewed by:
Anto Čartolovni,
Catholic University of Croatia, Croatia
Felice Cimatti,
University of Calabria, Italy
Corinna Peifer,
University of Lübeck, Germany

****Correspondence:***
Michael Holohan
m.holohan@tum.de

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *04 June 2021*
Accepted: *16 August 2021*
Published: *27 September 2021*

Citation:
Holohan M and Fiske A (2021) "Like I'm Talking to a Real Person": Exploring the Meaning of Transference for the Use and Design of AI-Based Applications in Psychotherapy. Front. Psychol. 12:720476. doi: 10.3389/fpsyg.2021.720476

INTRODUCTION

A first-year college student is having trouble adjusting to university life. There are so many new things to deal with, so many new demands and responsibilities. She is making new friends, but finds it hard to connect with them. Her grades are starting to slip and she feels like she is losing control of her life. When she eventually decides to check the campus health service website to see what mental health services are available, she finds that there is a long waitlist to see a counselor. However, the website suggests an alternative that is available

immediately and is entirely free: a text-based chatbot, powered by artificial intelligence (AI). Using an app the student downloaded onto her phone, the chatbot checks in regularly to ask how she is doing, helps her to identify the emotions she feels in difficult situations, and suggests some relaxation exercises to work through her anxiety. She likes that the chatbot is available around the clock and always texts back immediately. Even though she knows she is talking to a computer, she feels heard and even understood.

Like this example, chatbots such as Tess,[1] Wysa,[2] or Woebot[3] offer similar virtual psychotherapeutic services and have demonstrated promising results in reducing symptoms of depression and anxiety in trial studies (Fitzpatrick et al., 2017; Fulmer et al., 2018). AI-enabled virtual and robot therapy is increasingly being integrated into psychotherapeutic practice. Given the speed of research and development trajectories of AI-enabled applications in psychotherapy and the practice of mental healthcare, it is likely that therapeutic chatbots, avatars, and socially assistive devices will soon translate into clinical applications much more broadly.

However, this field is still nascent and there are many questions that remain to be considered or clarified. For example, what does it mean to interact with a robot for help with your mental health? What does it mean to form a personal connection in a therapeutic setting with something you know is not a person? This is an issue not just for the users who access these services, but also for the engineers and designers who are developing these interfaces: How to best design algorithms that help people work through their intimate problems in a way that fosters a connection between the person and the interface? How is the therapeutic connection established and how do you factor it into your design? Moreover, is the nature of the connection with a virtual therapist even comparable to that of a human therapist?

The companies developing AI-enabled therapeutic applications have designed the applications to look and feel much like in-person therapy. However, this surface similarity obscures the possibility that there may be significant differences between AI-directed and human-directed psychotherapy. Therefore, it is necessary to carefully examine the points of similarity and difference between AI-directed and human-directed psychotherapy. Doing so will allow us to better understand not only the limitations of AI applications vis-a-vis traditional psychotherapy, but also what is new and unique about such applications and what they might make possible.

One aspect that deserves particular attention is the sense of "personal" connection between user-patients and their chatbot therapist. This is because most modalities of psychotherapy have a concept of "transference," which describes a specific way that patients and therapists relate to each other within the therapeutic relation. In this article, we focus on transference as one example in order to highlight some fundamental issues related to the use of AI-enabled psychotherapy. Drawing on the work of Science and Technology Studies (STS)[4] scholar Karen Barad on material-discursive practices in human–non-human relations (Barad, 1999, 2007), we present a framework for conceptualizing the therapeutic setting in order to help those involved (psychotherapists, patients, support staff, caretakers, robotics engineers, developers, researchers, ethicists, administrators, legislators, etc.) better understand the nature of the AI-driven therapeutic encounter. This approach can help to inform further work in this field, in terms of therapeutic practice with existing AI applications, research into the effects of such practices, and the research and development of new AI applications.

In what follows, we first present a review of the literature on existing AI-enabled psychotherapeutic applications. We then outline the concept of transference in psychotherapy, putting it in conversation with Barad's theory of agential realism. We end with a discussion of the implications of transference in relation to AI-enabled psychotherapy and possibilities for further research.

THE CURRENT STATE OF AI-ENABLED PSYCHOTHERAPEUTIC APPLICATIONS

Work in embodied artificial intelligence (AI) has growing clinical relevance for diagnostic and therapeutic applications across several areas in medicine (Calderita et al., 2014; Broadbent, 2017; Liu et al., 2018). Such applications are no longer designed to just provide simple assistive services, but also perform higher-level, invasive, diagnostic, and therapeutic interventions that used to be offered exclusively by highly trained health professionals (Jahn et al., 2019). In the area of mental health, embodied AI is increasingly being integrated into psychotherapeutic practice (Fiske et al., 2019). It has been proposed to support a range of emotional, cognitive, and social processes (Eichenberg and Küsel, 2018) through the use of chatbots, virtual reality therapies, social robots, and more. In what follows, we briefly summarize the range of AI applications that are being researched, tested, and applied in the area of mental healthcare, with a specific focus on applications within psychotherapy. As such, we have intentionally excluded from this analysis all AI applications that do not interact with patients directly, and those that may have a virtual or robotic interface but do not employ AI, such as telemedicine therapy.

The most prominent domain of AI-driven psychotherapeutic applications is therapeutic apps, sometimes called "chatbots."

[1]https://www.x2ai.com/
[2]https://www.wysa.io/
[3]https://woebothealth.com/

[4]STS is an "interdisciplinary research field that studies how social, political and cultural values and structures affect scientific research and technological innovation, and how research and innovation in turn affect society, politics and culture" (Müller et al., 2021). STS scholars analyze how and under which conditions scientific knowledge and technologies are produced as well as the distinct social, political, economic, and historical contexts of research and technology development. For example, STS examines how new concepts such as biomarkers change knowledge production in psychiatry, why these biological parameters are used to pursue specific research and treatment goals and not others, and how the outcomes of this research might affect society in social, political, economic and normative ways.

Known by their first names, apps such as Tess, Sara, Wysa, Ada, or Woebot work *via* text or on internet platforms and have addressed conditions such as depression, anxiety, and autism. Many such applications respond to the user in a way that aims to mimic a human therapist, probing the user to explore emotions or thought patterns that they are experiencing. Others offer techniques for reducing anxiety or advice for dealing with difficult situations (Sachan, 2018; Dekker et al., 2020), help users implement problem-solving strategies and approach problems from different perspectives, or inform users of nearby psychiatric services when needed (Bendig et al., 2019). Recent reviews found over 40 chatbots addressing mental health concerns available, most with several purposes including therapy, training, and screening (Abd-alrazaq et al., 2019; Tudor Car et al., 2020).

The area of virtual reality is increasingly being proposed for use with patients experiencing psychosis (Craig et al., 2018), schizophrenia, and autism. One such example currently in clinical testing is the Avatar Project,[5] in which an intelligent algorithm is expressed through an avatar which interacts with a patient in order to address symptoms such as persistent auditory hallucinations. The use of avatars is also being explored in AI-assisted therapy for schizophrenia (Dellazizzo et al., 2018a,b) as well as in combination with real-time fMRI (de Pierrefeu et al., 2018). Studies of virtual human agents have also experimented with improving interviewing skills with individuals with autism or other developmental disabilities (Burke et al., 2018), promoting life skills and well-being for adolescents (Gabrielli et al., 2020), treating the fear of heights (Freeman et al., 2018; Donker et al., 2019), and risk prevention (Rein et al., 2018).

While some technologies might be used as part of supervised therapies, AI-driven psychotherapeutic applications such as chatbots are slowly but surely progressing toward a therapeutic role outside of settings where human mental health professionals are involved. It is therefore necessary to assess how important elements of the "traditional" relationship between therapist and client/patient are either retained, altered, or made anew in the relationship between user and chatbot therapist. One central element of the traditional therapeutic relationship is transference, which is of particular interest because it is a form of personal connection that is specific to the psychotherapeutic setting. In the next section, we will discuss the concept of transference, how it functions, and why it is relevant for the study and design of AI-directed therapies.

THE CONCEPT OF TRANSFERENCE IN PSYCHOTHERAPY

The concept of transference can be traced back to the earliest days of psychotherapy. Introduced by Sigmund Freud (1912/2001) in the context of psychoanalytic treatment, it is a foundational concept in many forms of psychotherapy. Transference refers to a phenomenon where a patient redirects emotions, feelings, or wishes that were originally directed toward other people in their life onto the therapist (Goldstein and Goldberg, 2004; Parth et al., 2017). Transference can manifest, for example, in a patient's speech, demeanor, attitude, or patterns of behavior (Fink, 2007). The appearance of transference is not an accident, but an inevitable aspect of the therapeutic process (Freud, 1912/2001; Friedman, 2019). Put another way: it is not a bug; it is a feature of the therapeutic relationship.

Transference is integral to the interpersonal relationship between patient and therapist and represents an important point of action in the psychotherapeutic process. Regardless of what the two parties are talking about at a given moment, there is always another relationship in the room, i.e., the patient's relationship to someone else in their life, either actual or imagined. However the patient speaks to and acts toward their therapist—including silences and elisions—the past is present in their speech and behavior in the form of these prior relationships that the patient brings (i.e., transfers) into the consulting room. The therapist must be able to acknowledge this transference and work with it, since it is as indispensable to the treatment as it is unavoidable. Transference can have multiple different effects in the therapeutic relationship. For example, transference can help foster the therapeutic alliance, especially in the early stages of the treatment. A positive transference can make it possible for the patient to face difficult subjects, by helping them feel supported and understood. Transference is often also an object of analysis itself, and identifying, discussing, and actively working through transference feelings is a significant part of most psychodynamic psychotherapies. Transference can also act as a form of resistance and as an obstacle to treatment by keeping the patient from feeling like they can discuss certain ideas or topics, or a strong negative transference can make it hard for the patient to attend sessions regularly or even cause them to terminate the treatment (for an overview of the different effects of transference and ways of working with it, see Fink, 1997; Corradi, 2006; and Fink, 2007, esp. chapter 7; for empirical studies of its usefulness, see Marmarosh, 2012; Hersoug et al., 2014; Suszek et al., 2015; Ulberg et al., 2021).

Depending on the specific theoretical orientation of the psychotherapy, working with transference may be more or less central to the treatment, but it nonetheless remains a tool in the therapist's tool kit. For example, imagine a patient for whom the therapist's haircut or tone of voice resembles the hair or voice of her father, with whom she has a poor relationship. Based on this trivial similarity, the patient begins, sometimes without even meaning to, to act toward her therapist with the same kind of denial and protest that she did with her father. This transference of feeling from the father onto the therapist can lead the patient to complain about the therapist, find it hard to trust him, or even start to miss sessions. Without identifying and working through this negative transference, the therapy is unlikely to make any progress. It is worth noting that while we have referred here to "positive" or "negative" transference feelings, more often transference represents a fusion of contradictory currents (positive and negative, love and hate,

[5]https://www.ucl.ac.uk/brain-sciences/news/2020/nov/new-avatar-project-help-auditory-hallucinations

admiration and fear, etc.) that are inextricably entangled with each other.

While the concept of transference is most commonly associated with psychoanalytic and psychodynamic psychotherapies, it is also discussed in other approaches such as cognitive behavioral therapy (CBT) (Prasko et al., 2010; Folk et al., 2016). This is particularly significant because existing chatbots like Woebot and Tess are designed on CBT principles (Fitzpatrick et al., 2017; Fulmer et al., 2018).

As of yet, there have been no studies of transference in AI-enabled psychotherapeutic settings. However, studies of specific chatbots demonstrate anecdotal evidence that some users develop a human-like connection with the chatbot that can be seen as suggestive of the kind of personal relationship out of which transference can develop. For example, one study participant wrote, "I love Woebot so much. I hope we can be friends forever. I actually feel super good and happy when I see that it 'remembered' to check in with me!" (Fitzpatrick et al., 2017). Another participant in a similar study of the chatbot Tess wrote "Based on our interactions I do somewhat feel like I'm talking to a real person and I do enjoy the tips you've given. In that sense, you're better than my therapist in that she doesn't necessarily provide specific ways I can better myself and problems" (Fulmer et al., 2018).

UNDERSTANDING PSYCHOTHERAPY THROUGH THE LENS OF KAREN BARAD'S AGENTIAL REALISM

Transference is both a product of the psychotherapeutic encounter and a mechanism though which treatment occurs. In order to better understand what this means and how it can be considered in AI-driven therapy, we turn to STS scholar Karen Barad's theory of agential realism, a conceptual approach to theorizing human–non-human relations (Barad, 1999, 2007). Barad's theory provides a framework for conceptualizing and understanding what the psychotherapeutic encounter consists of, what its elements are, and how those elements shape what is possible in the encounter. This makes it possible to analyze different kinds of situations and identify how substituting a chatbot for a human therapist might alter the situation. Barad's theory focuses on knowledge production, which relates to AI-driven psychotherapy in terms of how it creates knowledge about such things as emotional states, patterns of behavior, or unconscious desires, depending on the therapeutic tradition.

Barad builds on the theoretical and epistemological work of quantum physicist Niels Bohr, arguing that the knower does not stand apart from the object they seek to measure (Barad, 2007). As an illustrative example, she considers the well-known Heisenberg uncertainty principle, which states that it is impossible to measure both a particle's position and velocity at the same time. Bohr argued that this is because the experimental apparatus determines what can be measured and thus also the conceptual framework for understanding.

> For example, instruments with fixed parts are required to understand what we might mean by the concept 'position.' However, any such apparatus necessarily excludes other concepts, such as 'momentum,' from having meaning during this set of measurements, since these other variables require an instrument with moveable parts for their definition. Physical and conceptual constraints are co-constitutive. (Barad, 1999, p. 4)

The interaction between what is observed and the apparatus used to observe it are thus inseparable from each other. Together they produce what Barad calls phenomena, and these phenomena are constitutive of the apparatus as well as the products of that apparatus, by means of "physical-conceptual *intra-actions*" (Barad, 1999, p. 5).

An apparatus is the set of materials and practices that, by being put to use in a specific situation and for a specific purpose, create the conditions of possibility for what can happen in that situation. Barad's agential-realist framework, and especially the concepts of apparatus and phenomena, can be useful for thinking about the practice of psychotherapy: The tools one uses in the therapeutic encounter (e.g., AI and a specific interface such as a text-based chatbot) are formative and constitutive of the kind of therapy that becomes possible. This also applies to less drastic changes in the traditional therapeutic process – any practitioner who has used remote technologies such as Zoom during the Covid-19 pandemic will be all too familiar with how the introduction of new technologies into the "standard" forms of "in-person" treatment has had distinct, if often difficult-to-articulate effects.

Following Barad, transference is a phenomenon which emerges as a product of the therapeutic apparatus. In this sense, transference is simultaneously also "productive of" the material-discursive psychotherapeutic apparatus (i.e., the therapeutic encounter) itself: It contributes to the formation of the therapeutic relationship. Transference is thus an artefact of the process itself, inherent to it and understandable only (or mainly) within its framework.

Based on this, we can attempt a preliminary definition of what the traditional psychotherapeutic apparatus is composed of, in terms of material-discursive practices: the therapist, the patient, the consulting room, periodic meetings (scheduled weekly, bi-weekly, etc.), specific modes of speaking and interacting, and specific techniques for eliciting the therapeutic relationship, insight, emotional change, or conflict (these can be specific to different therapeutic schools, including CBT, psychodynamic psychotherapy, psychoanalysis, humanistic psychology, etc.). Also included in this would be different means of interaction, such as sitting face to face, the use of the couch, or any technological modes of mediation such as email, text messages, apps, video, or avatars. As we shift to AI-directed therapy, a new apparatus emerges. New modes of material-discursive practices come into being: the chatbot therapist, the user/patient, mediation *via* an app on a mobile device or tablet, specific text-based modes of interacting, always-on availability, etc.

(RE)THINKING TRANSFERENCE WITH AI

As we can see from the studies of the chatbots Woebot and Tess, there are preliminary indications in the literature that some users develop human-like connections with their chatbot: "I love Woebot so much. I hope we can be friends forever. I actually feel super good and happy when I see that it 'remembered' to check in with me!" (Fitzpatrick et al., 2017). These feelings of happiness, love or enjoyment demonstrate that some users do not necessarily treat chatbots like inanimate instruments for self-improvement, but can relate to them as if they were "talking to a real person" (Fulmer et al., 2018). Even a routine feature such as pre-scripted regular check-ins can be interpreted as the chatbot "remembering" the user. If these affective connections are being made, it is certainly conceivable that transference may also develop in such situations. It is even possible that this is already happening.

Transference is a useful phenomenon to consider not only because it is specific and essential to the psychotherapeutic apparatus, but because it occurs as a relationship between the patient and therapist. The apparatus enables and is enabled by a process of intra-action, or what feminist STS scholar Donna Haraway calls "becoming with," a form of entanglement where "The partners do not precede their relating": Chatbots become therapeutic only through their intra-action with users, who themselves become patients (Haraway, 2008, p. 17). It is therefore readily apparent that the apparatus has changed when the therapist is no longer a human, but a chatbot. Since the phenomenon of transference is crucial to the psychotherapeutic apparatus, we must ask how AI-driven innovations could be designed to account for and even foster opportunities for transference that might be useful and even novel. In other words, it will be necessary to conceive of transference not as an unanticipated byproduct of AI-directed psychotherapy, but to actively consider it in the design process. Here it helps to think of the psychotherapeutic encounter as an apparatus because it allows us to see how the material-discursive practices that make up the apparatus make possible or hinder certain intra-actions, thus creating different phenomena.

One place to start would be to ask what transference might look like in relation to a chatbot: What quality or qualities of the chatbot interface, for example, might become the kernel for a patient's transference? How might the patient be relating transferentially to the chatbot (through what words, behaviors, demeanor, etc.)? Does it matter if the chatbot operates through an avatar with "human-like" features? In approximating the responses of a human therapist, are there specific speech patterns, forms of questioning, or other features of AI communication that might give rise to specific forms of transference in the therapeutic encounter? For example, imagine the following scenario: A patient using a psychotherapeutic chatbot feels relief in not being judged, since they know they are interacting with a robot. On the one hand, this makes them feel safe, making it easier to talk about difficult topics. On the other hand, the patient might at the same time contrast this absence of judgement with the overly judgmental attitude of their mother, to whom they still attribute a strong degree of authority despite the fact that they suffer under her judgmental gaze. In this case, the patient might ultimately fail to take their chatbot therapist seriously, or even treat it with disdain because, through their transference, they ascribe a lack of authority to the chatbot, even though interacting with it makes them feel safe and cared for. It is important to note that patients are often unaware of transference when it happens.

In this example, the specific form of the apparatus produces the specific phenomenon of transference: The patient develops a relation to the chatbot precisely because they know they are talking to a robot who is incapable of judging them. However, as we can see, this transferential phenomenon might also make it difficult, if not impossible, to sustain the therapeutic relationship, potentially leading to its premature collapse. In this case, we might ask what would it mean for chatbot designers to take this into account? Would it be possible for the chatbot to register not just that there has been a shift in the patient's relationship to it, but that this is due to a resemblance with a person from the patient's past, and that the patient might be unaware of this aspect of their transference?

While this example is illuminating, it is also ultimately limited because it presumes that the form of the psychotherapeutic intra-action between a human and a chatbot will look very much like that between two humans. The AI–human psychotherapeutic apparatus and its phenomena remain in many ways undetermined, and the phenomena that it produces might look quite different from what we are used to or can easily imagine. As mentioned above, the apparatus determines the phenomena and thus the conceptual framework for understanding. So, how might the phenomenon of transference be constituted differently in an encounter with a chatbot versus a meeting with a therapist in their practice? For example, our scenario focused on the question of judgment, yet our understanding of what judgement means might need to be redefined or rearticulated in light of the specificities of an AI-driven chatbot. The question of judgment that we are familiar with in psychotherapy is a phenomenon produced by an apparatus based on human intra-actions. Humans judge each other. A psychotherapist is supposed to withhold their judgment, but a patient might justifiably wonder whether their therapist is actually capable of such a feat and, through transference, attribute judgment to their therapist even if none actually exists. In comparison, a chatbot is incapable of expressing personal judgment. Yet this might not cause the question of judgement to simply disappear. Instead, the question might shift to how societal norms are "baked in" to the chatbot's algorithm, since a chatbot's AI might be trained on a dataset that is structurally (algorithmically) biased (Manrai et al., 2016; Obermeyer et al., 2019; Panch et al., 2019a,b). As the makeup of the apparatus shifts from human–human to human–AI, the concept shifts from personal judgment to impersonal, structural bias.

The therapeutic relationship (even when produced by a chatbot) should never be understood to be a "simple" interaction between human and/or nonhuman actors, which is to say one modeled on general social interaction models (which are themselves, of course, far from simple). This requires a recognition of the assumptions and definitions that are at play in any

interactive design, including a (re)definition of any and all concepts with an eye to how they are produced by the design of the apparatus. Following Barad, such definitions or concepts do not preexist their emergence from and within the apparatus. In other words, it is not possible to say what concepts are or will be best suited to understanding the technologies to come except in and through designing and testing them.

We must ask how the inclusion of AI (either to augment or replace some aspect of the human therapist) changes the apparatus, and how this new mode of therapy changes and can be designed to change the phenomena that are produced, raising a series of important questions for psychotherapy and for AI developers: Does transference occur with the inclusion of AI in the therapeutic encounter? If so, what forms does this transference take and how does it shape the ensuing therapeutic relationship and therapeutic work? How can transference be accounted for, and addressed, within AI-driven therapy? How can transference be intentionally engaged by developers and engineers in the design of AI-driven therapeutic apps? Is it even transference as we currently understand it? Or is it some other kind of relation that may look like transference, but is in some way different? What new phenomena are unique to the new apparatus? In what ways does the therapeutic process that occurs in AI-driven encounters overlap with and differ from human therapeutic relationships?

One way to approach these questions would be to consider the agency of the non-human actors in this context. Here again, Barad's work is useful. For Barad, agency is "an enactment, not something that someone or something has" (Barad, 2007, p. 214). In other words, agency describes an effect that is not imposed from outside, but which is produced by something from within a given set of intra-actions. Thus, agency can be extended to non-human actors (in this case AI, algorithms, chatbot interfaces) because their presence and specificity have demonstrative effects. We must keep in mind that technology is not passive in the co-production of phenomena. As Barad likes to say, "The world kicks back" (Barad, 2007, p. 215). In other words, the agency of the non-human elements of the apparatus matters. We can try to design different ways of relating within the apparatus of AI-based psychotherapy, but in the end, what emerges is not pre-scripted; it needs to be the subject of empirical study. It is not possible to fully know what we are creating ahead of time, but we can be intentional about trying to create opportunities. This must be an iterative and recursive process, always going back to see how the human and non-human actors intra-act in their encounters, and what those encounters produce.

DISCUSSION

In order to better understand how using AI differs from human-directed psychotherapy, it is helpful to reflect on the realities and possibilities of AI-enabled apps and other psychotherapeutic interfaces, including the elements which enable the therapeutic encounter to occur through these platforms. This means exploring not only what AI-driven therapies cannot do, but also to what they can offer, make possible, and what the implications are for clinical psychotherapeutic practice. Considering the psychotherapeutic setting as an apparatus that structures the conditions of what is possible and that is productive of specific phenomena, *a la* Barad's theory of agential realism, we can see that shifting the elements of the therapeutic apparatus—such as the use of an app, platform, or AI technology—fundamentally reshapes the therapeutic encounter itself. This has direct application to the implementation and research and development of AI-enabled apps and any other interfaces that might be developed down the line.

As we have shown, there is a need for further research in this regard. First, there is a need for studies that investigate what psychotherapy becomes with the introduction of AI-enabled "therapists." This should include the consideration of the effects that specific interfaces such as chatbots and virtual avatars have on the psychotherapeutic apparatus. For example, there is the significant change in the apparatus introduced by the always-on aspect of mobile devices. A smartphone-based chatbot can be available anytime, day or night, with no limit to the length of the "session" (a term which depends entirely on the fact that it has an end). This is in stark contrast to the limited and prescribed availability of a traditional therapist, which is often an important feature of the therapeutic interaction. In addition, the always-on aspect affects the kind of "data" that the AI-driven app can collect (either actively or passively) about a patient through a smartphone's different sensors (microphone, GPS, gyroscope, accelerometer, ambient light sensor, camera, lidar, etc.) and usage histories (browser history, app usage, screentime metrics, etc.), which of course also raises new and specific issues regarding trust and privacy in AI-driven therapeutic apparatuses. It is not a matter of AI-enabled interactions being something less than traditional psychotherapy, but as potentially being something new altogether.

Empirical studies will be instrumental to understanding the complexities of the new and emergent AI-driven apparatuses. The examples and scenarios we have provided in this article have been hypothetical and speculative. While this kind of speculative thinking is valuable and necessary, it should be supported with empirical studies that identify and analyze how users actually relate to chatbot therapists in real-life situations as well as which assumptions, for example regarding the therapeutic relationship, about possible user groups and their needs go into the design of psychotherapeutic apps. This will be particularly important in order to clearly identify the kinds of novel effects and phenomena that a new apparatus might produce, especially those which may be quite different from what we can imagine ahead of time. Transference is one important aspect to be considered here: Which notions of transference go into the design of apps and which transference phenomena with AI-driven psychotherapeutic apps are produced as users begin to interact with them?

Within these empirical studies, it will be important to put emphasis on how people from different social positions (based on gender, ethnicity, sexuality, age, or socioeconomic status) might interact with these new opportunities differently. It is

known that, based on their specific social position, people have different relationships to issues of mental health, healthcare in general, as well as technology (Oudshoorn and Pinch, 2003; Epstein, 2007; Criado Perez, 2019). It is important to interrogate whose needs and interests are represented in the design of currently existing psychotherapeutic AI-driven applications as well as how these different user groups interact with, benefit from, or are put at risk by these new technologies. Intersectional analysis is paramount here. Transference, of course is shaped by social positionality and hence is an important topic of study in this context. In addition, from a health equity perspective, it will also be essential to investigate who might be the people that doctors and therapists refer to AI-driven therapeutic technologies versus who might be referred to traditional human therapists. In this context, AI might both hinder or promote health equity.

In this article, we have focused primarily on scenarios where an AI-driven chatbot replaces a human therapist. However, there are other instances where a chatbot may be used in addition to or as an augmentation of human-directed psychotherapy. This might occur deliberately, where a therapist suggests the use of a chatbot as part of the therapy. A current example of this is an app to treat substance use disorders, which is used as an addition to in-person treatment (Budney et al., 2019; Triberti et al., 2020). But it could also be that it is not a deliberate choice of the therapist to introduce AI-driven applications, such as situations where patients begin to use chatbots on their own. These intentional or unintentional triad situations might create an overlap of different apparatuses or the emergence of a hybrid apparatus of treatment. Such a situation can lead to confusion about the role of artificial entities in the complex therapist-patient relationship, an aspect of what has been characterized as a "third wheel effect" (Triberti et al., 2020). This is an important aspect that should be considered in the further study of psychotherapy, AI, and transference.

One possible way of addressing these therapeutic concepts as they emerge in AI-driven technologies is to integrate practitioners of psychotherapy as well as social scientists well-versed in the social study of technology and healthcare in the design process. This form of integration can be analogous to an approach recently promoted for ethically sound and socially robust AI applications in other fields of healthcare, the "embedded ethics and social science" approach. This approach combines participatory research practices that include the study of both technical development and user perspectives with empirical bioethical analysis (Fiske et al, 2020; McLennan et al., 2020). Embedded ethics integrates critical voices from the social sciences and fields of practice into the development process from the beginning, so as to anticipate, identify, and address ethical and social issues that arise during the process of developing healthcare technologies, including planning, ethics approval, designing, programming, piloting, testing, and implementation phases of the technology. Positioning these actors as participants in the development stages of healthcare technology, such as AI-driven psychotherapeutic apps, aims to promote the reflexive and equity-oriented design of novel technologies. It thereby helps to anticipate, rather than simply respond to, vital questions regarding the social impact of such technologies, such as the role of transference in the therapeutic encounter in new AI-driven healthcare technologies.

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

MH had the initial idea for the article, led the writing process, and wrote and edited the majority of the article. AF codeveloped the content and structure of the article, wrote parts of the article, and commented on as well as edited MH's work. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank the three reviewers and the editors for their constructive comments that helped to improve the article. We would also like to thank the editors for convening the research topic "On the Human in Human-Artificial Intelligence Interaction" and including this article. Finally, many thanks to Alena Buyx and Ruth Müller for their valuable feedback during the writing process.

REFERENCES

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A., Bewick, B., Gardner, P. H., and Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* 132:103978. doi: 10.1016/j.ijmedinf.2019.103978

Barad, K. (1999). "Agential realism: feminist interventions in understanding scientific practices," in *The Science Studies Reader*. ed. M. Biagioli (London: Routledge), 1–11.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke University Press.

Bendig, E., Erb, B., Schulze-Thuesing, L., and Baumeister, H. (2019). Die nächste Generation: Chatbots in der klinischen Psychologie und Psychotherapie zur Förderung mentaler Gesundheit – Ein Scoping-Review. *Verhaltenstherapie*, 29, 266–280. doi: 10.1159/000499492

Broadbent, E. (2017). Interactions with robots: the truths we reveal about ourselves. *Annu. Rev. Psychol.* 68, 627–652. doi: 10.1146/annurev-psych-010416-043958

Budney, A. J., Borodovsky, J. T., Marsch, L. A., and Lord, S. E. (2019). "Technological innovations in addiction treatment," in *The Assessment and Treatment of Addiction*. eds. I. Danovitch and L. Mooney (St. Louis: Elsevier), 75–90.

Burke, S. L., Bresnahan, T., Li, T., Epnere, K., Rizzo, A., Partin, M., et al. (2018). Using virtual interactive training agents (ViTA) with adults with autism and other developmental disabilities. *J. Autism Dev. Disord.* 48, 905–912. doi: 10.1007/s10803-017-3374-z

Calderita, L. V., Manso, L. J., Bustos, P., Suárez-Mejías, C., Fernández, F., and Bandera, A. (2014). THERAPIST: towards an autonomous socially interactive robot for motor and neurorehabilitation therapies for children. *JMIR Rehabil. Assist. Technol.* 1:e1. doi: 10.2196/rehab.3151

Corradi, R. (2006). A conceptual model of transference and its psychotherapeutic application. *J. Am. Acad. Psychoanal. Dyn. Psychiatry* 34, 415–439. doi: 10.1521/jaap.2006.34.3.415

Craig, T. K. J., Rus-Calafell, M., Ward, T., Leff, J. P., Huckvale, M., Howarth, E., et al. (2018). AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatry* 5, 31–40. doi: 10.1016/S2215-0366(17)30427-3

Criado Perez, C. (2019). *Invisible Women: Data Bias in a World Designed for Men*. New York: Abrams Press.

de Pierrefeu, A., Fovet, T., Hadj-Selem, F., Löfstedt, T., Ciuciu, P., Lefebvre, S., et al. (2018). Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity. *Hum. Brain Mapp.* 39, 1777–1788. doi: 10.1002/hbm.23953

Dekker, I., De Jong, E. M., Schippers, M. C., De Bruijn-Smolders, M., Alexiou, A., and Giesbers, B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Front. Psychol.* 11:1063. doi: 10.3389/fpsyg.2020.01063

Dellazizzo, L., du Sert, O. P., Phraxayavong, K., Potvin, S., O'Connor, K., and Dumais, A. (2018a). Exploration of the dialogue components in avatar therapy for schizophrenia patients with refractory auditory hallucinations: a content analysis. *Clin. Psychol. Psychother.* 25, 878–885. doi: 10.1002/cpp.2322

Dellazizzo, L., Potvin, S., Phraxayavong, K., Lalonde, P., and Dumais, A. (2018b). Avatar therapy for persistent auditory verbal hallucinations in an ultra-resistant schizophrenia patient: a case report. *Front. Psych.* 9:131. doi: 10.3389/fpsyt.2018.00131

Donker, T., Cornelisz, I., van Klaveren, C., van Straten, A., Carlbring, P., Cuijpers, P., et al. (2019). Effectiveness of self-guided app-based virtual reality cognitive behavior therapy for acrophobia: a randomized clinical trial. *JAMA Psychiat.* 76, 682–690. doi: 10.1001/jamapsychiatry.2019.0219

Eichenberg, C., and Küsel, C. (2018). Roboter in der Psychotherapie: Intelligente artifizielle Systeme. *Deutsches Ärzteblatt* 17, 365–367.

Epstein, S. (2007). *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Fink, B. (1997). *A Clinical Introduction to Lacanian Psychoanalysis: Theory and Technique*. Cambridge: Harvard UP.

Fink, B. (2007). *Fundamentals of Psychoanalytic Technique: A Lacanian Approach for Practitioners*. New York: W. W. Norton & Co.

Fiske, A., Henningsen, P., and Buyx, A. (2019). Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet. Res.* 21:e13216. doi: 10.2196/13216

Fiske, A., Tigard, D., Müller, R., Haddadin, S., Buyx, A., and McLennan, S. (2020). Embedded ethics could help implement the pipeline model framework for machine learning healthcare applications. *American Journal of Bioethics.* 20, 32–35. doi: 10.1080/15265161.2020.1820101

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment. Health* 4:e19. doi: 10.2196/mental.7785

Folk, J., Disabato, D., Goodman, F., Carter, S., Dimauro, J., and Riskind, J. (2016). Wise Additions Bridge the Gap Between Social Psychology and Clinical Practice: Cognitive-Behavioral Therapy as an Exemplar. *Journal of Psychotherapy Integration.* doi: 10.1037/int0000038

Freeman, D., Haselton, P., Freeman, J., Spanlang, B., Kishore, J., Albery, E., et al. (2018). Automated psychological therapy using immersive virtual reality for treatment of fear of heights: a single-blind, parallel-group, randomised controlled trial. *Lancet Psychiatry.* 5, 625–632. doi: 10.1016/S2215-0366(18)30226-8

Freud, S. (1912/2001). "The dynamics of transference," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XII (1911–1913): The Case of Schreber, Papers on Technique and Other Works*. ed. J. Strachey (London: Vintage), 97–108.

Friedman, L. (2019). *Freud's Papers on Technique and Contemporary Clinical Practice*. London: Routledge.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., and Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment. Health* 5:e64. doi: 10.2196/mental.9782

Gabrielli, S., Rizzi, S., Carbone, S., and Donisi, V. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. *JMIR Hum. Factors* 7:e16762. doi: 10.2196/16762

Goldstein, W. N., and Goldberg, S. T. (2004). *Using the Transference in Psychotherapy*. Plymoth: Rowman and Littlefield.

Haraway, D. J. (2008). *When Species Meet*. Minneapolis: University of Minnesota Press.

Hersoug, A. G., Ulberg, R., and Høglend, P. (2014). When is transference work useful in psychodynamic psychotherapy? Main results of the first experimental study of transference work (FEST). *Contemp. Psychoanal.* 50, 156–174. doi: 10.1080/00107530.2014.880314

Jahn, E., Reindl, A., Müller, M., and Haddadin, S. (2019). Roboterassistenten als Helfer der Senioren im Alltag der Zukunft? Altenheim (accepted).

Liu, C., Liu, X., Wu, F., Xie, M., Feng, Y., and Hu, C. (2018). Using artificial intelligence (Watson for oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. *J. Med. Internet Res.* 20:e11087. doi: 10.2196/11087

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., et al. (2016). Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* 375, 655–665. doi: 10.1056/NEJMsa1507092

Marmarosh, C. (2012). Empirically supported perspectives on transference. *Psychotherapy* 49, 364–369. doi: 10.1037/a0028801

McLennan, S., Fiske, A., Celi, L., Müller, R., Harder, J., Ritt, K., et al. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence.* 2, 1–3. doi: 10.1038/s42256-020-0214-1

Müller, R., Clare, A., Feiler, J., and Ninow, M. (2021). Between a rock and a hard place: Farmer's perspectives on gene editing in livestock agriculture in Bavaria. *EMBO Rep.* 22:e53205. doi: 10.15252/embr.202153205

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Oudshoorn, N., and Pinch, T. eds. (2003). *How Users Matter: The Co-construction of Users and Theory*. Cambridge: MIT Press.

Panch, T., Mattie, H., and Atun, R. (2019a). Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* 9:010318. doi: 10.7189/jogh.09.020318

Panch, T., Mattie, H., and Celi, L. A. (2019b). The "inconvenient truth" about AI in healthcare. *NPJ Digit. Med.* 2:77. doi: 10.1038/s41746-019-0155-4

Parth, K., Datz, F., Seidman, C., and Loeffler-Stastka, H. (2017). Transference and countertransference: a review. *Bull. Menn. Clin.* 81, 167–211. doi: 10.1521/bumc.2017.81.2.167

Prasko, J., Diveky, T., Grambal, A., Kamaradova, D., Mozny, P., Sigmundova, Z., et al. (2010). Transference and countertransference in cognitive behavioral therapy. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech. Repub.* 154, 189–197. doi: 10.5507/bp.2010.029

Rein, B. A., McNeil, D. W., Hayes, A. R., Hawkins, T. A., Ng, H. M., and Yura, C. A. (2018). Evaluation of an avatar-based training program to promote suicide prevention awareness in a college setting. *J. Am. Coll. Heal.* 66, 401–411. doi: 10.1080/07448481.2018.1432626

Sachan, D. (2018). Self-help robots drive blues away. *Lancet Psychiatry* 5:547. doi: 10.1016/S2215-0366(18)30230-X

Suszek, H., Wegner, E., and Maliszewski, N. (2015). Transference and its usefulness in psychotherapy in the light of empirical evidence. *Ann. Psychol.* 18, 345–380. doi: 10.18290/rpsych.2015.18.3-4en

Triberti, S., Durosini, I., and Pravettoni, G. (2020). A 'third wheel' effect in health decision making involving artificial entities: a psychological perspective. *Front. Public Health* 8:117. doi: 10.3389/fpubh.2020.00117

Tudor Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y. L., et al. (2020). Conversational agents in health care: scoping review and conceptual analysis. *J. Med. Internet Res.* 22:e17158. doi: 10.2196/17158

Ulberg, R., Hummelen, B., Hersoug, A. G., Midgley, N., Høglend, P., and Dahl, H.-S. (2021). The first experimental study of transference work–in teenagers (FEST–IT): a multicentre, observer- and patient-blind, randomised

controlled component study. *BMC Psychiatry* 21:106. doi: 10.1186/s12888-021-03055-y

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.

frontiers in Psychology

REVIEW
published: 28 September 2021
doi: 10.3389/fpsyg.2021.710982

Human, All Too Human? An All-Around Appraisal of the "Artificial Intelligence Revolution" in Medical Imaging

Francesca Coppola[1,2†], Lorenzo Faggioni[3†], Michela Gabelloni[3], Fabrizio De Vietro[3], Vincenzo Mendola[3], Arrigo Cattabriga[1], Maria Adriana Cocozza[1], Giulio Vara[1], Alberto Piccinino[1], Silvia Lo Monaco[1], Luigi Vincenzo Pastore[1], Margherita Mottola[4], Silvia Malavasi[4], Alessandro Bevilacqua[4], Emanuele Neri[2,3‡] and Rita Golfieri[1‡]*

[1] Department of Radiology, IRCCS Azienda Ospedaliero Universitaria di Bologna, Bologna, Italy, [2] SIRM Foundation, Italian Society of Medical and Interventional Radiology, Milan, Italy, [3] Academic Radiology, Department of Translational Research, University of Pisa, Pisa, Italy, [4] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Artificial intelligence (AI) has seen dramatic growth over the past decade, evolving from a niche super specialty computer application into a powerful tool which has revolutionized many areas of our professional and daily lives, and the potential of which seems to be still largely untapped. The field of medicine and medical imaging, as one of its various specialties, has gained considerable benefit from AI, including improved diagnostic accuracy and the possibility of predicting individual patient outcomes and options of more personalized treatment. It should be noted that this process can actively support the ongoing development of advanced, highly specific treatment strategies (e.g., target therapies for cancer patients) while enabling faster workflow and more efficient use of healthcare resources. The potential advantages of AI over conventional methods have made it attractive for physicians and other healthcare stakeholders, raising much interest in both the research and the industry communities. However, the fast development of AI has unveiled its potential for disrupting the work of healthcare professionals, spawning concerns among radiologists that, in the future, AI may outperform them, thus damaging their reputations or putting their jobs at risk. Furthermore, this development has raised relevant psychological, ethical, and medico-legal issues which need to be addressed for AI to be considered fully capable of patient management. The aim of this review is to provide a brief, hopefully exhaustive, overview of the state of the art of AI systems regarding medical imaging, with a special focus on how AI and the entire healthcare environment should be prepared to accomplish the goal of a more advanced human-centered world.

Keywords: artificial intelligence, medical imaging, ethics, medico-legal issues, patient data, communication, psychology

OPEN ACCESS

Edited by:
Ilaria Durosini, European Institute of Oncology (IEO), Italy

Reviewed by:
Michelangelo Casali, University of Milan, Italy
Robertas Damasevicius, Silesian University of Technology, Poland

***Correspondence:**
Lorenzo Faggioni
lfaggioni@sirm.org

†These authors have contributed equally to this work and share first authorship

‡These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *18 May 2021*
Accepted: *02 September 2021*
Published: *28 September 2021*

Citation:
Coppola F, Faggioni L, Gabelloni M, De Vietro F, Mendola V, Cattabriga A, Cocozza MA, Vara G, Piccinino A, Lo Monaco S, Pastore LV, Mottola M, Malavasi S, Bevilacqua A, Neri E and Golfieri R (2021) Human, All Too Human? An All-Around Appraisal of the "Artificial Intelligence Revolution" in Medical Imaging. Front. Psychol. 12:710982. doi: 10.3389/fpsyg.2021.710982

INTRODUCTION

The term artificial intelligence (AI) was coined in 1956 to differentiate the intelligence of machines (generated by software development programs) from natural, human intelligence. In the past decade, AI algorithms have begun to influence many activities based on computer platforms, having a significant impact on daily life. These new technologies have aroused great interest among biomedical scientists, since AI has proven to be able to simplify the work of researchers and healthcare professionals, and to provide crucial information for the management of patients (e.g., early diagnosis, prediction of individual prognosis, and therapy personalization), which would realistically be very difficult or impossible to obtain without the support of such systems (Yu et al., 2018).

Radiology is one of the medical specialties with a greater interest in AI, since the latter can offer radiologists new tools for quantitative analysis and image interpretation in addition to offering automation and standardization of processes and procedures, which allow saving time and effort during fatiguing and/or repetitive tasks, improving diagnostic performance, and optimizing the overall workflow (Curtis et al., 2018; Hosny et al., 2018; Pesapane et al., 2018a; Yu et al., 2018; European Society of Radiology (ESR), 2019b; Grassi et al., 2019). However, this enthusiasm is paralleled by concerns of psychological, ethical, and medico-legal nature (including those related to the involvement of AI systems in patient management and the responsibilities that this may entail), as well as by the fear that AI could revolutionize radiologists' jobs, possibly threatening their existence as specific professional figures (Gong et al., 2019; Savadjiev et al., 2019; van Hoek et al., 2019). More generally, it has been emphasized that as long as AI becomes more and more autonomous (e.g., able to talk, "think," and actively participate in decision making), its role within complex relationships such as those between patients and physicians may be unclear to the human interlocutors, and new obstacles to decision making could arise due to AI acting as a "third wheel" between them (Triberti et al., 2020).

This review illustrates the main characteristics of AI systems and the current issues related to their use in the radiology profession.

ARTIFICIAL INTELLIGENCE: BASIC CONCEPTS

Broadly speaking, AI encompasses the ability of hardware and software devices to autonomously mimic activities which have traditionally been deemed as specific to humans, such as learning and thinking. More than 60 years after its inception, AI has recently come back under the spotlight owing to the increasing availability of relatively low-cost computers capable of processing large amounts of data in real time, enabling the practical implementation of AI systems. In the medical field, AI mainly refers to the ability of systems to detect and analyze data related to patient clinical information and management with the aim of accomplishing a predetermined goal (Savadjiev et al., 2019).

Artificial intelligence systems can be broadly classified into strong (or general) and weak (or restricted). Strong AI systems can apply AI to resolve any problem, and as such, aim to mimic human intelligence. Conversely, weak AI denotes systems from which humans can take advantage to efficiently perform specific tasks (Zackova, 2015; Brink et al., 2017; Park and Park, 2018; European Society of Radiology (ESR), 2019b). More specifically, weak AI systems can improve their intrinsic ability to solve problems autonomously by means of progressive learning, starting from acquired information. This category includes most systems used in practice which are based on different machine learning (ML) techniques, including those of bio-inspired artificial neural networks.

Machine Learning

In 1959, Arthur Samuel gave a boost to the development of weak AI by introducing the concept of "ML," defined as a subclass of AI systems which help the machine to learn and make decisions based on the data. To this end, the machine builds its own model from a subset of data used for training (Bishop, 2006; Litjens et al., 2017). Consequently, ML can make predictions on new data based on previous training without the need of being specifically programmed or recall previously defined models.

Another notable feature of ML is that the system performance increases with increasing experience of the system itself. In classic ML (which is used for classifying and interpreting data related to image analysis), data are labeled by human experts and organized according to their properties using statistical methods (Chartrand et al., 2017). In order for an ML algorithm to successfully reproduce the process of analyzing an image (e.g., a chest X-ray) by a radiologist, it must first be trained with a supervised approach starting from different labeled learning datasets (which contain many heterogeneous types of radiographic abnormalities), reinforced with different datasets each containing a class of abnormal findings (e.g., cardiac, mediastinal, pulmonary, and bone) and, if necessary, additionally reinforced with specific datasets for various subclasses of anomalies (e.g., congenital heart disease).

In general, ML is the highest expression of the power of a computer system. However, as in human learning, ML can also encounter some problems. For example, if the training dataset is poorly representative of the characteristics to be analyzed, an ML algorithm could learn from the training dataset in too much detail, leading to the problem of overfitting. In this case, non-significant statistical fluctuations of the same sample are cataloged by the learning model as separate data, which subsequently causes a worsening of the performance in analyzing new data (Duda et al., 2001; European Society of Radiology (ESR), 2019b). In diagnostic imaging, overfitting can be amplified by the possibility of non-pathological anatomical variants (such as accessory bones, or congenitally absent or hypoplastic structures).

The need for numerical accuracy and precision in processing radiological images represents one of the main challenges for the applications of ML systems in diagnostic imaging. Accuracy is essential for addressing the complexity of semantic aspects (related to the enormous variety of normal and pathological

findings that an ML system could encounter in the analysis of images acquired on real patients) and technical issues due to differences among various imaging modalities. Another challenge is related to the large number of images which need to be processed from cross-sectional imaging modalities (even with the aid of semi-automatic algorithms), with magnetic resonance imaging (MRI) or multislice computed tomography (MSCT) being able to furnish hundreds or thousands of images per single dataset.

Artificial Neural Networks and Deep Learning

The term deep learning (DL) was first introduced in 1986 by Rina Dechter, and represents a form of ML which can yield better performance than classic ML (**Figure 1**). Compared to a traditional artificial neural network, in which the number of levels is limited and the nodes of one level are connected to those of the next level ("completely connected"), DL systems are generally made up of several specialized levels. The last levels are generally the only fully connected ones and combine functionalities learned to make decisions. Instead of requiring labeling or engineering of the properties, DL algorithms independently learn the most suitable characteristics for classifying the data provided, depending on the specific task (Chartrand et al., 2017; Philbrick et al., 2018).

The commonest approach in image processing is represented by convolutional neural networks (CNNs), a particular type of neural network developed for the recognition of patterns within images, which can accept two- or three-dimensional images as input (Domingos, 2012). While the first CNN was implemented in 1980 by Fukushima (1980), CNNs were formalized as we know them now by LeCun et al. (1998). The introduction of advanced graphic processing units with the ability to process enormous amounts of data in parallel has made CNNs an essential tool for the development of modern DL algorithms (Hinton and Salakhutdinov, 2006; Trebeschi et al., 2017).

The two factors which mainly affect the functionality of CNNs are the power of the hardware and, most importantly, the availability of adequate data for the learning process. If computer power increases progressively over years or months, and can therefore only be relatively limiting, time and cost constraints make it difficult to find a solution to the problem of the low availability of well-structured datasets for training, which represents an actual hurdle to the development and diffusion of these systems (Napel et al., 2018).

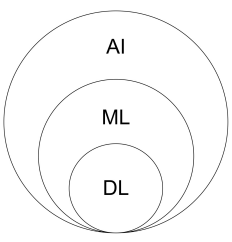

FIGURE 1 | Schematic representation of the hierarchical relationship between AI (which, in general, describes systems capable of imitating human behavior), ML (a subset of AI, which describes systems capable of learning by experience), and DL (a subset of ML, which describes systems based on neural networks).

Deep learning has proven to be a promising tool for the extraction of features from biomedical images (LeCun et al., 2015; Wang et al., 2017; Kermany et al., 2018; Lustberg et al., 2018). For that application, computational units are defined as levels integrated with each other to extract the intrinsic characteristics of the images. Using a CNN structured in a hierarchical manner, a DL system can, for example, extract the intrinsic characteristics of a neoplasm to build a model capable of providing prognostic or predictive information, having a clear potential impact on patient clinical management (Wang et al., 2019).

Artificial Intelligence in Medical Imaging: "Images Are More Than Pictures, They Are Data" (Gillies et al., 2016)

During its development, medical imaging has enjoyed great benefit from technological progress (Nance et al., 2013; Nguyen and Shetty, 2018), and the scientific relevance of the development of AI systems in radiology has been underscored by an ever-increasing number of publications on AI. For diagnostic imaging alone, the number of publications on AI has increased from about 100–150 per year in 2007–2008 to 1000–1100 per year in 2017–2018 (Tang, 2020).

Artificial intelligence systems can support medical decision-making processes related to requests for imaging tests, not only by means of the evaluation of the patient's medical record and the accuracy of the radiological examinations, but also by guiding the choice of the most suitable diagnostic modality. Of note, AI algorithms can be programmed to work in keeping with the appropriateness criteria developed and approved by scientific societies (such as those developed by the American College of Radiology) in order to maximize the adherence to validated criteria (Blackmore and Medina, 2006; American College of Radiology (ACR), 2021 Reporting and Data Systems).

Furthermore, AI has opened new perspectives on how to make the most of the information which can be obtained from biomedical imaging for a more in-depth understanding of the various pathological processes, aimed at more effective diagnostic and therapeutic management. Once trained with appropriate learning datasets, AI systems can analyze biomedical images with the aim of recognizing specific characteristics (either visible or invisible to the human eye) and build probabilistic models capable of detecting abnormal findings (Dodd, 2007; Sardanelli et al., 2010).

Automated image interpretation is one of the potential radiological applications of AI which has been received with the greatest enthusiasm. Rajpurkar et al. (2018) illustrated an AI algorithm with a comparable accuracy to that of human radiologists for diagnosing pneumonia on chest X-rays in a public dataset. Similar experiences have been reported for the detection of vertebral fractures on plain spinal radiography (Murata et al., 2020), the diagnosis of tuberculosis

(Lakhani and Sundaram, 2017), and the estimation of bone age (Dedouit et al., 2015). More generally, different DL methods have been applied to biomedical image analysis (Ierardi et al., 2016; Trimboli et al., 2018; Villanueva-Meyer et al., 2019) and successfully used with various imaging modalities, such as breast (Kallenberg et al., 2016; Zanotel et al., 2018; Geras et al., 2019; Hickman et al., 2021) and cardiac imaging (van Assen et al., 2020), MSCT (Lerouge et al., 2015; Kooi et al., 2017; Hu et al., 2020), MRI (Havaei et al., 2017; Ariji et al., 2019; **Figure 2**), as well as in interventional radiology (Gurgitano et al., 2021). AI can also be helpful to quantify lung involvement and predict prognosis in patients with COVID-19 pneumonia (Belfiore et al., 2020; Akram et al., 2021; Cappabianca et al., 2021), and Harmon et al. (2020) recently found that a series of DL algorithms trained in a diverse multinational cohort of 1280 patients can achieve up to 90.8% accuracy, with 84% sensitivity and 93% specificity in detecting COVID-19 pneumonia on chest CT examinations of 1337 patients. Other AI applications allow prioritizing the reporting of certain exams (e.g., urgent brain CT scans in patients with hemorrhagic stroke), thus optimizing the workflow and avoiding diagnostic delays, especially in situations in which members of the radiology department are busy with other tasks (Ngo et al., 2017; European Society of Radiology (ESR), 2019b). However, there are currently no commercial

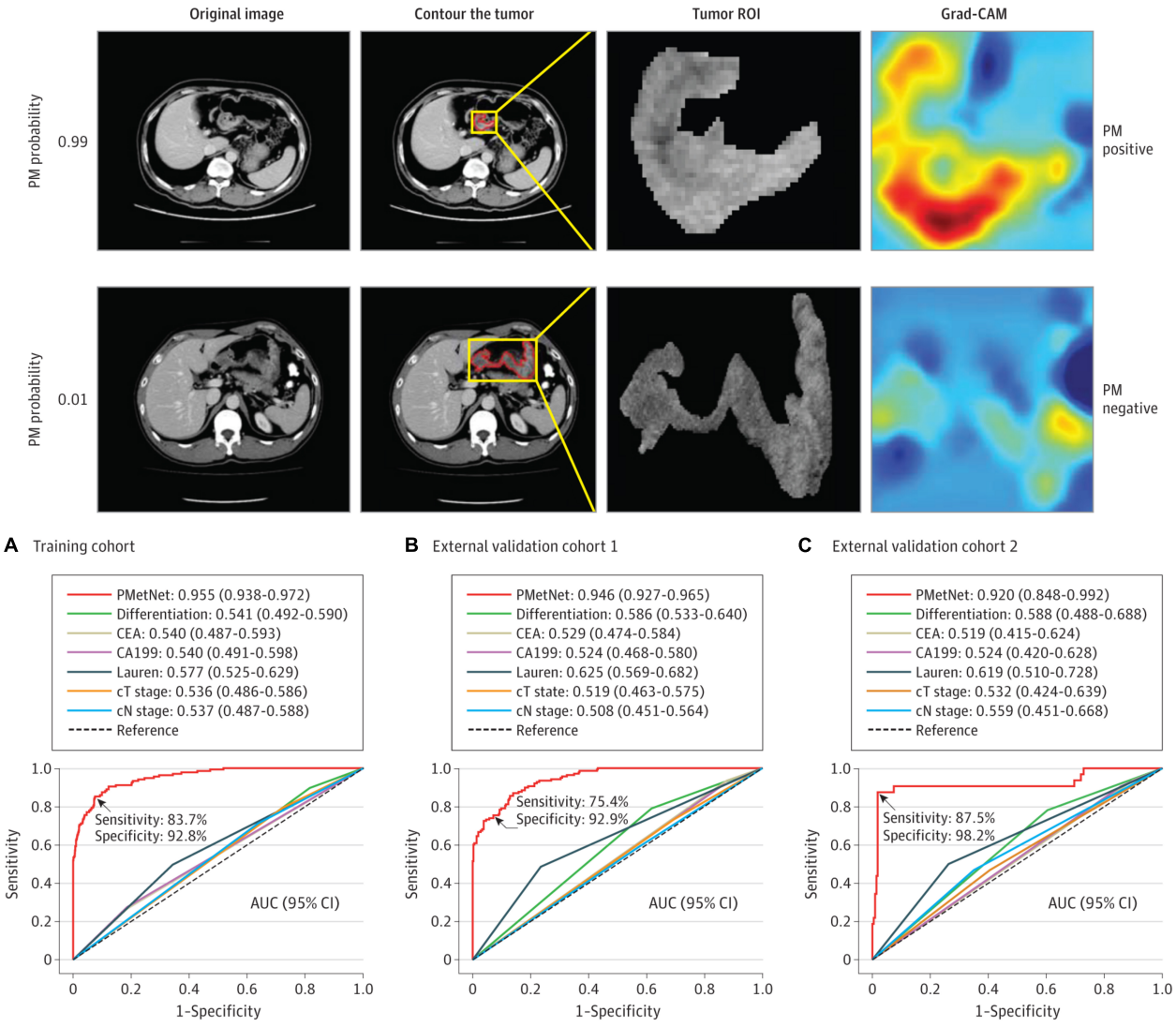

FIGURE 2 | Evaluation of a CNN-based DL algorithm (PMetNet) for detection of occult peritoneal metastases (PM) from gastric cancer on preoperative CT images. (Upper) Following tumor segmentation on original CT images (tumor contours highlighted in red, tumor region of interest [ROI] shown within magnified tumor location [boxed in yellow]), the gradient-weighted class activation mapping (Grad-CAM) was used to highlight which areas of the CT image are important in generating a particular prediction. The numbers 0.99 and 0.01 represent the predicted probability of PM. (Lower) Area under the receiver operating characteristic curves (AUCs) derived from the PMetNet and clinicopathological factors for the diagnosis of occult PM in the training cohort and in two validation cohort. The discrimination performance of PMetNet was substantially higher than conventional clinicopathological factors, and in multivariable logistic regression analysis, PMetNet was an independent predictor of occult PM. Adapted from Jiang et al. (2021) under Creative Commons Attribution 4.0 (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

solutions available which can independently interpret images and generate a report.

The ability of extracting structured and categorized data from existing radiological archives [radiology information system (RIS) and picture archiving and communication system (PACS)] is an essential requirement for the development and dissemination of AI in radiological environments. In fact, the training of AI systems usually requires enormous amounts of data which should be as accurate and correctly categorized as possible. However, to date, most radiological reports are written in the form of an unstructured narrative text which greatly complicates the extraction of information, even if the aim is to create AI systems based exclusively on clinical data. The latter could be resolved with the adoption of structured reporting (SR), which, if properly implemented, would allow the exchange of information with a common lexicon and semantics. While there is extensive evidence that SR has several advantages over conventional unstructured reporting (including better clarity, improved communication with patients and referring physicians, higher productivity, and ease of data mining), it has seen a relatively slow diffusion so far due to radiologists' fears, among others, that it could diminish their autonomy and professional reputation with respect to patients or non-radiologist specialists (Marcovici and Taylor, 2014; Faggioni et al., 2017; Coppola et al., 2021). In this context, constructive interaction between medical radiologists and other specialists, industries, and institutions would be desirable to promote a large-scale dissemination of the SR, offering a decisive stimulus for additional development of AI in the radiological field (Bosmans et al., 2015; Pinto Dos Santos and Baeßler, 2018; European Society of Radiology (ESR), 2019b).

A topic of interest for both the biomedical industry and research is the prospect of using AI to optimize biomedical image acquisition protocols, with potential advantages in terms of patient safety and health management costs. For example, some AI algorithms allow obtaining equivalent or even superior results as compared to commercial non-AI-based solutions for noise reduction in MSCT and positron emission tomography (PET) examinations, allowing the acquisition of diagnostic images with a significantly lower radiation exposure than conventional protocols (Zhu et al., 2018; Shan et al., 2019).

Radiomics is a field of research which has become very popular in the era of modern precision medicine. Radiomics refers to the established use of ML techniques applied to the analysis of radiological images. Radiomics is often defined as "the extraction of a large number of quantitative features from conventional biomedical images in order to obtain data that can be used in clinical decision support systems to improve diagnostic, prognostic, and predictive accuracy" (Choy et al., 2018). A radiomic model can reveal the value of biomarkers extracted from the images (which are quantifiable by means of applying mathematical and statistical models, even of considerable complexity), but it can also extend to a so-called hybrid model, including other data not from images (e.g., from clinical data or laboratory parameters). In any case, these models are able to provide information not obtainable with standard radiological semeiotics, such as those related to the early response to treatment, the prediction of the biological aggressiveness of a neoplasm, the existence of molecular targets for any targeted therapies, up to the prediction of the individual prognosis, and the personalization of therapies (Gillies et al., 2016; Lambin et al., 2017; Choy et al., 2018; Bi et al., 2019; Liu et al., 2019; Rogers et al., 2020; Zerunian et al., 2021). A notable feature of radiomics consists of the possibility of obtaining, in a repeatable and non-invasive way, information about a tissue in its entirety in contrast, for example, to what happens with a classic biopsy which is invasive and limited to a portion of tissue, with the risk of collecting a sample that is not representative of the heterogeneity of the lesion (Lambin et al., 2017; Abdollahi et al., 2019; Nazari et al., 2020; Zerunian et al., 2021). The radiomics approach can also be extended to the analysis of the genetic structure of a tissue (e.g., neoplastic), which is referred to as radiogenomics (King et al., 2013; Pinker et al., 2018; Story and Durante, 2018; Gabelloni et al., 2019; Lo Gullo et al., 2020). A radiomic biomarker is made up of a set of characteristics (or features) extracted from the image, it is represented by a mathematical equation, known as a "signature," and its value can be calculated using dedicated programs, starting from images acquired using routine protocols.

However, even if many of these tools are easy for operators to use and allow extracting radiomic features in a relatively short time, their diffusion is currently still limited by several factors. These include the enormous amount of work often necessary for image segmentation and the difficulty in ensuring the adequate quality of the data entered to obtain consistent results, also considering the inevitable differences both between the image acquisition protocols, and different machines and imaging centers (Stoyanova et al., 2016).

ARTIFICIAL INTELLIGENCE AND HUMANS

Several authors have hypothesized that AI systems might shortly be able to replace medical radiologists in their professional activity (Rizzo et al., 2018). The key question is: will AI be able to replace radiologists in the observation, characterization, and quantification tasks that they currently accomplish using their cognitive skills? (Beregi et al., 2018; Tajmir and Alkasab, 2018; Mendelson, 2019; Miller and Brown, 2019; Rubin, 2019). The short answer is: NO. However, as argued by Dr. Curtis Langlotz at the European Congress of Radiology in 2018: "AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't" (Krittanawong, 2018), and this concept could be generalized to all fields of healthcare (Meskó et al., 2018). In this context, it is important to point out that the final decision regarding patient diagnosis is still autonomous and the responsibility lies with the radiologist, not AI systems. What is most likely to change will be the use of information not only derived from morphological analysis in the formulation of the diagnosis, but also from the numerical values provided by AI. These refer directly to statistically significant distributions of the pixel values of the image which are not perceptible to the naked eye.

Both radiologists and AI systems must follow essential rules and principles for optimal patient management. Several issues are

related to the proper use of AI in clinical practice and include (but are not limited to) the following:

- Data (including generation, recording, maintenance, processing, dissemination, sharing, and use)
- AI algorithms used to process patients' data for a specific task
- Practices (including responsible innovation, programming, security, formulation, and implementation of ethical solutions)
- Communication (including the tools through which the information obtained from AI systems is provided to patients, as well as the management of psychological problems arising from them, which cannot be handed over to a computer system) (Meskó et al., 2018).

Various aspects of data ethics can be recognized, including informed consent, privacy and data protection, data ownership, objectivity in managing data, and the likelihood that a gap may exist between those who have the resources to manage and analyze large amounts of data and those who do not. In addition, the operation of AI systems integrated into big data networks raises ethical and legal issues related to patient-specific consent, data sharing, privacy and security protection, and the availability of multi-layered access to fully or partially anonymized health information.

Artificial Intelligence Overconfidence and Medico-Legal Issues

As stated by Kulikowski (2019), "whether a good ethical human can work with an AI and remain ethical is a major open problem for all of us that will have to be confronted not only scientifically, but also in a socially acceptable and humanistic way in clinical informatics." Hence, ethics should always guide radiologists (and physicians in general) in deciding when to rely on AI, so as to avoid improper applications of it which may have a harmful impact on both healthcare operators and patients.

One of the main biases which can hamper the use of AI in diagnostic imaging is the automation bias, which can be defined as the propensity to favor a machine-generated diagnosis over evidence derived from scientific knowledge and the physician's own expertise. This leads to the so-called omission and commission errors. Omission errors occur when the physician, deeming AI flawless, does not notice (or outright ignores) the fallacy of one of its tools. On the other hand, commission errors occur when a machine's decision is accepted, even in the face of contrary evidence. The risks of automation bias can be amplified in realities which suffer from a lack of medical personnel, since there may not be any radiologist to double-check the AI results (Geis et al., 2019). It has also been observed that automation could engender overreliance by its users (due to its advantages in terms of increased efficiency), and in the long term, lead to the so-called deskilling, with physicians losing their ability to autonomously perform tasks which have become automated (Cabitza et al., 2017). Harada et al. (2021) performed a randomized controlled study aimed to explore the prevalence of AI diagnoses in physicians' differential diagnoses when using an AI-driven diagnostic decision support system (DDSS) based on the information entered by the patient before the clinical encounter, showing that at least 15% of physicians' differential diagnoses were affected by the differential diagnosis list in the AI-driven DDSS. While many clinicians hope that AI will free them to focus on patient interaction, research on the overreliance of technology in medicine has found that the increased use of electronic health records has led to a prioritization of physician–technology interactions over physician–patient interactions, leading to decreased patient satisfaction, a scenario that could foreshadow the role of AI in patient care (Lu, 2016; Ross and Spates, 2020).

There is a still highly unmet need for specific guidelines, policies, and recommendations offering an ethical framework that can guide the use and implementation of AI technologies in an increasingly broad spectrum of clinical applications, which are progressively emerging as an effect of technological evolution, but also carry substantial psychological and ethical implications. Some of such potential applications include, for instance, AI in assisted reproductive technologies for human embryo selection *in vitro* fertilization (Dirvanauskas et al., 2019), and optimization of clinical trials of innovative stem cell and gene therapies in pediatric patients by precise planning of treatments, simplifying patient recruitment and retention, and lowering their complexity and costs (Sniecinski and Seghatchian, 2018). However, despite efforts by scientists, healthcare professionals, administrative managers, and lawmakers, so far very few countries worldwide have adequate and critical governance frames allowing best understanding and steering AI innovation trajectories in healthcare (Dzobo et al., 2020).

Such scenario is further complicated by the sweeping speed at which AI techniques are being developed or sometimes used, even before the publication of appropriate policies and guidelines, which might leave users confused about how to best integrate this new technology into their practice. This implies that updated regulatory policies and continuing education of all users (including adequate information to patients about the purposes, rights, and legal terms related to the use of AI for their health management) should be promoted, as AI systems are poised to become more widely available, complex and powerful. To this purpose, it is noteworthy that the majority of Singaporean radiology residents joining a national multiprogram survey thought that since AI will drastically change radiology practice, AI/ML knowledge should be taught during residency (84.8% of survey participants), and this was as important as imaging physics and clinical skills/knowledge curricula (80.0 and 72.8%, respectively) (Ooi et al., 2021). From a psychological standpoint, it has been observed that openness to experience is associated with higher trust toward robots and AI, as well as having a degree in technology or engineering, exposure to robots online, and robot use self-efficacy (Oksanen et al., 2020), highlighting the importance of technology knowledge in addition to personal differences in building AI confidence.

A key medico-legal aspect regarding the use of AI in healthcare is the responsibility for the decision-making processes upon which the patient's health depends. In the absence of specific

regulations, there may be ethical and medico-legal issues where an AI system is involved in the process and may suggest solutions (right or wrong); however, the final decision (also right or wrong) is, and will always be, the responsibility of the physician who is legally responsible (Mittelstadt and Floridi, 2016; Price et al., 2019, 2021; Reddy et al., 2019; Neri et al., 2020). Price et al. (2019, 2021) provided an in-depth analysis of the potential legal outcomes related to the use of AI in healthcare under current law (**Figure 3**).

Another problem is the opacity related to AI models being mostly a "black box" without a universal understanding of their inner workings, which are not acceptable for decision support solutions (especially in healthcare) and may lead to ethical and legal risks and liability issues, as well as undermine patients' and physicians' confidence into AI (Valiuškaitė et al., 2020; Tohka and van Gils, 2021). Putting this issue in the context of radiological practice, radiologists would be asked to monitor AI system outputs and validate AI interpretations, so they would risk carrying the ultimate responsibility of validating something they cannot understand (Neri et al., 2020). There is a clear difference between statistical and clinical validation, and hence achieving adequate informed consent is problematic when the algorithmic decision-making process is opaque to clinicians, patients, or courts (Martinez-Martin et al., 2018; Arnold, 2021). Actually, while the number of published articles on the applications of AI in medical imaging and other medical specialties is steadily increasing, so far only few AI applications have been validated for clinical use, partly due to the difficulty of using AI projects on a large scale in real-life clinical practice, poor adherence to scientific quality standards (Nagendran et al., 2020; Park et al., 2020), and clinical validation issues. While the ongoing development of AI has generated considerable hype and highly optimistic expectations in the scientific community, such enthusiasm is often curbed by the reality of proper performance assessment, which is not trivial (requiring an understanding of the problem and data) and is often costly (data needs to be reserved) and time consuming (Tohka and van Gils, 2021). According to some researchers, the overall problem is wide and ultimately originates from inappropriate experimental design and hypothesis testing procedures, including so-called Hypothesizing After the Results are Known (aka HARKing) practices (Gencoglu et al., 2019; Tohka and van Gils, 2021).

The process of technology and infrastructure development requires close multidisciplinary cooperation among governmental institutions, research centers, healthcare professionals, and industry. In this context, a potential solution could involve enrolling AI experts in radiology units to act as a link between AI systems and radiologists who, in turn, should be trained to use those systems independently. To ensure that the approach to an innovative and potentially destructive technology is properly managed, radiologists will have to develop strategies not based on prejudice, but specifically adapted to the peculiar characteristics of AI systems, their technical/scientific development, their implementation by the industry, and their actual diffusion.

Data Confidentiality and Regulation Policies

Using AI systems for diagnosing diseases (including life-threatening or invalidating diseases, with a potentially dramatic impact on the physical and psychological well-being of patients and their families) and finding the most appropriate therapeutic approach implies that these systems should have the highest grade of reliability and dependability. To this end, the following requirements should be met:

- The largest possible amount of data (both imaging and non-imaging-related) should be shared.
- The quality and integrity of these data should be as high as possible, avoiding errors due to poor image quality, mislabeling, and over- and underfitting.
- The anonymity and depersonalization of data must be guaranteed so as to ensure that the individual(s) who has/have consented to their use can be traced.

In recent years, several authors have discussed the requisites for the reliable innovation of AI by means of attaining the

Scenario	AI recommendation	AI accuracy	Physician action	Patient outcome	Legal outcome (probable)	Empirical Study Findings (Tobia et al.)
1	Standard of care	Correct	Rejects	Injury	Liability	☑ ★
2		Incorrect (standard of care is incorrect)	**Follows**	Injury	No liability	☑ ★★★★
3	Nonstandard care	Correct (standard of care is incorrect)	Rejects	Injury	*No liability*	⚠ ★★
4		Incorrect	**Follows**	Injury	*Liability*	☒ ★★★

FIGURE 3 | Comparison of potential legal outcomes under current law according to analysis of Price et al. (2019) and empiric study findings of Tobia et al. (2021). ★ = agreement that physician decision was reasonable (highest is ★★★★; lowest is ★). Greater agreement indicates lower likelihood of liability; ☑ = study results confirming Price et al.'s analysis of current tort law; ⚠ study results suggesting that jury outcome may also be liability; ☒ study results suggesting that jury might decide no liability. Reproduced from Price et al. (2021). © SNMMI.

most important ethical principles. These principles have been embodied in the laws of the various countries throughout the world, and while an overarching political vision and long-term strategy for the development of a "good AI" society are currently lacking, this process has been characterized in broad terms by the search for a tradeoff between AI technological innovation and regulation (Cath et al., 2018; Pesapane et al., 2018b; Monreale, 2020).

As mentioned by Cath et al. (2018), the political attitude of the United States toward the implementation of AI in healthcare can be summarized by the sentence: "Letting a thousand flowers bloom," whereas that of the European Union (EU) can be described as: "European standards for robotics and AI," and the UK approach ("Keep calm and commission on") stands in an approximately intermediate position between the US and the EU policies. In the United States, a "Silicon Valley" model oriented toward the more liberal regulation of ethical issues and based on the "move fast, break things first, apologize later" approach has prevailed (Armitage et al., 2017) and, in 2020, the Trump administration published a guide for AI application which discouraged any action resulting in limiting innovation and technological progress (Vought, 2020). At the other extreme, the EU policy points to strictly codifying regulations based on ethical principles. While this policy has raised the objection that it could hinder AI innovation, the European Commission sees the codification of ethical principles for AI use as a competitive advantage which will promote consumer confidence in their products and harmonize their adoption across the EU (Monreale, 2020). Data protection in the EU is regulated by General Data Protection Regulation (GDPR) EU 2016/679 and other EU directives for confidential data protection, which is of paramount importance in case of the AI/ML systems being trained on personal healthcare data (Voigt and Von dem Bussche, 2017). In this respect, the principles of "privacy by design" described by Cavoukian (2009) and updated by Monreale et al. (2014) could be applied, in the perspective of promoting research and innovation while taking care and full responsibility of the protection of the personal data, rights, and freedom of EU citizens.

It is especially important that the requirements for privacy protection are fulfilled during the process of data extraction for the training of ML algorithms, avoiding the potential risks related to illegal access either to confidential data during training or to the ML model used for clinical patient management (Brundage et al., 2018; Pesapane et al., 2018b; Monreale, 2020). On the other hand, current data protection laws may pose a significant limitation for researchers who develop and use ML algorithms, resulting in a lack of generalization of training which has so far prevented a more widespread application of such algorithms into clinical practice by healthcare providers across the world. This holds especially true for rare diseases, for which the accuracy of ML algorithms could be limited due to the relatively small amount of data for algorithms to train on and data collection is inherently slow due to a low disease prevalence. Similarly, algorithms which predict outcomes from genetic findings may lack generalizability if there are only a limited number of studies in certain populations (Kelly et al., 2019; Abayomi-Alli et al., 2021; Tohka and van Gils, 2021).

Machine learning models are programmed based on de-identified data, i.e., those which do not directly allow identifying an individual in a univocal manner. Unfortunately, in some contexts, de-identification is not sufficient to protect a person's privacy, since individuals could be indirectly reidentified by means of the correlation of the de-identified data with public data (Jones, 2018), prompting the adoption of more advanced solutions aimed at fully protecting patient anonymity, such as k-anonymity (Sweeney, 2002). From a more general ethical and legal viewpoint, while patient data stored in electronic health records may be de-identified and, through data linkage, generate beneficial research outcomes, this may create a tension between beneficence (for the public) and private confidentiality, overriding contemporary notions of privacy and confidentiality according to the duty of "easy rescue," particularly in circumstances of minimal risk as defined by research regulators (Porsdam Mann et al., 2016; Arnold, 2021). Moreover, a study from the University of California, Berkeley, suggests that progress in AI has rendered the privacy standards set by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) obsolete (Na et al., 2018). The important conclusion is that privacy standards associated with the current legal and regulatory framework should be revisited and reworked, such that the advances of AI and its impact on data privacy as it pertains to healthcare are factored in (Na et al., 2018; Ahuja, 2019; Kulkarni, 2021).

An additional risk regarding a breach of patient confidentiality could derive from so-called "membership inference attacks," i.e., malicious attacks toward AI algorithms which are aimed at detecting the confidential data used to build the algorithm (Shokri et al., 2017). Actually, the implementation of AI systems means access to sensitive health data, which intrinsically always carries the risk of cyberattacks, posing a substantial risk on the privacy of patients (especially those with lower education and financial income; Bilen and Özer, 2021) and requiring a guaranteed level of robustness against such attacks (Catak et al., 2021; Zhou et al., 2021). Attacks on AI systems can undermine diagnostic accuracy, administer lethal drug doses, or sabotage critical moves in an operation, and in the area of diagnostic imaging, they can manipulate data entering AI systems (so-called "input attacks"), leading to false diagnosis and altered patient care and/or reimbursement (Finlayson et al., 2019; Kiener, 2020; Myers et al., 2020). The malware can obtain personal information by means of query and repersonalization of the data within the algorithm, and most strategies aimed at offering protection against such privacy violations rely on methods based on differential privacy, i.e., a privacy model based on the concept of data perturbation (Bugliesi et al., 2006). While several solutions have been proposed to forecast, prevent, and mitigate threats from malicious uses of AI technologies, a coordinated action of all involved stakeholders (including researchers, engineers, and AI users) has been advocated to manage what is expected to become a long-term equilibrium between AI attackers and defenders (Brundage et al., 2018).

Accessibility of Artificial Intelligence Services

An important issue that should receive special attention is related to the possibility that the access to AI systems may be not equal for all patients or healthcare professionals. In fact, smaller facilities and academic centers with fewer resources may lack the means to acquire (and skills to use) complex and more performing AI systems. Furthermore, if AI were to be developed and exclusively owned by large entities in the private sector, this would likely further restrict its spread to a wider public. Collaboration between academic institutions and the public and private sectors has been advocated to foster the development of both workforce and AI applications in healthcare (Mikhaylov et al., 2018; Ishii et al., 2020).

The so-called "digital divide" can be classified as global, social, and democratic (Srinuan and Bohlin, 2011); in any case, it invariably implies that affected subjects are excluded from the benefits of technological progress and innovation. A realist review in general practice by Huxley et al. (2015) showed that while digital communication technology offers increased opportunities for marginalized groups to access health care, it cannot remove all barriers to care for these groups, and actually they will likely remain disadvantaged relative to other population groups after their introduction. There is a risk that such phenomenon may occur in a previously unseen fashion with AI, with factors including age, gender, health condition, level of education, or financial income possibly leading to unequal access to AI systems.

An increasingly recognized issue is the potential for bias of AI systems with respect to certain population subgroups with a lack of diversity (e.g., in age, ethnicity, socioeconomic background, etc.) if algorithms have been developed on datasets which under- or over-represent them (Caliskan et al., 2017; Reddy et al., 2019; DeCamp and Lindvall, 2020; Hickman et al., 2021). Algorithmic bias may occur in ML systems for healthcare, perhaps predicting a greater likelihood of disease on the basis of gender or race when those are not actual causal factors (Davenport and Kalakota, 2019). Other issues could arise from discriminatory behaviors toward socially weaker individuals, from the need to gain the physicians and patients' trust in a context where AI systems process biomedical data and play a crucial role in clinical management, or from the duty of providing concrete rights of access to services to each patient. The communication of medical information, rules for the use of data, and requirements for institutional review committees may need to include new possibilities for patient data management (Kohli and Geis, 2018).

Communication and Psychological Issues

As outlined previously, while AI is supposed to offer radiologists substantial aid in their professional activity, predictions that it will replace radiologists in a more or less distant future are unfounded since the professional role of radiologists involves many tasks which cannot be accomplished by computers alone, including carrying out interventional radiology procedures, performing a clinical-radiological correlation in image interpretation, interpreting complex findings, and communicating them to colleagues and/or patients (Russell and Bohannon, 2015; Price et al., 2019). However, radiologists will have to improve their relationship with patients in the AI era to avoid any patient discomfort due to a lack of empathy and of a human reference figure during all the steps of a radiological procedure. These range from the patient's admission to the communication and discussion of imaging findings (the latter being a source of considerable psychological stress for patients and, hence, a task which could not be assigned to any, however perfect, AI algorithm). Moreover, reaching a diagnosis may often involve the use of multiple imaging techniques which are proposed by the radiologist (in combination with clinical and laboratory data, as well as with other non-radiological tests), and the overall interpretation of imaging findings is a complex task which requires a global assessment of the patient's condition and as such cannot be demanded to a computer system. In this context, the ability of AI systems to not only improve the detection and characterization of diseases (e.g., cancer) but also guide treatment and predict individual patient outcomes and prognosis (Bi et al., 2019; Rogers et al., 2020) can create additional issues related to the complexity of communicating and discussing topics with a high emotional impact (Butow and Hoque, 2020).

While AI allows saving time regarding the diagnostic and therapeutic decision process, the latter could actually be delayed if the role of AI is not taken into account in the consultation between physician and patient (Pravettoni and Triberti, 2019) (who may wish, and has the right to, know the implications of using AI in his/her clinical management), or if AI conclusions need to be revised by human doctors (especially when important decisions are to be made based on such conclusions) (Triberti et al., 2020). Moreover, as mentioned above, the poor explainability of most current AI systems (which are undoubtedly characterized by a high degree of complexity) and their lack of transparency could engender anxiety, distrust, or outright hostility with respect to AI in patients and clinicians (Bi et al., 2019). The relationship between patient and physician is a complex and profound psychosocial interaction characterized by mutual knowledge, trust, loyalty, and regard, so that human interaction will remain essential for patient-centered care due to the uniqueness of "human touch," consisting of peculiar features (such as empathy or the ability to be in tune with other people's thoughts and feelings) (Honavar, 2018). To this regard, it is known that a better communication between patients and physicians is associated with lower patient anxiety, fewer malpractice claims, and improved quality of life (Levinson et al., 2010). As to patients' trust in AI performance, Juravle et al. (2020) reported three online experiments showing that given the option of receiving their diagnosis from AI or human physicians, patients trusted those latter more for both first diagnoses and a second opinion for high risk diseases, and their trust in AI did not increase when they were told that AI outperformed the human doctor, but the trust in AI diagnosis increased significantly when participants could choose their doctor.

Owing to their pivotal role in the diagnostic process, radiologists are often the first healthcare professionals who are asked by patients about their imaging findings (and hence find

themselves to deal with patients' emotional reaction), and from whom patients expect a direct communication of their imaging findings (Berlin, 2007; Capaccio et al., 2010; Cox and Graham, 2020). The use of AI systems with the ability to provide additional information which may have a significant impact on patient management and overall life (e.g., eligibility to specific treatment options, prognosis, etc.) will entail for radiologists more stringent requirements in terms of communication skills and psychological balance, as well as a high degree of constructive interaction and feedback with other medical and non-medical specialists involved in patient care.

A detailed knowledge of the main features of AI (including its technical background, its current fields of application, and its psychological and legal implications) is a preliminary condition for its usage as a mature professional tool (Kobayashi et al., 2019; Savadjiev et al., 2019; Sogani et al., 2020). In a nationwide online survey among members of the Italian Society of Medical and Interventional Radiology (SIRM), most radiologists (77%) were favorable to the adoption of AI in their working practice, with a lower diagnostic error rate and work optimization being main perceived advantages, whereas the risk of a poorer professional reputation compared with non-radiologists was seen as one major downside (60% of survey respondents). However, about 90% of surveyed radiologists were not afraid of losing their job due to AI, and less than 20% of them were concerned that computers will replace radiologists for reporting of imaging examinations (Coppola et al., 2021). To this respect, it is worth mentioning that while most medical students surveyed by Gong et al. (2019) were discouraged from considering the radiology specialty out of anxiety that AI could potentially displace radiologists, in that same study prior significant exposure to radiology and high confidence in AI understanding were associated with a lower anxiety level, suggesting that professional education can have a significant impact on the psychological attitude of physicians toward AI.

Moreover, in the aforementioned SIRM survey and in a EuroAIM survey aimed at assessing the perceived impact of AI in radiology among European Society of Radiology (ESR) members, most respondents believed that if AI systems will allow radiologists to save time, such time should be used to interact with other clinicians or patients, thus improving personal interaction and communication (European Society of Radiology (ESR), 2019a; Coppola et al., 2021). Similar findings were reported in a French survey including 70 radiology residents and 200 senior radiologists, whose main expectations about AI included a lower risk of imaging-related medical errors and an increase in the time spent with patients (Waymel et al., 2019).

In light of the above, AI could alleviate radiologists' traditional work burden by undertaking tasks that could better be performed by computers, while giving them the opportunity to invest time and resources for other tasks that are better or uniquely accomplished by humans, such as interpreting imaging findings in the full width and complexity of a real clinical context, enhancing communication with patients and clinicians, supervising the correct operation and usage of AI systems, and being actively engaged in research (including AI-assisted data mining for big data handling and management of large-scale clinical trials) and quality optimization of the whole healthcare process. Like pathologists (who also extract medical information from images), radiologists will have an inescapable opportunity to leave once for all the stigma of "invisibility" which has often overshadowed the perception of their professional role by patients and clinicians in the past (Glazer and Ruiz-Wibbelsmann, 2011), and to take on a pivotal role in patient care as information specialists, adapting incrementally to AI and retaining their own services for cognitively challenging tasks and interaction with patients and clinicians (Jha and Topol, 2016; Recht and Bryan, 2017). Likewise, also clinicians need not fear AI as a potential enemy who could harm their professional reputation in the patients' eyes or their jobs in the future, but they should leverage its power to tackle computationally and labor-intensive tasks better than humans and to concentrate on those tasks which require human action (Ahuja, 2019). Therefore, an enhanced professional role could be envisaged for both radiologists and clinicians, requiring more advanced and specific skills (Recht and Bryan, 2017; Krittanawong, 2018; Ahuja, 2019; Waymel et al., 2019), despite fears that AI taking over professional tasks once performed by humans could, in the long run, lead to deskilling of human physicians (Bisschops et al., 2019; Campbell et al., 2020; Panesar et al., 2020). AI could actually help radiologists and clinicians make the most of their own specialty knowledge and competence in a medical science of rapidly increasing complexity (where "diseases do not respect boundaries" between medical specialties and require the cooperation of multiple specialists; Deo, 2021), avoiding misunderstandings and "turf wars" due to poor communication and confusion regarding their specialty-specific roles in patient management, and possibly fostering the adoption of AI-augmented multidisciplinary teams (including software engineers and data scientists among participants) for clinical decision making (Di Ieva, 2019; Lee and Lee, 2020; Martín-Noguerol et al., 2021).

Other potential issues of AI in the physician–patient relationship include misunderstanding (since a disagreement between the physician and AI can cause confusion, and the patient may not recognize who has the real authority in the care management) and alienation due to the physician or patient feeling excluded from the contribution of AI. To this regard, it should be considered that AI is deficient in emotional intelligence, whereas a physician has skills, beliefs, and subjective perceptions which can shape the communication with the patient, thus seeking an adequate patient's understanding of the disease and its related treatment options as the main aim of the communication process (Oh et al., 2017; Keskinbora, 2019; Pravettoni and Triberti, 2019; Triberti et al., 2020).

It has been observed that once digital and objective data will have become accessible to both caregivers and patients, the so-called "digital health" (of which AI is a major component) will lead to an equal level of physician–patient relationship with shared decision-making and a democratization of care (Meskó et al., 2017). However, it is possible that while some patients could accept or even require AI as an additional tool for decision making in their own medical care, others would not accept its use in decision-making (Meskó et al., 2018), thus stressing

the need for setting out shared policies aimed to a rational utilization of AI in patient management. A recent study on patients' perception about the use of AI for skin cancer screening as assessed by means of a semistructured interview revealed that most of them were favorable to AI and believed that it may improve the quality of care, but only if implemented in a manner which preserves the integrity of the human physician–patient relationship (Nelson et al., 2020). Again, direct physician–patient communication must be considered as an integral part of care delivery which cannot be substituted by a machine, and as Krittanawong has pointed out: "*AI cannot engage in high-level conversation or interaction with patients to gain their trust, reassure them, or express empathy, all important parts of the doctor–patient relationship*" (Krittanawong, 2018).

eXplainable Artificial Intelligence and Causability: Forthcoming Steps for Artificial Intelligence to Enter Maturity?

In conclusion, while it is undeniable that AI will sooner or later affect healthcare and the professional role and work of healthcare providers, physicians should neither uncritically accept nor unreasonably resist developments in AI, but they must actively engage and contribute to an iterative discourse to preserve humanitarian concerns in future models of care (Arnold, 2021). In this context, it is clear that the sustainable use of AI involves keeping in mind its fields of applicability and limitations, thus envisaging a future where its capabilities and advantages integrate (rather than supplant) human intelligence.

The main future goal is to make AI capable of interacting with operators in a meaningful and easily accessible manner. In this context, eXplainable Artificial Intelligence (xAI) has emerged as a new discipline which tries to fulfill the need for causability in the medical domain; in the same way that usability encompasses measurements for the quality of use, causability encompasses measurements for the quality of explanations produced by xAI. Multi-modal causability is especially important in the medical domain, since results are often achieved by means of multiple different modalities. The key for future human–AI interfaces is to map explainability with causability, and to allow a domain expert to ask questions so as to understand why AI has come up with a result, and also to ask "what-if" questions (counterfactuals) to gain insight into the underlying independent explanatory factors of a result (Holzinger, 2021).

AUTHOR CONTRIBUTIONS

FC, LF, and MG: conceptualization. MG, GV, AP, and SLM: methodology. FDV, VM, AC, MAC, and LVP: literature review. FC, LF, FDV, and VM: writing–original draft preparation. MG, MM, SM, and AB: writing–review and editing. EN and RG: supervision and guarantee of scientific integrity. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

The authors wish to thank Lorenzo Tumminello for providing insights about the psychological impact of AI on physicians and patients and Gerard Goldsmith for kindly revising the English language of the manuscript.

REFERENCES

Abayomi-Alli, O. O., Damaševičius, R., Maskeliūnas, R., and Misra, S. (2021). Few-shot learning with a novel voronoi tessellation-based image augmentation method for facial palsy detection. *Electronics* 10:978. doi: 10.3390/electronics10080978

Abdollahi, H., Mofid, B., Shiri, I., Razzaghdoust, A., Saadipoor, A., Mahdavi, A., et al. (2019). Machine learning-based radiomic models to predict intensity-modulated radiation therapy response, Gleason score and stage in prostate cancer. *Radiol. Med.* 124, 555–567. doi: 10.1007/s11547-018-0966-4

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* 7:e7702. doi: 10.7717/peerj.7702

Akram, T., Attique, M., Gul, S., Shahzad, A., Altaf, M., Naqvi, S. S. R., et al. (2021). A novel framework for rapid diagnosis of COVID-19 on computed tomography scans. *Pattern Anal. Appl.* 24, 965–965. doi: 10.1007/s10044-020-00950-0

American College of Radiology (ACR) (2021). *Reporting and Data Systems*. Available online at: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems [Accessed July 22, 2021]

Ariji, Y., Fukuda, M., Kise, Y., Nozawa, M., Yanashita, Y., Fujita, H., et al. (2019). Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. *Oral. Surg. Oral Med. Oral Pathol. Oral. Radiol.* 127, 458–463. doi: 10.1016/j.oooo.2018.10.002

Armitage, A., Cordova, A., and Siegel, R. (2017). *Design-Thinking: The Answer To The Impasse Between Innovation And Regulation. UC Hastings Research Paper No. 250*. Available online at: https://ssrn.com/abstract=3024176 or http://dx.doi.org/10.2139/ssrn.3024176 (Accessed August 22, 2017).

Arnold, M. H. (2021). Teasing out artificial intelligence in medicine: an ethical critique of artificial intelligence and machine learning in medicine. *J. Bioeth. Inq.* 18, 121–139. doi: 10.1007/s11673-020-10080-1

Belfiore, M. P., Urraro, F., Grassi, R., Giacobbe, G., Patelli, G., Cappabianca, S., et al. (2020). Artificial intelligence to codify lung CT in Covid-19 patients. *Radiol. Med.* 125, 500–504. doi: 10.1007/s11547-020-01195-x

Beregi, J.-P., Zins, M., Masson, J.-P., Cart, P., Bartoli, J.-M., Silberman, B., et al. (2018). Radiology and artificial intelligence: an opportunity for our specialty. *Diagn. Interv. Imaging* 99, 677–678. doi: 10.1016/j.diii.2018.11.002

Berlin, L. (2007). Communicating results of all radiologic examinations directly to patients: has the time come? *AJR Am. J. Roentgenol.* 189, 1275–1282. doi: 10.2214/AJR.07.2740

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J. Clin.* 69, 127–157. doi: 10.3322/caac.21552

Bilen, A., and Özer, A. B. (2021). Cyber-attack method and perpetrator prediction using machine learning algorithms. *PeerJ. Comput. Sci.* 7:e475. doi: 10.7717/peerj-cs.475

Bishop, C. M. (2006). *Pattern recognition And Machine Learning*. Available online at: https://cds.cern.ch/record/998831 [Accessed July 22, 2021].

Bisschops, R., East, J. E., Hassan, C., Hazewinkel, Y., Kamiński, M. F., Neumann, H., et al. (2019). Advanced imaging for detection and differentiation of colorectal neoplasia: European Society of Gastrointestinal Endoscopy (ESGE) guideline – update 2019. *Endoscopy* 51, 1155–1179. doi: 10.1055/a-1031-7657

Blackmore, C. C., and Medina, L. S. (2006). Evidence-based radiology and the ACR appropriateness criteria. *J. Am. Coll. Radiol.* 3, 505–509. doi: 10.1016/j.jacr.2006.03.003

Bosmans, J. M. L., Neri, E., Ratib, O., and Kahn, C. E. Jr. (2015). Structured reporting: a fusion reactor hungry for fuel. *Insights Imaging* 6, 129–132. doi: 10.1007/s13244-014-0368-7

Brink, J. A., Arenson, R. L., Grist, T. M., Lewin, J. S., and Enzmann, D. (2017). Bits and bytes: the future of radiology lies in informatics and information technology. *Eur. Radiol.* 27, 3647–3651. doi: 10.1007/s00330-016-4688-5

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). *The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv [CS.AI]*. Available online at: http://arxiv.org/abs/1802.07228 (accessed July 22, 2021).

Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I. (2006). "Automata, Languages and Programming," in *Proceedings of the 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006* (Berlin: Springer).

Butow, P., and Hoque, E. (2020). Using artificial intelligence to analyse and teach communication in healthcare. *Breast* 50, 49–55. doi: 10.1016/j.breast.2020.01.008

Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA* 318, 517–518. doi: 10.1001/jama.2017.7797

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Campbell, C. G., Ting, D. S. W., Keane, P. A., and Foster, P. J. (2020). The potential application of artificial intelligence for diagnosis and management of glaucoma in adults. *Br. Med. Bull.* 134, 21–33. doi: 10.1093/bmb/ldaa012

Capaccio, E., Podestà, A., Morcaldi, D., Sormani, M. P., and Derchi, L. E. (2010). How often do patients ask for the results of their radiological studies? *Insights Imaging* 1, 83–85. doi: 10.1007/s13244-009-0003-1

Cappabianca, S., Fusco, R., De Lisio, A., Paura, C., Clemente, A., Gagliardi, G., et al. (2021). Clinical and laboratory data, radiological structured report findings and quantitative evaluation of lung involvement on baseline chest CT in COVID-19 patients to predict prognosis. *Radiol. Med.* 126, 29–39. doi: 10.1007/s11547-020-01293-w

Catak, F. O., Ahmed, J., Sahinbas, K., and Khand, Z. H. (2021). Data augmentation based malware detection using convolutional neural networks. *PeerJ Comput. Sci.* 7, e346. doi: 10.7717/peerj-cs.346

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'Good Society': the US, EU, and UK approach. *Sci. Eng. Ethics* 24, 505–528. doi: 10.1007/s11948-017-9901-7

Cavoukian, A. (2009). *Privacy by design: the 7 foundational principles. Information and privacy commissioner of Ontario, Canada 5, 12.* Available online at: http://dataprotection.industries/wp-content/uploads/2017/10/privacy-by-design.pdf [Accessed July 22, 2021].

Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., et al. (2017). Deep learning: a primer for radiologists. *Radiographics* 37, 2113–2131. doi: 10.1148/rg.2017170077

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., et al. (2018). Current applications and future impact of machine learning in radiology. *Radiology* 288, 318–328. doi: 10.1148/radiol.2018171820

Coppola, F., Faggioni, L., Regge, D., Giovagnoni, A., Golfieri, R., Bibbolino, C., et al. (2021). Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol. Med.* 126, 63–71. doi: 10.1007/s11547-020-01205-y

Cox, J., and Graham, Y. (2020). Radiology and patient communication: if not now, then when? *Eur. Radiol.* 30, 501–503. doi: 10.1007/s00330-019-06349-8

Curtis, C., Liu, C., Bollerman, T. J., and Pianykh, O. S. (2018). Machine learning for predicting patient wait times and appointment delays. *J. Am. Coll. Radiol.* 15, 1310–1316. doi: 10.1016/j.jacr.2017.08.021

Davenport, T., and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc. J.* 6, 94–98. doi: 10.7861/futurehosp.6-2-94

DeCamp, M., and Lindvall, C. (2020). Latent bias and the implementation of artificial intelligence in medicine. *J Am. Med. Inform. Assoc.* 27, 2020–2023. doi: 10.1093/jamia/ocaa094

Dedouit, F., Saint-Martin, P., Mokrane, F.-Z., Savall, F., Rousseau, H., Crubézy, E., et al. (2015). Virtual anthropology: useful radiological tools for age assessment in clinical forensic medicine and thanatology. *Radiol. Med.* 120, 874–886. doi: 10.1007/s11547-015-0525-1

Deo, A. (2021). *Will AI deskill doctors?*. Available online at: https://healthcare2020plus.com/?p=98 [Accessed July 22, 2021].

Di Ieva, A. (2019). AI-augmented multidisciplinary teams: hype or hope? *Lancet* 394:1801. doi: 10.1016/S0140-6736(19)32626-1

Dirvanauskas, D., Maskeliūnas, R., Raudonis, V., Damaševičius, R., and Scherer, R. (2019). HEMIGEN: human embryo image generator based on generative adversarial networks. *Sensors (Basel)* 19:3578. doi: 10.3390/s19163578

Dodd, J. D. (2007). Evidence-based Practice in Radiology: Steps 3 and 4—Appraise and apply diagnostic radiology literature. *Radiology* 242, 342–354. doi: 10.1148/radiol.2422051679

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. New Delhi: Wiley-Interscience.

Dzobo, K., Adotey, S., Thomford, N. E., and Dzobo, W. (2020). Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. *OMICS* 24, 247–263. doi: 10.1089/omi.2019.0038

European Society of Radiology (ESR) (2019a). Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. *Insights Imaging* 10:105. doi: 10.1186/s13244-019-0798-3

European Society of Radiology (ESR) (2019b). What the radiologist should know about artificial intelligence - an ESR white paper. *Insights Imaging* 10:44. doi: 10.1186/s13244-019-0738-2

Faggioni, L., Coppola, F., Ferrari, R., Neri, E., and Regge, D. (2017). Usage of structured reporting in radiological practice: results from an Italian online survey. *Eur. Radiol.* 27, 1934–1943. doi: 10.1007/s00330-016-4553-6

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science* 363, 1287–1289. doi: 10.1126/science.aaw4399

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36, 193–202. doi: 10.1007/bf00344251

Gabelloni, M., Faggioni, L., and Neri, E. (2019). Imaging biomarkers in upper gastrointestinal cancers. *BJR Open* 1:20190001. doi: 10.1259/bjro.20190001

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., et al. (2019). Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *J. Am. Coll. Radiol.* 16, 1516–1521. doi: 10.1016/j.jacr.2019.07.028

Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., et al. (2019). *HARK Side Of Deep Learning – From Grad Student Descent To Automated Machine Learning*. Available online at: https://arxiv.org/abs/1904.07633v1 (Accessed July 22, 2021).

Geras, K. J., Mann, R. M., and Moy, L. (2019). Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology* 293, 246–259. doi: 10.1148/radiol.2019182627

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology* 278, 563–577. doi: 10.1148/radiol.2015151169

Glazer, G. M., and Ruiz-Wibbelsmann, J. A. (2011). The invisible radiologist. *Radiology* 258, 18–22. doi: 10.1148/radiol.10101447

Gong, B., Nugent, J. P., Guest, W., Parker, W., Chang, P. J., Khosa, F., et al. (2019). Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. *Acad. Radiol.* 26, 566–577. doi: 10.1016/j.acra.2018.10.007

Grassi, R., Miele, V., and Giovagnoni, A. (2019). Artificial intelligence: a challenge for third millennium radiologist. *Radiol. Med.* 124, 241–242. doi: 10.1007/s11547-019-00990-5

Gurgitano, M., Angileri, S. A., Rodà, G. M., Liguori, A., Pandolfi, M., Ierardi, A. M., et al. (2021). Interventional radiology ex-machina: impact of artificial intelligence on practice. *Radiol. Med.* 2021, 1–9. doi: 10.1007/s11547-021-01351-x

Harada, Y., Katsukura, S., Kawamura, R., and Shimizu, T. (2021). effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: an exploratory analysis of data from a randomized controlled study. *Int. J. Environ. Res. Public Health* 18:5562. doi: 10.3390/ijerph18115562

Harmon, S. A., Sanford, T. H., Xu, S., Turkbey, E. B., Roth, H., Xu, Z., et al. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat. Commun.* 11:4080. doi: 10.1038/s41467-020-17971-2

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Hickman, S. E., Baxter, G. C., and Gilbert, F. J. (2021). Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br. J. Cancer* 125, 15–22. doi: 10.1038/s41416-021-01333-w

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Holzinger, A. (2021). Explainable AI and multi-modal causability in medicine. *I-Com* 19, 171–179. doi: 10.1515/icom-2020-0024

Honavar, S. G. (2018). Patient–physician relationship – communication is the key. *Indian J. Ophthalmol.* 66, 1527–1528. doi: 10.4103/ijo.IJO_1760_18

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510. doi: 10.1038/s41568-018-0016-5

Hu, H. T., Shan, Q. Y., Chen, S. L., Li, B., Feng, S. T., Xu, E. J., et al. (2020). CT-based radiomics for preoperative prediction of early recurrent hepatocellular carcinoma: technical reproducibility of acquisition and scanners. *Radiol. Med.* 125, 697–705. doi: 10.1007/s11547-020-01174-2

Huxley, C. J., Atherton, H., Watkins, J. A., and Griffiths, F. (2015). Digital communication between clinician and patient and the impact on marginalised groups: a realist review in general practice. *Br. J. Gen. Pract.* 65, e813–e821. doi: 10.3399/bjgp15X687853

Ierardi, A. M., Fontana, F., Giorlando, F., De Marchi, G., Pinto, A., Radaelli, A., et al. (2016). Evaluation of tablet ultrasound for routine abdominal interventional procedures. *Radiol. Med.* 121, 675–680. doi: 10.1007/s11547-016-0641-6

Ishii, E., Ebner, D. K., Kimura, S., Agha-Mir-Salim, L., Uchimido, R., and Celi, L. A. (2020). The advent of medical artificial intelligence: lessons from the Japanese approach. *J. Intensive Care Med.* 8:35. doi: 10.1186/s40560-020-00452-5

Jha, S., and Topol, E. J. (2016). Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* 316, 2353–2354. doi: 10.1001/jama.2016.17438

Jiang, Y., Liang, X., Wang, W., Chen, C., Yuan, Q., Zhang, X., et al. (2021). Noninvasive prediction of occult peritoneal metastasis in gastric cancer using deep learning. *JAMA Netw. Open* 4:e2032269. doi: 10.1001/jamanetworkopen.2020.32269

Jones, H. (2018). *Geoff hinton dismissed the need for explainable AI: 8 experts explain why he's wrong. Forbes Magazine*. Available online at: https://www.forbes.com/sites/cognitiveworld/2018/12/20/geoff-hinton-dismissed-the-need-for-explainable-ai-8-experts-explain-why-hes-wrong/ [Accessed July 22, 2021].

Juravle, G., Boudouraki, A., Terziyska, M., and Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. *Prog. Brain. Res.* 253, 263–282. doi: 10.1016/bs.pbr.2020.06.006

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., et al. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans. Med. Imaging* 35, 1322–1331. doi: 10.1109/TMI.2016.2532122

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17:195. doi: 10.1186/s12916-019-1426-2

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131.e9. doi: 10.1016/j.cell.2018.02.010

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *J. Clin. Neurosci.* 64, 277–282. doi: 10.1016/j.jocn.2019.03.001

Kiener, M. (2020). Artificial intelligence in medicine and the disclosure of risks. *AI Soc.* doi: 10.1007/s00146-020-01085-w

King, A. D., Chow, K.-K., Yu, K.-H., Mo, F. K. F., Yeung, D. K. W., Yuan, J., et al. (2013). Head and neck squamous cell carcinoma: diagnostic performance of diffusion-weighted MR imaging for the prediction of treatment response. *Radiology* 266, 531–538. doi: 10.1148/radiol.12120167

Kobayashi, Y., Ishibashi, M., and Kobayashi, H. (2019). How will "democratization of artificial intelligence" change the future of radiologists? *Jpn. J. Radiol.* 37, 9–14. doi: 10.1007/s11604-018-0793-5

Kohli, M., and Geis, R. (2018). Ethics, artificial intelligence, and radiology. *J. Am. Coll. Radiol.* 15, 1317–1319. doi: 10.1016/j.jacr.2018.05.020

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., et al. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* 35, 303–312. doi: 10.1016/j.media.2016.07.007

Krittanawong, C. (2018). The rise of artificial intelligence and the uncertain future for physicians. *Eur. J. Intern. Med.* 48, e13–e14. doi: 10.1016/j.ejim.2017.06.017

Kulikowski, C. A. (2019). Beginnings of artificial intelligence in medicine (AIM): computational artifice assisting scientific inquiry and clinical art - with reflections on present AIM challenges. *Yearb. Med. Inform.* 28, 249–256. doi: 10.1055/s-0039-1677895

Kulkarni, A. (2021). *AI In Healthcare: Data Privacy And Ethics Concerns*. Available online at: https://www.lexalytics.com/lexablog/ai-healthcare-data-privacy-ethics-issues [Accessed July 22, 2021].

Lakhani, P., and Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 284, 574–582. doi: 10.1148/radiol.2017162326

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14, 749–762. doi: 10.1038/nrclinonc.2017.141

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lee, K., and Lee, S. H. (2020). Artificial intelligence-driven oncology clinical decision support system for multidisciplinary teams. *Sensors (Basel)* 20:4693. doi: 10.3390/s20174693

Lerouge, J., Herault, R., Chatelain, C., Jardin, F., and Modzelewski, R. (2015). IODA: an input/output deep architecture for image labeling. *Pattern Recognit.* 48, 2847–2858. doi: 10.1016/j.patcog.2015.03.017

Levinson, W., Lesser, C. S., and Epstein, R. M. (2010). Developing physician communication skills for patient-centered care. *Health Aff. (Millwood)* 29, 1310–1318. doi: 10.1377/hlthaff.2009.0450

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, Z., Wang, S., Dong, D., Wei, J., Fang, C., Zhou, X., et al. (2019). The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics* 9, 1303–1322. doi: 10.7150/thno.30309

Lo Gullo, R., Daimiel, I., Morris, E. A., and Pinker, K. (2020). Combining molecular and imaging metrics in cancer: radiogenomics. *Insights Imaging* 11, 1. doi: 10.1186/s13244-019-0795-6

Lu, J. (2016). Will medical technology deskill doctors? *Int. Educ. Stud.* 9, 130–134. doi: 10.5539/ies.v9n7p130

Lustberg, T., van Soest, J., Gooding, M., Peressutti, D., Aljabar, P., van der Stoep, J., et al. (2018). Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother. Oncol.* 126, 312–317. doi: 10.1016/j.radonc.2017.11.012

Marcovici, P. A., and Taylor, G. A. (2014). Journal Club: structured radiology reports are more complete and more effective than unstructured reports. *AJR Am. J. Roentgenol.* 203, 1265–1271. doi: 10.2214/AJR.14.12636

Martinez-Martin, N., Dunn, L. B., and Roberts, L. W. (2018). Is it ethical to use prognostic estimates from machine learning to treat psychosis? *AMA J. Ethics* 20, E804–E811. doi: 10.1001/amajethics.2018.804

Martín-Noguerol, T., Paulano-Godino, F., López-Ortega, R., Górriz, J. M., Riascos, R. F., and Luna, A. (2021). Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. *Clin. Radiol.* 76, 317–324. doi: 10.1016/j.crad.2020.11.113

Mendelson, E. B. (2019). Artificial intelligence in breast imaging: potentials and limitations. *AJR Am. J. Roentgenol.* 212, 293–299. doi: 10.2214/AJR.18.20532

Meskó, B., Drobni, Z., Bényei, É, Gergely, B., and Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *mHealth* 3:38. doi: 10.21037/mhealth.2017.08.07

Meskó, B., Hetényi, G., and Győrffy, Z. (2018). Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Serv. Res.* 18:545. doi: 10.1186/s12913-018-3359-4

Mikhaylov, S. J., Esteve, M., and Campion, A. (2018). Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration. *Philos. Trans. A Math. Phys. Eng. Sci.* 376:20170357. doi: 10.1098/rsta.2017.0357

Miller, D. D., and Brown, E. W. (2019). How cognitive machines can augment medical imaging. *AJR Am. J. Roentgenol.* 212, 9–14. doi: 10.2214/AJR.18.19914

Mittelstadt, B. D., and Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* 22, 303–341. doi: 10.1007/s11948-015-9652-2

Monreale, A. (2020). *Rischi Etico-Legali Dell'intelligenza Artificiale. DPCE Online 44*. Available online at: http://www.dpceonline.it/index.php/dpceonline/article/view/1083 [Accessed July 22, 2021].

Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., and Pedreschi, D. (2014). Privacy-by-design in big data analytics and social mining. *EPJ Data Sci.* 3, 10. doi: 10.1140/epjds/s13688-014-0010-4

Murata, K., Endo, K., Aihara, T., Suzuki, H., Sawaji, Y., Matsuoka, Y., et al. (2020). Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Sci. Rep.* 10:20031. doi: 10.1038/s41598-020-76866-w

Myers, T. G., Ramkumar, P. N., Ricciardi, B. F., Urish, K. L., Kipper, J., and Ketonis, C. (2020). Artificial intelligence and orthopaedics: an introduction for clinicians. *J. Bone Joint. Surg. Am.* 102, 830–840. doi: 10.2106/JBJS.19.01128

Na, L., Yang, C., Lo, C. C., Zhao, F., Fukuoka, Y., and Aswani, A. (2018). Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Netw. Open* 1:e186040. doi: 10.1001/jamanetworkopen.2018.6040

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., et al. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368:m689. doi: 10.1136/bmj.m689

Nance, J. W. Jr., Meenan, C., and Nagy, P. G. (2013). The future of the radiology information system. *AJR Am. J. Roentgenol.* 200, 1064–1070. doi: 10.2214/AJR.12.10326

Napel, S., Mu, W., Jardim-Perassi, B. V., Aerts, H. J. W. L., and Gillies, R. J. (2018). Quantitative imaging of cancer in the postgenomic era: radio(geno)mics, deep learning, and habitats. *Cancer* 124, 4633–4649. doi: 10.1002/cncr.31630

Nazari, M., Shiri, I., Hajianfar, G., Oveisi, N., Abdollahi, H., Deevband, M. R., et al. (2020). Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning. *Radiol Med.* 125, 754–762. doi: 10.1007/s11547-020-01169-z

Nelson, C. A., Pérez-Chada, L. M., Creadore, A., Li, S. J., Lo, K., Manjaly, P., et al. (2020). Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol.* 156, 501–512. doi: 10.1001/jamadermatol.2019.5014

Neri, E., Coppola, F., Miele, V., Bibbolino, C., and Grassi, R. (2020). Artificial intelligence: who is responsible for the diagnosis? *Radiol. Med.* 125, 517–521. doi: 10.1007/s11547-020-01135-9

Ngo, T. A., Lu, Z., and Carneiro, G. (2017). Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. *Med. Image Anal.* 35, 159–171. doi: 10.1016/j.media.2016.05.009

Nguyen, G. K., and Shetty, A. S. (2018). Artificial intelligence and machine learning: opportunities for radiologists in training. *J. Am. Coll. Radiol.* 15, 1320–1321. doi: 10.1016/j.jacr.2018.05.024

Oh, C., Lee, T., Kim, Y., Park, S., Kwon, S., and Suh, B. (2017). "Us vs. them: understanding artificial intelligence technophobia over the Google DeepMind challenge match," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems CHI '17* (New York, NY: Association for Computing Machinery), 2523–2534. doi: 10.1145/3025453.3025539

Oksanen, A., Savela, N., Latikka, R., and Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human-technology interactions online. *Front. Psychol.* 11:568256. doi: 10.3389/fpsyg.2020.568256

Ooi, S. K. G., Makmur, A., Soon, A. Y. Q., Fook-Chong, S., Liew, C., Sia, S. Y., et al. (2021). Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. *Singapore Med. J.* 62, 126–134. doi: 10.11622/smedj.2019141

Panesar, S. S., Kliot, M., Parrish, R., Fernandez-Miranda, J., Cagle, Y., and Britz, G. W. (2020). Promises and perils of artificial intelligence in neurosurgery. *Neurosurgery* 87, 33–44. doi: 10.1093/neuros/nyz471

Park, J. E., Kim, D., Kim, H. S., Park, S. Y., Kim, J. Y., Cho, S. J., et al. (2020). Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *Eur. Radiol.* 30, 523–536. doi: 10.1007/s00330-019-06360-z

Park, W. J., and Park, J.-B. (2018). History and application of artificial neural networks in dentistry. *Eur. J. Dent.* 12, 594–601. doi: 10.4103/ejd.ejd_325_18

Pesapane, F., Codari, M., and Sardanelli, F. (2018a). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* 2:35. doi: 10.1186/s41747-018-0061-6

Pesapane, F., Volonté, C., Codari, M., and Sardanelli, F. (2018b). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* 9, 745–753. doi: 10.1007/s13244-018-0645-y

Philbrick, K. A., Yoshida, K., Inoue, D., Akkus, Z., Kline, T. L., Weston, A. D., et al. (2018). What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *AJR Am. J. Roentgenol.* 211, 1184–1193. doi: 10.2214/AJR.18.20331

Pinker, K., Shitano, F., Sala, E., Do, R. K., Young, R. J., Wibmer, A. G., et al. (2018). Background, current role, and potential applications of radiogenomics. *J. Magn. Reson. Imaging* 47, 604–620. doi: 10.1002/jmri.25870

Pinto Dos Santos, D., and Baeßler, B. (2018). Big data, artificial intelligence, and structured reporting. *Eur. Radiol. Exp.* 2:42. doi: 10.1186/s41747-018-0071-4

Porsdam Mann, S., Savulescu, J., and Sahakian, B. J. (2016). Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. *Philos. Trans. A. Math. Phys. Eng. Sci.* 374:20160130. doi: 10.1098/rsta.2016.0130

Pravettoni, G., and Triberti, S. (2019). *Il Medico 4.0: Come Cambia La Relazione Medico-Paziente Nell'era Delle Nuove Tecnologie*. Palm Beach Gardens, FL: EDRA.

Price, W. N. II, Gerke, S., and Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. *JAMA* 322, 1765–1766. doi: 10.1001/jama.2019.15064

Price, W. N. II, Gerke, S., and Cohen, I. G. (2021). How much can potential jurors tell us about liability for medical artificial intelligence? *J. Nucl. Med.* 62, 15–16. doi: 10.2967/jnumed.120.257196

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., et al. (2018). Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 15:e1002686. doi: 10.1371/journal.pmed.1002686

Recht, M., and Bryan, R. N. (2017). Artificial intelligence: threat or boon to radiologists? *J. Am. Coll. Radiol.* 14, 1476–1480. doi: 10.1016/j.jacr.2017.07.007

Reddy, S., Fox, J., and Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* 112, 22–28. doi: 10.1177/0141076818815510

Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., et al. (2018). Radiomics: the facts and the challenges of image analysis. *Eur. Radiol. Exp.* 2, 36. doi: 10.1186/s41747-018-0068-z

Rogers, W., Thulasi Seetha, S., Refaee, T. A. G., Lieverse, R. I. Y., Granzier, R. W. Y., Ibrahim, A., et al. (2020). Radiomics: from qualitative to quantitative imaging. *Br. J. Radiol.* 93, 20190948. doi: 10.1259/bjr.20190948

Ross, P., and Spates, K. (2020). Considering the safety and quality of artificial intelligence in health care. *Jt. Comm. J. Qual. Patient Saf.* 46, 596–599. doi: 10.1016/j.jcjq.2020.08.002

Rubin, D. L. (2019). Artificial intelligence in imaging: the radiologist's role. *J. Am. Coll. Radiol.* 16, 1309–1317. doi: 10.1016/j.jacr.2019.05.036

Russell, S., and Bohannon, J. (2015). Artificial intelligence. Fears of an AI pioneer. *Science* 349:252. doi: 10.1126/science.349.6245.252

Sardanelli, F., Hunink, M. G., Gilbert, F. J., Di Leo, G., and Krestin, G. P. (2010). Evidence-based radiology: why and how? *Eur. Radiol.* 20, 1–15. doi: 10.1007/s00330-009-1574-4

Savadjiev, P., Chong, J., Dohan, A., Vakalopoulou, M., Reinhold, C., Paragios, N., et al. (2019). Demystification of AI-driven medical image interpretation: past, present and future. *Eur. Radiol.* 29, 1616–1624. doi: 10.1007/s00330-018-5674-x

Shan, H., Padole, A., Homayounieh, F., Kruger, U., Khera, R. D., Nitiwarangkul, C., et al. (2019). Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat. Mach. Intell.* 1, 269–276. doi: 10.1038/s42256-019-0057-9

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). "Membership inference attacks against machine learning models," in *Proceedings of the 2017*

IEEE Symposium on Security and Privacy (SP), San Jose, CA, 3–18. doi: 10.1109/SP.2017.41

Sniecinski, I., and Seghatchian, J. (2018). Artificial intelligence: a joint narrative on potential use in pediatric stem and immune cell therapies and regenerative medicine. *Transfus. Apher. Sci.* 57, 422–424. doi: 10.1016/j.transci.2018.05.004

Sogani, J., Allen, B. Jr., Dreyer, K., and McGinty, G. (2020). Artificial intelligence in radiology: the ecosystem essential to improving patient care. *Clin. Imaging* 59, A3–A6. doi: 10.1016/j.clinimag.2019.08.001

Srinuan, C., and Bohlin, E. (2011). *Understanding The Digital Divide: A Literature Survey And Ways Forward*. Available online at: https://www.econstor.eu/handle/10419/52191 [Accessed July 22, 2021].

Story, M. D., and Durante, M. (2018). Radiogenomics. *Med. Phys.* 45, e1111–e1122. doi: 10.1002/mp.13064

Stoyanova, R., Pollack, A., Takhar, M., Lynne, C., Parra, N., Lam, L. L. C., et al. (2016). Association of multiparametric MRI quantitative imaging features with prostate cancer gene expression in MRI-targeted prostate biopsies. *Oncotarget* 7, 53362–53376. doi: 10.18632/oncotarget.10523

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowledge Based Syst.* 10, 557–570. doi: 10.1142/S0218488502001648

Tajmir, S. H., and Alkasab, T. K. (2018). Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence. *Acad. Radiol.* 25, 747–750. doi: 10.1016/j.acra.2018.03.007

Tang, X. (2020). The role of artificial intelligence in medical imaging research. *BJR Open* 2:20190031. doi: 10.1259/bjro.20190031

Tobia, K., Nielsen, A., and Stremitzer, A. (2021). When does physician use of AI increase liability? *J. Nucl. Med.* 62, 17–21. doi: 10.2967/jnumed.120.256032

Tohka, J., and van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: a tutorial. *Comput. Biol Med.* 132:104324. doi: 10.1016/j.compbiomed.2021.104324

Trebeschi, S., van Griethuysen, J. J. M., Lambregts, D. M. J., Lahaye, M. J., Parmar, C., Bakers, F. C. H., et al. (2017). Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci. Rep.* 7:5301. doi: 10.1038/s41598-017-05728-9

Triberti, S., Durosini, I., and Pravettoni, G. (2020). A "third wheel" effect in health decision making involving artificial entities: a psychological perspective. *Front. Public Health* 8:117. doi: 10.3389/fpubh.2020.00117

Trimboli, R. M., Codari, M., Bert, A., Carbonaro, L. A., Maccagnoni, S., Raciti, D., et al. (2018). Breast arterial calcifications on mammography: intra- and inter-observer reproducibility of a semi-automatic quantification tool. *Radiol. Med.* 123, 168–173. doi: 10.1007/s11547-017-0827-6

Valiuškaitė, V., Raudonis, V., Maskeliūnas, R., Damaševičius, R., and Krilavičius, T. (2020). Deep learning based evaluation of spermatozoid motility for artificial Insemination. *Sensors (Basel)* 21:72. doi: 10.3390/s21010072

van Assen, M., Muscogiuri, G., Caruso, D., Lee, S. J., Laghi, A., and De Cecco, C. N. (2020). Artificial intelligence in cardiac radiology. *Radiol. Med.* 125, 1186–1199. doi: 10.1007/s11547-020-01277-w

van Hoek, J., Huber, A., Leichtle, A., Härmä, K., Hilt, D., von Tengg-Kobligk, H., et al. (2019). A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *Eur. J. Radiol.* 121:108742. doi: 10.1016/j.ejrad.2019.108742

Villanueva-Meyer, J. E., Chang, P., Lupo, J. M., Hess, C. P., Flanders, A. E., and Kohli, M. (2019). Machine learning in neurooncology imaging: from study request to diagnosis and treatment. *AJR Am. J. Roentgenol.* 212, 52–56. doi: 10.2214/AJR.18.20328

Voigt, P., and Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st Edn. Cham: Springer.

Vought, R. T. (2020). *Re: Guidance For Regulation Of Artificial Intelligence Applications*. Available online at: https://www.acr.org/-/media/ACR/Files/Advocacy/Regulatory-Issues/acr_comments_draft-OMB-memo_3-12-2020.pdf [Accessed July 22, 2021].

Wang, S., Liu, Z., Rong, Y., Zhou, B., Bai, Y., Wei, W., et al. (2019). Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother. Oncol.* 132, 171–177. doi: 10.1016/j.radonc.2018.10.019

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., et al. (2017). Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. *Med. Image Anal.* 40, 172–183. doi: 10.1016/j.media.2017.06.014

Waymel, Q., Badr, S., Demondion, X., Cotten, A., and Jacques, T. (2019). Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagn. Interv. Imaging* 100, 327–336. doi: 10.1016/j.diii.2019.03.015

Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nat. Biomed. Eng.* 2, 719–731. doi: 10.1038/s41551-018-0305-z

Zackova, E. (2015). "Intelligence explosion quest for humankind," in *Beyond artificial intelligence: the disappearing human-machine divide*, eds J. Romportl, E. Zackova, and J. Kelemen (Cham: Springer International Publishing), 31–43. doi: 10.1007/978-3-319-09668-1_3

Zanotel, M., Bednarova, I., Londero, V., Linda, A., Lorenzon, M., Girometti, R., et al. (2018). Automated breast ultrasound: basic principles and emerging clinical applications. *Radiol. Med.* 123, 1–12. doi: 10.1007/s11547-017-0805-z

Zerunian, M., Caruso, D., Zucchelli, A., Polici, M., Capalbo, C., Filetti, M., et al. (2021). CT based radiomic approach on first line pembrolizumab in lung cancer. *Sci. Rep.* 11:6633. doi: 10.1038/s41598-021-86113-5

Zhou, X., Ma, Y., Zhang, Q., Mohammed, M. A., and Damaševičius, R. (2021). A reversible watermarking system for medical color images: balancing capacity, imperceptibility, and robustness. *Electronics* 10:1024. doi: 10.3390/electronics10091024

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature* 555, 487–492. doi: 10.1038/nature25988

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 Psychology

REVIEW
published: 30 September 2021
doi: 10.3389/fpsyg.2021.696346

On the Commoditization of Artificial Intelligence

Abdullah A. Abonamah, Muhammad Usman Tariq and Samar Shilbayeh*

Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates

As artificial intelligence's potential and pervasiveness continue to increase, its strategic importance, effects, and management must be closely examined. Societies, governments, and business organizations need to view artificial intelligence (AI) technologies and their usage from an entirely different perspective. AI is poised to have a tremendous impact on every aspect of our lives. Therefore, it must have a broader view that transcends AI's technical capabilities and perceived value, including areas of AI's impact and influence. Nicholas G. Carr's seminal paper "IT Does not Matter (Carr, 2003) explained how IT's potential and ubiquity have increased, but IT's strategic importance has declined with time. AI is poised to meet the same fate as IT. In fact, the commoditization of AI has already begun. This paper presents the arguments to demonstrate that AI is moving rapidly in this direction. It also proposes an artificial intelligence-based organizational framework to gain value-added elements for lowering the impact of AI commoditization.

Keywords: artificial intelligence, AI commoditization, AI business value, AI strategy, AI operations

OPEN ACCESS

Edited by:
Davide La Torre, SKEMA Business School, Sophia Antipolis Campus, France

Reviewed by:
Filipe Portela, University of Minho, Portugal
Jane Heather, Eastern New Mexico University, United States

****Correspondence:***
Muhammad Usman Tariq
m.tariq@adsm.ac.ae

Specialty section:
This article was submitted to Organizational Psychology, a section of the journal Frontiers in Psychology

Received: *16 April 2021*
Accepted: *16 August 2021*
Published: *30 September 2021*

Citation:
Abonamah AA, Tariq MU and Shilbayeh S (2021) On the Commoditization of Artificial Intelligence. Front. Psychol. 12:696346. doi: 10.3389/fpsyg.2021.696346

INTRODUCTION

Artificial intelligence is a technological concept in operational management, philosophy, humanities, statistics, mathematics, computer sciences, and social sciences. Artificial intelligence aims to create computers or machines to carry out jobs that generally need human intelligence. The sub-discipline of artificial intelligence is machine learning, which then directs to statistical learning. Artificial intelligence is a branch of computer science that allows the machine to mimic human intelligence and execute tasks that humans can perform more efficiently. This term may also be applied to any machine that exhibits traits associated with a human mind, such as learning and problem-solving. This section is about how AI has developed over the years and about many important discoveries and inventions that have brought us to where AI is today.

In 1947, Alan Turing raised a life-changing yet straightforward query, "Can machines think?" Turing gave quite possibly the earliest public lecture (London, 1947) to mention computer intelligence, saying, "What we want is a machine that can learn from experience," and that the "possibility of letting the machine alter its instructions provides the mechanism for this." (Britannica, 2021). Decades later, scientists demonstrated that computers could, in fact, exhibit some form of intelligence. In the midst of the 1950s and 1970s, the computer industry discovered its foothold when computers began operating faster, became more approachable, and were less costly. An article in 1970 Life Magazine predicted that machines would shortly have a similar intellect to human beings in only 3–5 years. Still, extensive evolution in storage adequacy and computing capacity was required for that to take place. During the 1980s, there was an evolution of two effective techniques. The first one was the "expert system," which imitated human's aptitude

to make decisions. Computers started to utilize reasoning depended on "rules" - an "if-then/else" procedure used to respond to queries. The second was "machine learning," which made computers learn through experience. In 1997, Dragon Systems made and adopted software for natural language speech recognition on Windows. In the 2000s, speed and storage options such as "cloud, catapulting the utilization of computers into the mainstream," became more widespread providing artificial intelligence with its turn in the spotlight. Recently, advancements in artificial intelligence technologies have accelerated at a very rapid pace. The acceleration may be explained by three significant industry developments (Hildt, 2019).

- **"Graphics Processing Units (GPU)":** Requirement propelled by the gaming and video world produced enhanced and less costly GPUs. In addition, these GPUs have increased the processing power of AI algorithms.
- **"Big Data and Machine Learning":** Artificial intelligence uses machine learning algorithms to perform data analytics by building learning models to be used in "intelligent" ways. The learning models are based on the idea that machines can learn from data, identify patterns, and predict future states that help in decision-making with little human intervention.
- **"Deep learning":** A subset of machine learning that can learn from the data without human intervention. For example, in classical machine learning (non-deep), machines need human interventions to label the unlabeled data. On the other hand, "deep" machine learning can leverage labeled datasets to inform its algorithm, but it does not necessarily require a labeled dataset to enable the unsupervised machine to train without any interventions.

It is widely acknowledged that AI can perform human-similar tasks with some level of intelligence, such as understanding verbal communication, driving cars, and distinguishing pictures. However, in comparison with human intelligence and understanding AI, the following section will shed light on three AI types.

Types of Artificial Intelligence: ASI, AGI, ANI

There are three types of artificial intelligence: ASI (artificial super intelligence), AGI (artificial general intelligence), and ANI (artificial narrow intelligence).

- **Artificial Superintelligence:** This is a hypothetical ability of an intelligent agent to possess intelligence substantially exceeding that of the brightest and most gifted human minds. Currently, it is not technologically possible to produce machines that possess superintelligence properties.
- **Artificial General Intelligence (AGI):** This is the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can perform. Currently, it is a major focus of much artificial intelligence research. However, there is no existing intelligent agent that possesses the AGI properties (ref).
- **Artificial Narrow Intelligence (ANI):** ANI, also known as "weak" AI, is the most common today. Narrow AI can perform a single task—whether it is driving a car, playing chess, or recognizing spoken or written words. ANI systems are designed to focus on their tasks in real-time. With continuous learning from their environment, they build knowledge over time and become experts in performing their assigned tasks. However, these systems cannot perform tasks outside the single-task environment that they are designed for.

Artificial narrow intelligence is the most coherent kind of artificial intelligence to be utilized by most people. The following are some common examples of artificial narrow intelligence:

- **"Self-driving cars":** A self-driving car, also known as an autonomous vehicle, driverless car, or robo-car, is a vehicle that is capable of sensing its environment and moving safely with little or no human input. Self-driving cars combine a variety of sensors to perceive their environment. These sensors include radar, lidar, sonar, GPS, odometry, and inertial measurement units. Advanced control systems interpret the sensors' data to identify appropriate navigation paths and obstacles and relevant signage.
- **"Voice assistant devices":** A voice assistant is a digital assistant that uses voice recognition and natural language processing to listen and respond to verbal commands. Voice assistant devices are easy to use using voice-activated commands. From playing music to scheduling appointments, voice assistant devices make daily tasks easier. Some of these devices enable you to monitor your house on your smartphone, turn the lights on with a simple command, and access all your smart devices using just your voice. Alexa, Siri, and Google Assistant are the most common examples of these voice assistant devices. Voice-powered devices rely on artificial intelligence technologies to perform their voice recognition functions.
- **Robotics:** An AI application that can perform some tasks that need some level of intelligence, such as sensing obstacles and changing the path accordingly. It can be used for carrying goods in factories, cleaning offices, and inventory management.

Artificial general intelligence aims to advance artificial intelligence one step ahead, where machines can perform tasks at the human intelligence level. To achieve that goal, artificial general intelligence automata must pass a series of tests. It begins with the Turing Test, which Alan Turing originally designed in 1950. The Turing Test is a test of a machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from a human being. If a machine gets a 70% or higher score, it is considered an artificial intelligence agent. The second test of artificial general intelligence's efficacy is done through the Coffee Test. This test asks the intelligent agent to get into a home environment, prepare coffee, and master the art of brewing it. Next, the College Robot Test requires the AGI robot to enroll in college and successfully pass all classes (Goertzel, 2017). Finally, the robot can appear in an Employment Test, where it has to clear a vocational test, including writing and driving exams (Keyes et al., 2021).

As mentioned above, AI capabilities are achievable by embedding some human intelligence. The following section will highlight different human-like intelligence forms and concentrates on learning as one of the most notable.

Artificial Intelligence Learning Approaches

Artificial intelligence, defined as the machine's ability to mimic the human brain by performing tasks, needs some intelligence (Abbass, 2019). It includes learning, reasoning, problem-solving, and perception. Learning is the machine's ability to conclude or memorize knowledge without this information being fed into it. Human learning is distinguished into different forms. The simplest one is trial and error. However, the human brain can learn in a more complicated way. Similarly, machines are designed to learn using different learning approaches. These are machine learning (ML), deep learning (DL), and reinforced learning (RL).

- Machine learning is a subset of AI that involves techniques that enable machines to learn from the given data for pattern detection and future prediction (Agrawal et al., 2019).
- Deep learning is a subset of machine learning that makes a machine observe patterns and classify information, letting it "think" in a more advanced complicated way without the need for any human interventions.
- Reinforcement learning is similar to deep learning except that, in this case, machines learn through trial and error using data from their own experience.

Below is a close look at some artificial intelligence technologies and how they perform organizations' tasks.

- A "machine learning platform" can utilize information from various data sources- such as training and development tools, together with other algorithms to forecast and sort information (Bauguess, 2017).
- Deep learning is a machine learning technique that utilizes pattern classification and recognition to function with large data sets.
- Neural networks are machine learning techniques, which use statistical algorithms designed according to neuron behavior in the human brain.
- Cognitive computing is a kind of computing that utilizes high-grade understanding and reasoning. It is not contemplated as machine learning as it uses various artificial intelligence technologies to extract outcomes.
- Computer vision allows computer systems to function and act as a human eye. It examines the condition of digital pictures and videos to generate symbolic and numeric information for decision-making procedures.
- Natural language generation involves generating text from numeric characters. Organizations mainly utilize this procedure for reports, customer service, and business intelligence summaries.
- Graphical processing units (GPUs) are a division of an electronic circuit that amplifies picture formation on a display device. GPUs are essential for artificial intelligence to function successfully.
- Internet of Things (IoT) is a network of inter-connected devices that produce and share data, such as medical devices, smart speakers, appliances, and wearable technology. Artificial intelligence relies on these devices' data to make significant business intelligence decisions.
- Advanced algorithms are complicated algorithms that are continuously being generated and combined to furnish present intelligent processing.
- An application programming interface (API) is a technology that organizations utilize to acquire artificial intelligence services. Similarly, artificial intelligence uses data flows to support firms to make sense of data to help in organizational measures (Lawless et al., 2019).

Today, AI has become a mature technology and an increasingly important part of the modern fabric of life. AI is already deployed in different application domains.

The paper is organized as follows: In section Literature Analysis, the summary of the AI historical background emphasizing AI advantages and applications is provided. Section AI Commoditization discusses AI commoditization and presents some simple recommendations on "how to utilize AI to attain competitive advantages." Section Are we ready for AI? answers the question: "Are we ready for AI?" and in section Summary of Findings and Conclusion, conclusions are provided.

LITERATURE ANALYSIS

In recent years, there has been an abundance of research articles published in the area of AI. In the following section, we briefly present how AI has evolved during recent years:

The Evolution of Artificial Intelligence

Artificial intelligence has been shown to be useful in fulfilling the following requirements (Davenport and Ronanki, 2018). A paper published in 1955 referred to a famous economist who wrote in 1828 regarding the probability of motor cars as replacements for horses: "Nevertheless no machine will ever be able to perform what even the worst horses can—the service of carrying people and goods through the bustle and throng of a great city." People could never have imagined self-driving cars, intelligent mobiles, video calls, intelligent robots, pilotless airplanes, and supercomputers. Nevertheless, artificial intelligence that would have been considered science fiction <190 years ago are now available in today's era, and some, like self-driving motor cars, will most probably be in extensive use within the next 5 years (Katz, 2017). The challenge is to attempt to forecast future technologies based on artificial intelligence without repeating the errors of similar myopic scholars, who could not understand the significant computational evolution of the latest technologies (Dignum, 2019). There are two perceptions to be made. First, 190 years is a short period by ancient standards, and in the period, the world went from horses that were the most significant source of transportation to self-driving motor cars and from slide rules and abacuses to intelligent devices in our pockets (Roff, 2019). Secondly, the time frame between technological evolution and practical, general use is continuously being decreased. For

example, there were more than 200 years from when Newcomen initiated the first "workable steam engine" in 1707 to when Henry Ford manufactured a dependable and cost-efficient motor car in 1908. It took more than 90 years between the time electricity was initiated and its general use by companies to enhance factory productivity. There were 20 years, non-etheless, between, ENIAC, the first computer ever, and the 360 system of IBM that was mass-produced and was budget-friendly for small business organizations. While it took 10 years from 1973 when Dr. Martin Cooper made the first mobile call through a handheld device and its public inauguration by Motorola.

The grand and most swift development started with smartphones when they first emerged in 2002. Smartphones have witnessed tremendous progress, with the latest versions consisting of significant enhancements every year by Samsung, various Chinese companies, and Apple (Villaronga et al., 2018). Smartphones, in addition to their technical features, now integrate the characteristics of artificial intelligence. These include speech recognition, furnishing personalized information in spoken language, finishing words during text typing, and various other features that need embedded artificial intelligence, offered by a pocket computer considerably smaller than a packet of cigarettes. The development has gone from intelligent computers to intelligent machines and toward programs based on artificial intelligence. A thermoregulator is a simple mechanical device that exhibits some primary but valuable intelligence that makes temperature constant at some preferred predetermined level (Haibe-Kains et al., 2020).

From digital computers to artificial intelligence tools: The Intel Pentium Microprocessor, launched in 1993, integrated music capabilities and graphics and unlatched computers to many cost-effective applications expanding more than data processing (Dunjko and Briegel, 2018). These technologies signal the start of a new phase that now involves smart personal assistants recognizing and responding to natural languages, robots capable of seeing and performing an array of smart operations, self-driving motor cars, and a range of other abilities close to that of human capability. The technology optimists determine that in <26 years, computers will have shifted from calculating 0 and 1 digits to using advanced neural network algorithms that allow the speaking of natural languages, vision, and understanding.

Tech optimists believe there is no doubt that in the next 21 years, augmented artificial intelligence technological advancement will lead to a leap in deep learning that emulates the way youngsters learn, instead of arduous guidance by custom-made programs directed for particular applications that are dependent on logic, decision trees, and if-then logic (Galbusera et al., 2019). For example, DeepMind depends on a neural program using deep learning that comprehends by itself how to play various Atari games, like Breakout, or better than humans, without detailed guidance for doing it, but by playing a dozen games and revamping itself every time. The program distinctly instructed AlphaGo that beat Go champion Lee Sodol in 2016 (Luckin, 2017). In addition, however, it will develop a new project base to understand how to play Starcraft, a complex game dependent on long-term plans and robust skillful decisions to stay ahead of the rival, which DeepMind plans to be its next target for progressing deep learning. Deep learning is a concept that seems to be the leading edge of research and funding attempts to enhance artificial intelligence, as its success has created a burst of activity in capital funding that gave more than $1.5 Billion to 125 projects for start-ups in the first quarter of 2019, in comparison to 31 projects in a comparable quarter of 2017 (Hall and Pesenti, 2017).

Google had five deep learning projects underway in 2019. Today it is continuing more than 4,000, according to their representative, in all its significant sectors, involving Gmail, self-driving cars, Android, YouTube, translation, and maps. IBM's Watson system utilized artificial intelligence, but not deep learning, when it defeated the two jeopardy champions in 2011. However, there has been a boost in all of Watson's 35 constituent services due to deep learning (Semmler and Rose, 2017). Shareholders who were not aware of deep learning 7 years ago today are considerate of start-ups that do not integrate AI into their programs (Cabitza et al., 2020).

For survival, it is necessary to develop advanced software applications to keep away from menus by integrating natural-language processing and clicking deep learning. How distant can deep learning go? According to tech optimists, there are no restraints for three causes (Leslie, 2019). The first one is, as development is accessible to everyone realistically to use through Open Source software, researchers will focus their attempts on new, stronger algorithms resulting in additive learning. The second one is that deep learning algorithms will be proficient in recollecting what they have learned and implementing it in the same way but in distinct situations (Sejnowski, 2020). Last and equivalently vital, in the future, intelligent computer systems will have the capability to develop new software by themselves, at first maybe not so advanced ones, but enhancing with time as learning will be integrated as a segment of their capabilities. Non-biological intelligence is expected to match the extent and refinement of human intelligence in almost a quarter of a century. This event, known as the "singularity," is estimated to happen by 2045; it will bring the emergence of a new community to outmatch biological limits and boost our creativity (Corea, 2017). There will be no difference between machine and human, virtual reality, and human reality in this new era. For some individuals, this forecast is astonishing, with wide-ranging inferences should they become a reality (Galanos, 2018).

Real-World Artificial Intelligence Applications

The possibilities of artificial intelligence are more abundant than people realize. For example, people in real life interact with cyber assistants on their favorite shopping websites, request Facebook to create an advertisement for the promotion of a business, direct Alexa to play their favorite songs, ask Google about the direction of some desired place (Braga and Logan, 2017). These are methods of using artificial intelligence to make life easier and furnish more to consumers. The following are some more:

- Amazon furnishes transactional artificial intelligence with algorithms that continuously become more progressed.

Presently, it can forecast people's buying habits and give information about the products.
- Pandora's musical DNA procedure utilizes more than 500 musical characteristics from songs that experienced musicians have humanly searched to suggest the latest songs to users according to their choices.
- "Nest" is a thermostat that is "voice-controlled by Alexa" according to preferred heating or cooling temperature (Gabriel, 2020).

Artificial Intelligence Advantages

Artificial intelligence attaches value to computer systems' existent capabilities by continuously and accurately furnishing digital information and tasks (Hagras, 2018). Artificial intelligence can support improvement through:

- Less human error and fewer mistakes
- Enhanced business decisions with an approach to actual-time data
- Automatic tasks and procedures
- Enhanced operational efficiency and productivity
- Quality-lead generation
- Enhanced data learning with improved access to large data pools
- Enhanced service with consumer knowledge (Wirtz et al., 2019)

Moving into insights, the following are five elaborated ways artificial intelligence can work for a firm:

- **Data analysis and collection:** Artificial intelligence makes data analysis and collection budget-friendly, instinctive, and well-timed so the firm can automatically comprehend more about their consumers, safe repeats, and new establishments (Parson et al., 2019).
- **Smart hiring:** Machine learning algorithms can discover best practices for an organization's particular hiring requirements and generate a list of short-listed and best candidates (Wang, 2019).
- **Back-Office Efficiency:** Artificial intelligence can manage tasks such as scheduling, accounting, and some other day-to-day operations very quickly, without any mistakes (Došilović et al., 2018).
- **Customer Service:** Virtual consumer service representatives function 24/7 and can furnish help to existing and prospective customers without any human guidance or control (Siau and Wang, 2018).
- **Targeted Marketing:** Classifying and arranging all accessible data about the service or product is an artificial intelligence distinction- allowing the organization to focus on marketing that particularly highlights customers' requirements (Došilović et al., 2018).

AI COMMODITIZATION

In his HBR article, Nicholas Carr argues that IT has become a commodity just like electricity, the telephone, the steam engine, the telegraph, and the railroad (Carr, 2003). Because of their very nature, commodities do not provide any strategic differentiation. Carr (2003) suggests that IT can be used to supplement and improve strategy implementation, but it is not the foundation of competitive advantage. The research poses the question: "*Does the commoditization of IT apply to artificial intelligence?*" The argument is that there is emerging evidence that the answer is "yes." The argument is given and based on that AI commodity model is shown in **Figure 1** below.

Argument 1- AI Scarcity Is Vanishing

Many firms have created C-level positions for technology managers. Some organizations have appointed strategy consulting companies to identify how to take advantage of their artificial intelligence investments for strategic differentiation (Wilner, 2018). The simple supposition, the increased strategic value of AI must match its degree of effectiveness and pervasiveness. It is both a logical as well as an intuitive supposition. However, the supposition is incorrect. According to economic theory, what makes a resource strategic-what provides it with the capability to be the foundation for a continuous competitive edge- is not effectiveness and pervasiveness but scarcity. Machine learning, deep learning, and reinforced learning that form the basic building blocks of AI have become readily available commodities. Currently, AI solutions are privately owned for the most part. However, as it becomes increasingly available, accessible, and affordable, the competitive edge of being "AI-enabled" starts to dissolve. Essentially, if everyone is equipped with the same capabilities, then there is no competitive advantage anymore.

Argument 2- AI On-Demand Availability

Many big cloud providers are offering AI services out of the box. Instead of requiring dedicated teams of data scientists, companies can purchase and consume AI on-demand as a service. These services avoid the complexity of building ML models and pre-requisite knowledge for domains such as speech recognition, text analytics, or image recognition, among others. Furthermore, as fully-managed services, these AI capabilities require no DevOps on the customer's part. Without oversimplifying, for speech-to-text, for instance, you need to upload audio files and click "transcribe." Then retrieve the text output for whatever downstream analytics you would like to do on that text.

In the future, AI features will be built-in in all applications. In other words, we will see a convergence toward AI parity and performance, much like we see that email platforms are more-or-less the same, even if they are made by different companies (e.g., Hotmail, Gmail, Yahoo, AOL, etc.). In this way, AI will become standard and ordinary. But, ultimately, AI capability becomes a commodity. When a resource becomes vital to the competition but insignificant to strategy, the risks it generates become more significant than the benefits it delivers (Vähäkainu and Lehto, 2019). This implies that companies' differentiation (sustainable value) must be derived from something else—something proprietary, something not readily available to others.

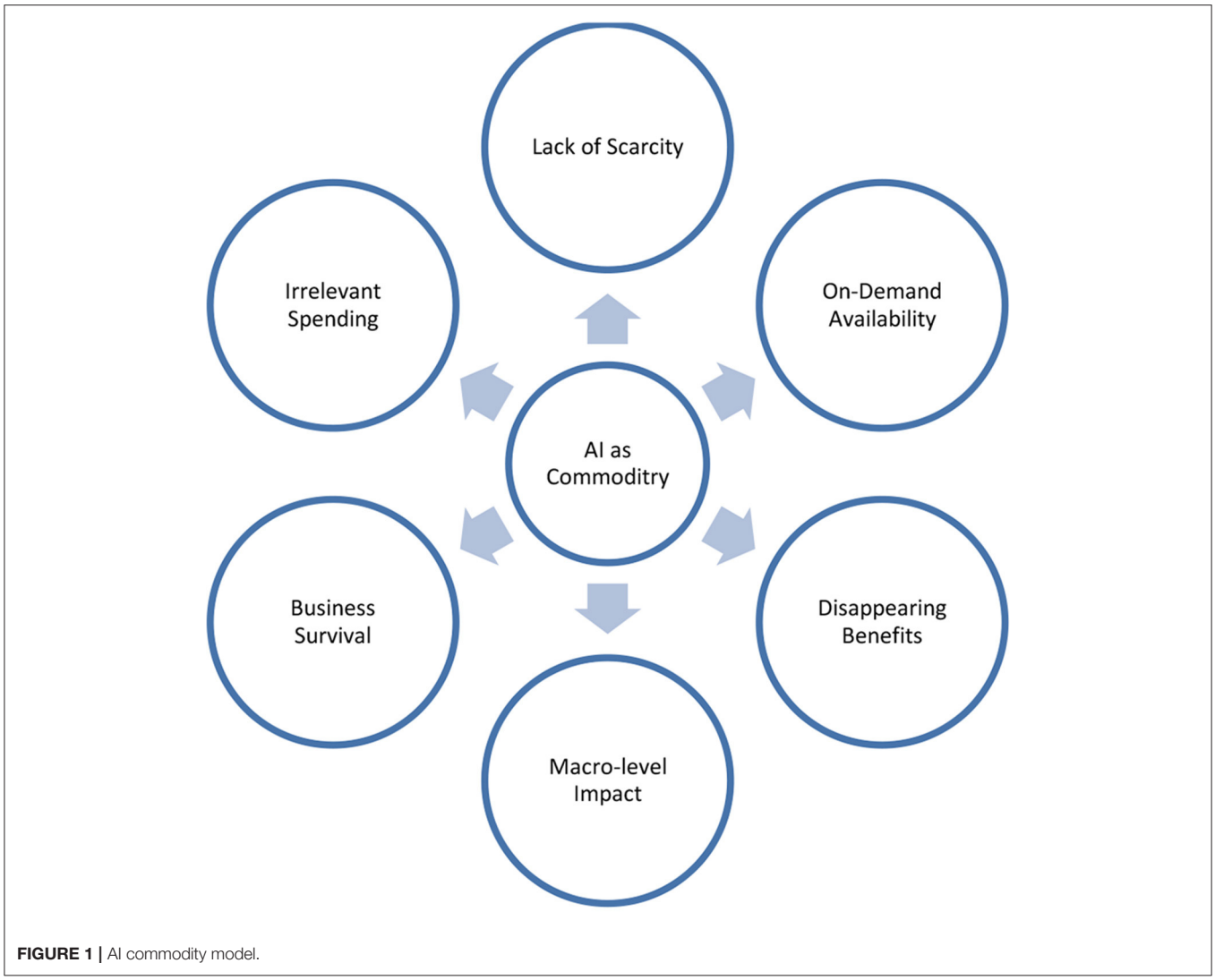

FIGURE 1 | AI commodity model.

Argument 3- AI Disappearing Benefits

Many scholars have emphasized the importance of the integration of artificial intelligence, especially machine learning, in existing technologies. However, many comparisons have emphasized both economies and investment formats linked with the technologies-the "boom to bust cycle or the roles of artificial intelligence in re-structuring the overall operation of firms." There is a significantly more minor discussion about the influence or no influence and competition at the company level for artificial intelligence.

Argument 4- AI Macro-level Impact

Artificial intelligence appears to have had significant market changes. With the advancement and evolution of artificial intelligence, the ways of living and working are also transforming, but it also raises an inevitable question about the impact of artificial intelligence on the economy, customers, and businesses. Employees want to get information about what changes artificial intelligence can bring to their income and job. In contrast, companies are concerned about how they can gain benefits from the opportunities that artificial intelligence offer and which areas need more investment. After all these considerations, the main question is how to create artificial intelligence so that it is transparent and responsible enough to gain the trust of customers and business stakeholders. Even without the elevation in labor demand due to economic factors, artificial intelligence will need new roles and jobs. In addition, with the jobs in the application and development of artificial intelligence, the technologies will need to be created, operated, maintained, and regulated.

Argument 5- AI Is Becoming Vital to Business Survival

When a system becomes vital to the competition but insignificant to strategy, the risks it generates become more than the benefits it furnishes. Consider what happened since the introduction of computers. No firm develops its strategy around its computer usage in today's world, but even a shortage of systems can cause devastation. The operational risks have created an association

with artificial intelligence- technical defects, discontinuance, service interruption, undependable partners or dealers, security vulnerability, terrorism- and some have become amplified as firms have shifted from strictly controlled artificial intelligence systems to shared, open ones. In today's world, a disturbance in an information technology system can completely paralyze the company's ability to make and deliver its products; the disruption in artificial intelligence systems can negatively affect customers' connection and build a negative reputation (Kaplan and Haenlein, 2020). Still, some firms have performed an efficient job of recognizing and moderating their vulnerabilities. Panicking about what might go in the wrong direction may not be as alluring a job as hypothesizing about the future but is the more important job at present. In the distant future, even though the most considerable artificial intelligence risk facing most firms is more unimaginative than devastation. It is clearly, excessive spending (Huang and Rust, 2018). Artificial intelligence may be a product, and its costs may decrease swiftly enough to guarantee that any new skills are quickly shared, but the very real fact that so many firms are interlinked with so many firm operations means that it will commence using a massive amount of collective spending. For many firms, just remaining in the market will result in huge artificial intelligence expenditure (Kim, 2020). What is significant- and this remains true for any product output-is the ability to separate vital investments from those that are permissive, avoidable, and even inefficacious. At an upper level, more robust cost management needs more diligence in assessing anticipated returns for the investments of systems, more innovation in investigating cheaper and simpler replacements, and a broader directness to externalization and other partnerships. Nevertheless, many firms can also derive vital savings by clearly eliminating waste. Personal computer systems are an excellent example (Cockburn et al., 2018).

Argument 6- AI Spending Is Irrelevant

Each year, firms buy more than 115 million personal computers, most of which substitute previous models. The vast majority of employees who utilize personal computers depend on only some simple applications- web-browsing, email, word processing, and spreadsheets. These applications have been technologically intelligent for years, and they need only some fragment of the computing power furnished by the microprocessors in today's world. Nonetheless, firms spend substantial amounts across the board on software and hardware updates (Townsend and Hunt, 2019). Much of that investment is compelled by the strategies of vendors. Huge software and hardware distributors have become very efficient at delivering the latest artificial intelligence capabilities and features in ways that force firms into purchasing the latest computers, networking equipment, and applications much more regularly than required. The time has arrived for artificial intelligence system buyers to endorse and confer contracts that ensure their computer investments' durable effectiveness and force rigid restrictions on reforming costs. If vendors resist, firms should be ready to find cheaper solutions involving open-source apps and essential elements that network personal computers, even if they give up on features. If a firm requires proof of the type of money that could be saved, it requires only a check at Microsoft's profit margin (Vähäkainu and Lehto, 2019).

Additionally, to be compliant in their buying, the firms have not efficiently utilized artificial intelligence. That is precisely the case with Big Data, accounting for more than half of many firms' artificial intelligence expenditures (Duan et al., 2019). A large quantity of the stored data on business networks has significantly less to do with production or serving consumers- it involves emails and files, including video clips and MP3s. Artificial intelligence's world assesses that as much as 80% of a regular Windows network is famished, which is an unnecessary expense for firms (Carter and Nielsen, 2017). Limiting staff ability to save files randomly and continually may look objectionable to various managers, but it can create an actual effect on the bottom line. Now that artificial intelligence has become the principal capital expense for most firms, there is no reason for waste and negligence (Floridi, 2020). Given the swift momentum of technological advancement, delaying artificial intelligence investments can be another solid way to dimmish costs, decreasing the company's opportunity to be burdened with problematic or soon-to-be-archaic technology. Many firms, specifically during the 2000s, boosted their artificial intelligence investments because they desired to gain an advantage at first or because they had a fear of being left behind. Except in some scenarios, both the desire and fear were indefensible (Greene et al., 2019). Some firms may be distressed that being niggardly with AI dollars will deface their competitive positions in the market. However, various studies show that businesses that integrate AI spending continuously have more significant expenditures which infrequently convert into higher financial outcomes. The contradiction is true. The most immoderate payers seldom bring the best results. Many firms spend too much on AI but get significantly less in response (Russell, 2017).

Based on the above arguments, a commoditization model is devised that shows the leading factors of commoditization.

To manage the commoditization issue, an artificial intelligence-based organizational framework is proposed that can help to add value for organizations facing issues due to commoditization of artificial intelligence as shown in **Figure 2**.

AI Organizational Framework

The following provides the overview of each phase of the artificial intelligence organization framework and its sub-components.

Macro Level Impact

Leadership and Coaching

When conjoined with an actual life-experienced teacher or coach, two vital types of artificial intelligence are now being adopted to assist learners to practice actual-world skills in their work as the latest form of performance support such as deep learning and expert systems.

Rewards and Recognition

By any means, recognition and reward program owners should be responsible for applying and understanding advanced analytics and artificial intelligence to the sector of human motivation, performance, and engagement.

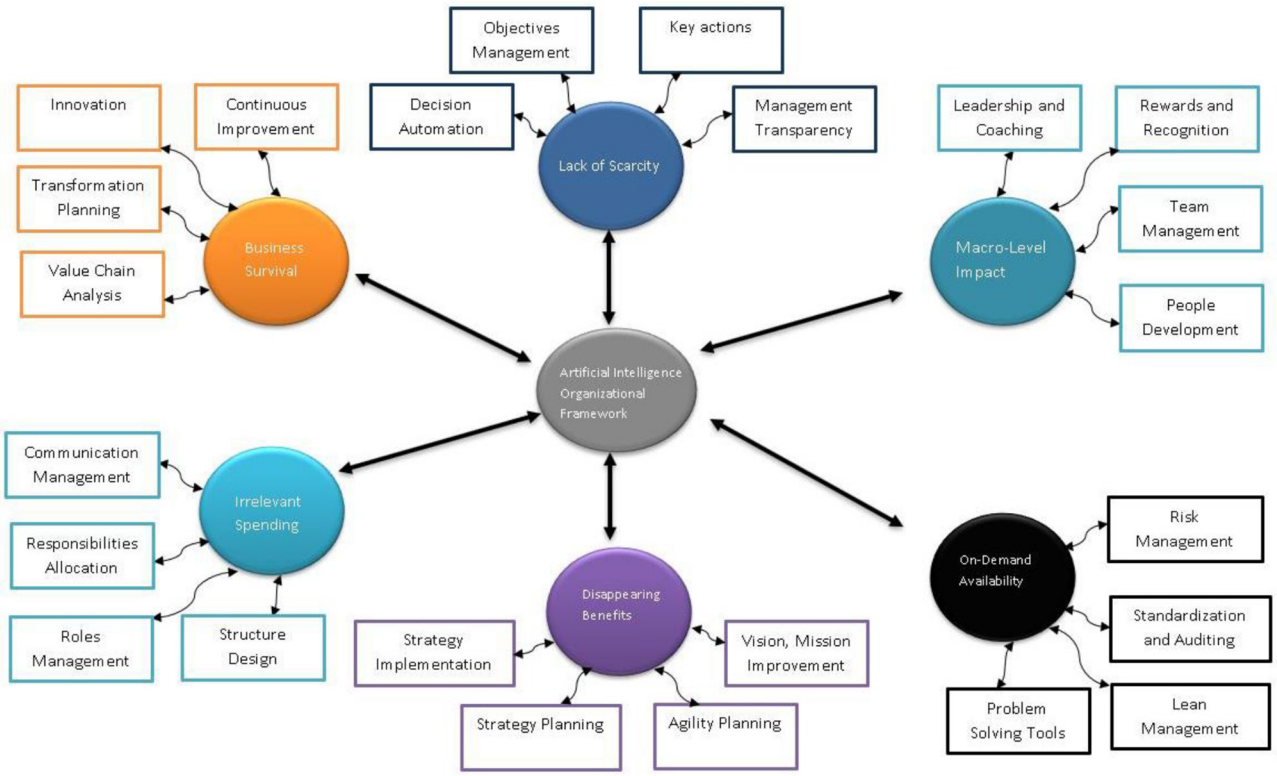

FIGURE 2 | Artificial intelligence organizational framework.

Team Management

With the adoption of chatbots and natural language processing, artificial intelligence can recognize who requires positive feedback and let the employer know or give the feedback. It can also identify who requires more training and make a fact-filled meeting schedule for the managers and employees.

People Development

Eventually, implementing artificial intelligence in training, development, and learning will let people gain content-based training according to their skills, traits, and preferences. First, however, artificial intelligence needs to create programs attainable to people even with various kinds of disabilities.

On-Demand Availability

Risk Management

Artificial intelligence can benefit risk management in different areas. For example, artificial intelligence can permit risk managers to respond quickly to the latest and evolving exposures, from the capability to process massive amounts of data to automate some repetitive and arduous risk management stages.

Standardization and Auditing

Auditing and standardization can be transformed with machine learning, heading to enhanced accuracy and productivity. With the use of machine learning, auditing firms can automate massive volumes of data to recognize incongruities and risky transactions that humans can analyze later.

Lean Management

Enhancing efficiencies and eliminating waste is a principle of artificial intelligence and also of lean management. Artificial intelligence and lean management can be revolutionary as firm leaders integrate employee experience in creating novel roles and technological structures.

Problem-Solving Tools

Artificial intelligence can solve problems by performing differential equations, logical algorithms, and using polynomial equations, and completing them utilizing modeling paradigms. As a result, there can be different solutions to one problem, which various heuristics attained (Iphofen and Kritikos, 2019).

Disappearing Benefits

Vision, Mission Improvement

Artificial intelligence is an emerging technology and may face unpredicted challenges; many firms want to take the risk and adopt artificial intelligence technology. Some of the main objectives are to lessen the operational costs, improve customer experience, and upsurge revenue (Ramamoorthy and Yampolskiy, 2018).

Agility Planning

The present era's competitive environment is in a condition of continuous acceleration, and it is dubious about slowing down ever. The breadth of technology, pace, and scale modification in previous years has developed such productive ground for innovation that it has essentially modified the ways firms succeed. As a result, the firms winning in today's world are the most agile, forecasting and responding to the change faster (Petrović, 2018).

Strategy Planning

Artificial intelligence has a significant impact on firms and plays an essential role in management. Artificial intelligence modifies the ways of business management and strategic planning. It helps the firms attain a competitive edge in the market and helps them achieve great success.

Strategy Implementation

Any artificial intelligence strategy will work as an unceasingly evolving parameter to confirm the selected artificial intelligence programs are created and function to business objectives, integrating innovation in all business functions. For the transition in the data-driven or at an increased level of artificial intelligence-driven solutions, firms will have to implement a culture of experimentation inclination and continuous enhancement by starting small and applying short and incremental cycles, evolving true artificial intelligence revolution over time.

Irrelevant Spending

Structure Design

Organizations can use artificial intelligence to monitor the conflicting issues in the organization's structure that will take into account the capabilities, strategy, and unique characteristics. As a result, it will minimize irrelevant spending and increase growth and profitability.

Roles Management

Artificial intelligence in firms is an advancement that can allow managers to become excellent. Artificial intelligence can be used in many facets, from enhancing relationships with staff and consumers to distinguishing patterns in excessive data volume to repetitive tasks.

Responsibilities Allocation

Responsibilities allocation is one of the secrets to boosting organizational benefits by managing as many tasks as possible. Numerous computational multi-agent systems utilize the capability of agents for responsibilities allocation.

Communication Management

Artificial intelligence has powered chatbots to automate and manage communication, and they are substitutes for dealing with humans. These chatbots can control communication in various ways to engage with consumers, for example, responding to queries or providing assistance.

Business Survival

Value Chain Analysis

Using artificial intelligence in value chain analysis, managers can improve their decision-making procedures by forecasting unexpected abnormalities, building up bottlenecks, and finding solutions to restructuring manufacturing schedules that tend to be increasingly inconstant because of dependencies in production operations.

Transformation Planning

Incredible precision can be attained through transformation planning by implementing artificial intelligence using deep neural networks. Furthermore, it helps to get the most out of data by utilizing the latest learning algorithms, proving it is a flexible technology (Smith, 2019).

Innovation

Artificial intelligence and innovation together can enhance many business areas. For example, in customer service, chatbots are now communicating with online consumers to improve customer service. In addition, artificial intelligence is being utilized for the HR department to accelerate the recruitment procedure, and for the marketing department, artificial intelligence-powered tools are used to customize the consumer experience (Liu, 2018).

Continuous Improvement

Using artificial intelligence, systems can check thousands of mathematical models of manufacturing and outcome possibilities and be more accurate about the analysis during the adoption of new information, for example, new products, supply chain disruptions, and unexpected changes in demand. Thus, it helps in continuous improvement for firms to attain a competitive edge in the market.

Lack of Scarcity

Decision Automation

Artificial intelligence can boost human intelligence and allow intelligent decision-making. It helps in the detection of wrong decisions and accelerates the whole decision-making procedure. Artificial intelligence enables the automation of decision-making without human interference (Mozer et al., 2019).

Objectives Management

Artificial intelligence has become an essential facet for many firms. It helps in many tasks in the firm and aids in modernizing business procedures and objective management, supporting the firm to perform more efficiently and achieve the firm's objectives more proficiently.

Key Actions

Many firms focus on the outcomes of artificial intelligence. For firms concerned with minute details, there are four key actions to recognize: collaborative filtering, categorization, machine learning, and classification. These four key actions also signify the steps of the analytical procedure (Lutz, 2019).

Management Transparency

Management transparency is the way firms and leaders behave and think. The firms using artificial intelligence require more openness, accountability, and communication between employees and managers (Lui and Lamb, 2018).

ARE WE READY FOR AI?

Artificial intelligence is changing the rules of competition within industries worldwide. Opportunities associated with AI are considered the most important technological growth regarding its vast potential for adding business value and competitive advantage (Miailhe and Hodes, 2017).

AI applications and adoption offer each business entity as many new challenges as it does opportunities. AI technology has already transformed businesses everywhere, small or large, developed or start-ups. AI has the potential to level the playing field. However, it is essential to understand the other factors that will help drive successful AI capabilities. These include cognitive-based technologies like machine learning, natural language processing, and robotics. Those cognitive-based technologies will ultimately have a tremendous impact on every business level. It is acknowledged that a proper understanding of the issue and keen insights into this change is not an option. It is "a must."

It raises a critical question that will be addressed in this section, the question is "***Are we ready for AI?"***

In light of IT Masters definition appraised in Westerman et al. (2014), and to answer the question "Are we ready for AI?," we focus on two different AI application scenarios found in today's business organizations. First, Westerman refers to organizations that apply AI technologies seamlessly to every aspect of their businesses as "AI Masters." They know exactly where and how to invest in AI opportunities, knowing the impact of those investments. AI Masters can see AI as the best way to increase business value, increase efficiency gains, and gain a competitive advantage. It could be achieved by applying AI solutions that ultimately impact customer relations, customer engagements, internal/external business operations, customer expectations, and even business models. AI Masters deeply understand the implications of evolving AI-driven automation ecosystems far beyond narrow AI applications. Accordingly, AI is not limited to changing how business works. It is also fundamentally transforming the traditional thinking and meaning of innovation.

Some organizations adopt new AI technologies without having a real strategy and without fully determining how AI technologies will be integrated within the organization. As a result, those organizations invest large amounts in their AI solutions. However, because they lack a strong and clear vision of the future, they will waste most of what they have invested. In this scenario, organizations have already invested in AI applications such as robots, AI power assistants such as virtual assistant shoppers and chatbots, fraud preventions, etc. However, they lack strong leadership capabilities, which lead to the definite inability to realize the concrete benefits of their AI investments.

It is widely acknowledged that "replacing a man with a machine" is not a fashion that every organization should follow, bearing in mind that "*why to replace*" is as important as "*how to replace*." These organizations tend to be mature in answering the second question but immature in answering the first one question. It leads to a narrowly overlaid AI adoption strategy as they mainly focus on using AI to change the way they provide the services without understanding why to they need to change it. In this way, they fail to answer questions such as why analyze data? Why predict performance? And why transform? As a result, the AI trend in applications in this scenario does not respond adequately to rapidly evolving intelligence capability. It will negatively affect their understanding of the broader AI trends on the horizon and their future AI preparations.

SUMMARY OF FINDINGS AND CONCLUSION

Artificial intelligence is becoming a commodity as more solutions are readily available for the end-users. This paper addressed the important aspects of AI as a commodity and provided a closer look at the various perspectives on the usage of AI in various fields. The findings show that AI commoditization is in the near future and must be avoided to sustain strategic differentiation. Otherwise, AI will have the same fate as previous technologies, i.e., information technology. The arguments on AI commoditization have been discussed in detail to show the leading factors of commoditization. Compiling leadership and AI capabilities will achieve greater performance than either dimension can deliver on its own. The organizations that excelled in AI capability and leadership capability may have higher financial outcomes. More relevant and rigorous studies are needed to shed light on the importance of improving the organizations' AI-focused leadership capabilities and how to enhance the business model to adapt the AI application, bearing in mind that the inability of the organizations to enhance their leadership/AI technical capabilities will lead to ultimate failure and unwanted consequences. The previous discussion has drawn us to define some other parameters that may affect AI application success, mainly the leadership capability that is as important as the AI technological capability.

Good AI leadership capability is the lever that uses AI technology for real business transformation. They do not apply the "bottom-up" AI applications model. Instead, they have a very strong "top-down" leadership model by setting the AI directions, building momentum, measuring the initiatives, and ensuring that the organizations can follow through. More specifically, the top-down leadership model occurs by setting up clear AI goals, then engaging their employees by energizing them to drive through the AI journey. Leaders know what they are aiming for. Even though they believe in the workforce as one key asset in their organizations, at the same time, they know exactly when "replacing a machine with a human" should happen. The AI commodity model links the arguments with the organizational framework used to add value for the organizations facing the commoditization issue. The AI-based organizational framework addressed the value added

to each phase of the commodity model. The framework can be used to drill down to the possible options available for organizations and end-users to have a clear pathway for the usage of AI responsibly. The history and arguments validate that shortly the value of AI solutions will not impact businesses. For this purpose, an artificial intelligence-based organizational framework is proposed that provides an overview of value-added features that can help organizations lower the impact of AI commoditization.

AUTHOR CONTRIBUTIONS

MT worked on the conception and design of the idea which was given by AA. MT and AA worked on the writing literature review, perspectives, and overall organization of the paper. SS worked on the technical aspects of artificial intelligence and machine learning. All three worked together on writing different sections of the paper and contributed to read, revise, and approved the submitted version.

REFERENCES

Abbass, H. A. (2019). Social integration of artificial Intelligence: functions, automation allocation logic, and human-autonomy trust. *Cognit. Comput.* 11, 159–171. doi: 10.1007/s12559-018-9619-0

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Exploring the impact of artificial Intelligence: prediction versus judgment. *Inform. Econ. Policy* 47, 1–6. doi: 10.1016/j.infoecopol.2019.05.001

Bauguess, S. W. (2017). The role of big data, machine learning, and AI in assessing risks: a regulatory perspective. machine learning, and ai in assessing risks: a regulatory perspective (June 21, 2017). SEC Keynote Address: OpRisk North America.

Braga, A., and Logan, R. K. (2017). The emperor of strong AI has no clothes: limits to artificial intelligence. *Information* 8:156. doi: 10.3390/info8040156

Britannica (2021). Available online at: https://www.britannica.com/

Cabitza, F., Campagner, A., and Balsano, C. (2020). Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. *Ann. Transl. Med.* 8:63. doi: 10.21037/atm.2020.03.63

Carr, N. G. (2003). IT doesn't matter. *Educause Rev.* 38, 24–38.

Carter, S., and Nielsen, M. (2017). Using artificial Intelligence to augment human Intelligence. *Distill* 2–e9. doi: 10.23915/distill.00009

Cockburn, I. M., Henderson, R., and Stern, S. (2018). *The impact of artificial Intelligence on innovation* (No. w24449). National bureau of economic research.

Corea, F. (2017). *Artificial Intelligence and Exponential Technologies: Business Models Evolution and New Investment Opportunities.* New York, NY: Springer.

Davenport, T. H., and Ronanki, R. (2018). *Artificial Intelligence for the Real World.* Available online at: https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and use AI in a Responsible Way*. New York, NY: Springer Nature.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). "Explainable artificial Intelligence: a survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (New York, NY: IEEE), 0210–0215.

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. (2019). Artificial Intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *Int. J. Inf. Manage.* 48, 63–71. doi: 10.1016/j.ijinfomgt.2019.01.021

Dunjko, V., and Briegel, H. J. (2018). Machine learning and artificial Intelligence in the quantum domain: a review of recent progress. *Rep. Progr. Phys.* 81:074001. doi: 10.1088/1361-6633/aab406

Floridi, L. (2020). "What the near future of artificial Intelligence could be," in *The 2019 Yearbook of the Digital Ethics Lab* (Cham: Springer), 127–142.

Gabriel, I. (2020). Artificial Intelligence, values, and alignment. *Minds Mach.* 30, 411–437. doi: 10.1007/s11023-020-09539-2

Galanos, V. (2018). "Artificial Intelligence does not exist: lessons from shared cognition and the opposition to the nature/nurture divide," in *IFIP International Conference on Human Choice and Computers* (Cham: Springer), 359–373.

Galbusera, F., Casaroli, G., and Bassani, T. (2019). Artificial Intelligence and machine learning in spine research. *JOR spine* 2:e1044. doi: 10.1002/jsp2.1044

Goertzel, B. (2017). *What Counts as a Conscious Thinking Machine* (accessed December 26, 2017).

Greene, D., Hoffmann, A. L., and Stark, L. (2019). "Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial Intelligence and machine learning," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 258.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer* 51, 28–36. doi: 10.1109/MC.2018.3620965

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., et al. (2020). Transparency and reproducibility in artificial Intelligence. *Nature* 586, E14–E16. doi: 10.1038/s41586-020-2766-y

Hall, W., and Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*. Department for Digital, Culture, Media and Sport and Department for Business, Energy and Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth.

Hildt, E. (2019). Artificial Intelligence: Does consciousness matter? *Front. Psychol.* 10:1535. doi: 10.3389/fpsyg.2019.01535

Huang, M. H., and Rust, R. T. (2018). Artificial Intelligence in service. *J. Serv. Res.* 21, 155–172. doi: 10.1177/1094670517752459

Iphofen, R., and Kritikos, M. (2019). Regulating artificial Intelligence and robotics: ethics by design in a digital society. *Contemp. Soc. Sci.* 16, 1–15. doi: 10.1080/21582041.2018.1563803

Kaplan, A., and Haenlein, M. (2020). Rulers of the world, unite! the challenges and opportunities of artificial Intelligence. *Bus. Horiz.* 63, 37–50. doi: 10.1016/j.bushor.2019.09.003

Katz, Y. (2017). *Manufacturing an Artificial Intelligence Revolution*. Available at SSRN 3078224.

Keyes, O., Hitzig, Z., and Blell, M. (2021). Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplin. Sci. Rev.* 46, 158–175. doi: 10.1080/03080188.2020.1840224

Kim, H. S. (2020). Decision-making in artificial Intelligence: is it always correct?. *J. Korean Med. Sci.* 35:e1. doi: 10.3346/jkms.2020.35.e1

Lawless, W. F., Mittu, R., Sofge, D., and Hiatt, L. (2019). Artificial Intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI. *AI Magazine* 40, 5–13. doi: 10.1609/aimag.v40i3.2866

Leslie, D. (2019). *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. Available at SSRN 3403301

Liu, H. Y. (2018). The power structure of artificial Intelligence. *Law Innov. Technol.* 10, 197–229. doi: 10.1080/17579961.2018.1527480

London (1947). *Alan Turing and the beginning of AI*. Britannica. Available online at: https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nat. Hum. Behav.* 1, 1–3. doi: 10.1038/s41562-016-0028

Lui, A., and Lamb, G. W. (2018). Artificial Intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. *Inform. Commun. Technol. Law* 27, 267–283. doi: 10.1080/13600834.2018.1488659

Lutz, C. (2019). Digital inequalities in the age of artificial Intelligence and big data. *Hum. Behav. Emerg. Technol.* 1, 141–148. doi: 10.1002/hbe2.140

Miailhe, N., and Hodes, C. (2017). The third age of artificial Intelligence. field actions science reports. *J. Field Act.* 17, 6–11.

Mozer, M. C., Wiseheart, M., and Novikoff, T. P. (2019). Artificial Intelligence to support human instruction. *Proc. Nat. Acad. Sci.* 116, 3953–3955. doi: 10.1073/pnas.1900370116

Parson, E., Re, R., Solow-Niederman, A., and Zeide, E. (2019). Artificial Intelligence in strategic context: an introduction. *Public Law Res. Pap.* 39, 19–45. doi: 10.2139/ssrn.3476384

Petrović, V. M. (2018). Artificial Intelligence and virtual worlds–toward human-level AI agents. *IEEE Access* 6, 39976–39988. doi: 10.1109/ACCESS.2018.2855970

Ramamoorthy, A., and Yampolskiy, R. (2018). Beyond mad? the race for artificial general Intelligence. *ITU J.* 1, 1–8.

Roff, H. M. (2019). Artificial Intelligence: Power to the people. *Ethics Int. Affairs* 33, 127–140. doi: 10.1017/S0892679419000121

Russell, S. (2017). Artificial intelligence: the future is superintelligent. *Nature* 548, 520–521. doi: 10.1038/548520a

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial Intelligence. *Proc. Nat. Acad. Sci.* 117, 30033–30038. doi: 10.1073/pnas.1907373117

Semmler, S., and Rose, Z. (2017). Artificial Intelligence: application today and implications tomorrow. *Duke L. and Tech. Rev.* 16:85.

Siau, K., and Wang, W. (2018). Building trust in artificial Intelligence, machine learning, and robotics. *Cut. Bus. Technol. J.* 31, 47–53.

Smith, B. C. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: Mit Press.

Townsend, D. M., and Hunt, R. A. (2019). Entrepreneurial action, creativity, and judgment in the age of artificial Intelligence. *J. Bus. Ventur. Insights* 11:e00126. doi: 10.1016/j.jbvi.2019.e00126

Vähäkainu, P., and Lehto, M. (2019). "Artificial Intelligence in the cyber security environment," in *ICCWS 2019 14th International Conference on Cyber Warfare and Security: ICCWS 2019* (Oxford: Academic Conferences and publishing limited), 431.

Villaronga, E. F., Kieseberg, P., and Li, T. (2018). Humans forget, machines remember: artificial Intelligence and the right to be forgotten. *Comput. Law Secur. Rev.* 34, 304–313. doi: 10.1016/j.clsr.2017.08.007

Wang, P. (2019). On defining artificial Intelligence. *J. Artific. Gene. Intell.* 10, 1–37. doi: 10.2478/jagi-2019-0002

Westerman, G., Bonnet, D., and McAfee, A. (2014). *Leading Digital: Turning Technology into Business Transformation*. Boston, MA: Harvard Business Press.

Wilner, A. S. (2018). Cybersecurity and its discontents: Artificial intelligence, the Internet of Things, and digital misinformation. *Int. J.* 73, 308–316. doi: 10.1177/0020702018782496

Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *Int. J. Pub. Admin.* 42, 596–615. doi: 10.1080/01900692.2018.1498103

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership