Computers in Human Behavior Reports 17 (2025) 100577

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Review

# Computational persuasion technologies, explainability, and ethical-legal implications: A systematic literature review

Davide Calvaresi [a,*], Rachele Carli [b], Simona Tiribelli [c], Berk Buzcu [a,d], Reyhan Aydogan [e,d], Andrea Di Vincenzo [a], Yazan Mualla [f], Michael Schumacher [a], Jean-Paul Calbimonte [a,g]

[a] University of Applied Sciences and Arts Western Switzerland HES-SO, Switzerland
[b] University of Luxembourg, Luxembourg
[c] Università di Macerata, Italy
[d] Ozyeugin University, Turkey
[e] Delft University of Technology, Netherlands
[f] UTBM, CIAD UMR, France
[g] The Sense Innovation and Research Center, Switzerland

ARTICLE INFO

*Keywords:*
Computational persuasion
eHealth
Behavior change
Ethics
Legal implications
Systematic literature review

ABSTRACT

This paper conducts a systematic literature review (SLR) to evaluate the effectiveness of computational persuasion technology (CPT) in the eHealth domain. Over the past fifteen years, CPT has been used in various scenarios, from promoting healthy diets to supporting chronic disease management. Despite the proliferation of intelligent systems and Web-based applications, the ethical and legal nuances of these technologies have become increasingly significant. The review follows a structured methodology, assessing 92 primary studies through sixteen research questions covering demographics, application scenarios, user requirements, objectives, functionalities, technologies, advantages, limitations, proposed solutions, ethical and legal implications, and the role of explainable AI (XAI). The findings indicate that while CPT holds promise in inducing behavioral change, many prototypes remain untested on a large scale (60% of surveyed studies only developed at a conceptual level), and long-term effectiveness is still uncertain (36% report attaining their goals, but none focuses on long-term assessment). The study highlights the need for more comparative analyses of persuasion models and tailored approaches to meet diverse user needs. Ethical and legal concerns, such as patient consent, data privacy, and potential for users' manipulation, are under-explored and require deeper investigation. The paper recommends a bottom-up regulatory approach to create more effective and flexible ethical and legal guidelines for CPT applications.

In conclusion, significant advancements have been made in CPT for eHealth, but ongoing research is essential to address current limitations, enhance user acceptability and adherence, and ensure ethical and legal soundness.

## 1. Introduction

Healthcare systems have evolved over the past decades, moving toward patient-centered care to improve medical indicators and quality of life in general. People have progressively become more autonomous in adopting healthy behaviors, mainly through active health education, ensuring appropriate follow-up of care, and monitoring by health professionals. The emergence of digital health solutions has been pivotal in this transformation, in particular for personalized interventions that focus on health behavioral change. The active participation of individuals in improving their health has been based on increasingly advanced persuasion techniques based on behavioral theories (Taj, Klein, & van Halteren, 2019). Moreover, Persuasion Technologies (PT) are addressed from trustworthiness (e.g., interpretability/explainability), ethical, and legal perspectives. Providing the user with (textual/graphical) *explanations* can shed light on the system's decision-making process (Graziani et al., 2023; Gunning & Aha, 2019). Such transparency generally serves two key purposes: (i) building user trust and (ii) fostering a dialogic

* Corresponding author.
*E-mail addresses:* davide.calvaresi@hevs.ch (D. Calvaresi), rachele.carli2@unibo.it (R. Carli), simona.tiribelli@unimc.it (S. Tiribelli), berk.buzcu@hevs.ch (B. Buzcu), reyhan.aydogan@ozyegin.edu.tr (R. Aydogan), andrea.divincenzo@master.hevs.ch (A. Di Vincenzo), yazan.mualla@utbm.fr (Y. Mualla), michael.schumacher@hevs.ch (M. Schumacher), jean-paul.calbimonte@hevs.ch (J.-P. Calbimonte).

https://doi.org/10.1016/j.chbr.2024.100577
Received 16 July 2024; Received in revised form 19 December 2024; Accepted 19 December 2024
Available online 28 December 2024

experience where users feel more engaged in the decision-making process. Indeed, the process of persuasion is intended as "an activity that involves one party trying to induce another party to believe something or to do something" (Hunter, 2018). However, distinguishing persuasion from other forms of non-legitimate will alteration is imperative. Indeed, unlike manipulation and coercion (Carli, Najjar, & Calvaresi, 2022), persuasion influences the architecture of choices, *leaving* individuals with all the alternatives they already possess and, potentially, enriching them (Thaler & Sunstein, 2009). Thus, persuasion technologies have always been used by caregivers to encourage patients to adopt positive health behaviors (e.g., quitting smoking, eating healthily, practicing sports, etc.) and have been demonstrated essential to guide people and help them to avoid harmful actions to themselves (Adaji & Adisa, 2022; Sara & Mostafa, 2019). To develop Computational Persuasive Technologies (CPT) that have a direct impact on the effectiveness of a desired behavioral change, without leading to manipulative dynamics that distort individual perception and/or will, is still an open challenge. Several existing studies assess the impact of CPT. Nonetheless, to date, the results are divergent as the development of CPT must still be considered at their early stages. Insofar, off-the-shelves scientific studies propose mostly conceptual/prototype-level analysis, with only a few practical tests. This makes the interpretation of results and the large-scale use of these technologies complex. As a result, even an accurate and satisfactory analysis of the possible ethical and legal implications and the social impact is also lacking and difficult to achieve — particularly for the long-run effects. However, ethical and legal implications of CPT in healthcare need to be considered. Indeed, while the corpus of ethical and legal scholarship on CPT in healthcare – as we will show – is still limited, AI ethics and legal research on AI in healthcare have skyrocketed in the last few years (Babic, Gerke, Evgeniou, & Cohen, 2021; Bouderhem, 2024; Fosch-Villaronga, Chokoshvili, Vallevik, Ienca, & Pierce, 2021; Gerke, Minssen, & Cohen, 2020; Giovanola & Tiribelli, 2023; Mennella, Maniscalco, De Pietro, & Esposito, 2024; Morley et al., 2020; Reddy, Allan, Coghlan, & Cooper, 2020). Particularly, core ethical concerns in this debate revolve around key themes such as the transparency of AI systems, data privacy, and the protection of patient autonomy. The main research in AI ethics in healthcare addresses transparency as accessibility. On the one side, transparency is considered pivotal to ensuring health AI technology is accurate and fair (Giovanola & Tiribelli, 2023); on the other side, transparency is also questioned as it conflicts with AI systems security (Tiribelli, Monnot, et al., 2023). Indeed, open-source approaches make AI systems vulnerable to cyberattacks and data breaches that are problematic due to the sensitivity of health matters and the importance of patients' privacy on health issues. Thus, research focuses prominently on intelligibility over transparency, that is, how to make AI systems understandable by users, and on explainability (XAI) methods when "black boxes" AI models are involved. Protecting patients' autonomy is also a major ethical concern healthcare AI raises. AI ethics and policy scholarship mainly focus on how to ensure that patients and consumers are fully informed and understand the risks and benefits of a particular health AI technology and voluntarily consent to it (Liao, 2023). Autonomy is mainly addressed through the protection of the patient's right to decide to opt-out from AI use and control over AI systems, prevention of AI human overreliance and manipulation risks, and the implementation of value alignment design (i.e., the design of health AI systems aligned to user's values, preferences, and goals Tiribelli & Calveresi, 2024). Hence, if and how these issues are considered in the scholarship on health CPT is to be investigated.

This work provides an in-depth reflection by conducting a systematic literature review (SLR) focused on assessing the CPT effectiveness in changing user behavior. In particular, it addresses sixteen research questions, including aspects such as demographics, application domains, end-users, requirements, objectives, technologies, strengths, limitations, explanation generation implementations, and future challenges of the solutions found in the literature. Furthermore, it seeks to extrapolate possible ethical/legal issues from the technical characteristics and analyses already carried out in the literature. By doing so, it is possible to (i) raise the necessary attention of future research and (ii) suggest regulatory approaches and solution strategies – where achievable – starting from the theoretical principles already present in doctrine and jurisprudence for similar or comparable cases. The goal is to provide a tool for researchers, software engineers, innovation managers, and other practitioners to investigate the current state of the art and discuss the open challenges.

The remainder of the paper is structured as follows: Section 2 presents the methodology applied for performing the systematic literature review, including the review planning phase, the definition of the protocol, and the research questions. Section 3 analyzes the outcomes of the applied methodology structured according to the research questions. Section 4 discusses the obtained results, projecting them into the stated (by the primary studies) and envisioned (by the authors of this paper) future directions. Finally, Section 5 concludes the paper.

## 2. Review methodology

The approach employed in this paper aims to be rigorous and reproducible. It relies on the methodology outlined by Kitchenham et al. (2009), and comprises three stages: (P1) Planning the review, consisting of defining the main generic question(s) and deriving structured research questions, characterizing the search protocol, and validating the protocol; (P2) Performing the review, which entails the collection and selection of literature, elaboration, and disagreement resolution; (P3) Dissemination, including analysis, documentation, reporting, and summary of learned lessons (see Fig. 1).

### 2.1. Review planning

This section describes the definition of the research questions, the development of the protocol, the search strategy, inclusion and exclusion criteria, and disagreement resolution. Over the years, several research studies have addressed computational persuasion techniques in the healthcare domain. These efforts are very different in terms of the goals they pursue, their specific subdomains of application, their degree of technological advancement, the persuasion models they rely on, etc. In this context, the main research question set for this literature review is: *What are the challenges addressed by computational persuasion technologies, and to what extent do they contribute to the e-health domain?*

To further investigate this question and its implications, we formulated a set of structured research questions, following the Goal-Question-Metric (GQM) methods (Galster, Weyns, Tofan, Michalik, & Avgeriou, 2014; Kitchenham et al., 2010).

**SRQ1** Demographics. *What is the temporal and geographical distribution of research works in computational persuasion?*

**SRQ2** Abstraction. *What is the abstraction level of the elaborated scientific contributions? E.g., at which level the contribution is: conceptual (C), prototype (P), or tested (T)?*

**SRQ3** Application scenarios. *Within the e-health domain, in which application scenarios (e.g., chronic diseases, nutrition, etc.) have computational persuasion solutions been employed?*

**SRQ4** Users. *Who are the users (recipients) of computational persuasion solutions? E.g., oncology patients, diabetic users, people affected by chronic diseases, etc.*

**SRQ5** Requirements. *What are the requirements standing behind the employment of computational persuasion technologies?*

**SRQ6** Objectives. To investigate what the CPT targeted to increase the effectiveness of health-related interventions, we set: *Which are the goals of the CPT solutions?*

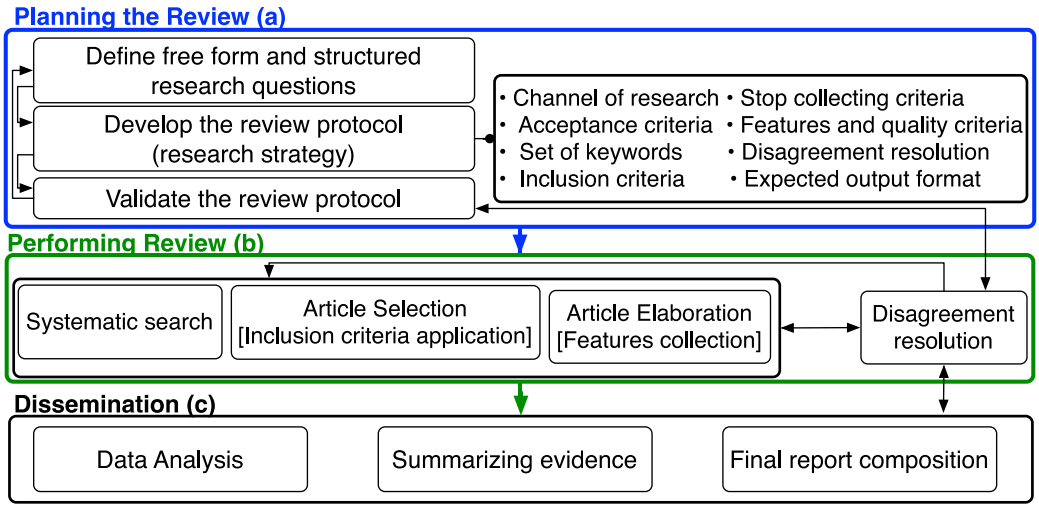



**Fig. 1.** Systematic Literature Review phases (Kitchenham et al., 2009).

**SRQ7** Functionalities realized. *Which computational persuasion functionalities have been implemented?*

**SRQ8** Technology. *Which underlying technologies are employed by computational persuasion systems?*

**SRQ9** Explanations in CPT *What are the role and involvement of explanations in persuasive systems in e-Health?*

**SRQ10** Advantages. *Which advantages are provided by computational persuasion technologies for their users?*

**SRQ11** Drawbacks. *Which limitations affect existing computational persuasion solutions?*

**SRQ12** Proposed solutions. *Which are the solutions identified to overcome the limitations identified in SQR11?*

**SRQ13** Functionalities envisioned and future challenges. *Which computational persuasion functionalities are envisioned to be realized as future work?*

**SRQ14** Adverse effects of persuasion. *To what risks may the user of persuasive systems be exposed?*

**SRQ15** Legal implications. *What legal problems may arise from the use of persuasive systems in e-Health?*

**SRQ16** Ethical implications. *Which ethical implications could affect the existing computational persuasion solutions?*

### 2.2. Review protocol

The *search strategy* included the selection the following information sources: IEEE Xplore,[1] ScienceDirect,[2] ACM Digital Library,[3] Citeseerx,[4] Pubmed.[5] The keyword selection relied on the reviewers' background and knowledge in the context of computational persuasion, and they include the following: *Persuasive design, persuasion dialogues, persuasive, computational models of argument, health behavior change, behavior change theory, behavior change intervention, behavior counseling, e-health, lifestyle intervention, supportive care, argumentation strategies, transparency, fairness, accountability, ethics, bias, privacy, autonomy, and manipulation*.

The purpose and scope of the review demanded combining the identified keywords instead of using them individually. For each combination, new articles related to the research questions were selected according to their relevance. The process stopped once the combinations had reached saturation and no more relevant articles could be found.

#### 2.2.1. Inclusion and exclusion criteria

The initial search collected 220 papers. In turn, we have filtered the paper by assessing their titles and abstract against the following inclusion criteria:

- *Time:* Papers published between 2010–2024.
- *Context:* The primary studies we considered are those conducted in the crucial areas of patient monitoring, home care support, healthcare, and behavior change.
- *Purpose:* The purpose should be related to one of these goals: improve patient life, patient empowerment, autonomy, and adherence.
- *Users:* The beneficiaries of the solutions presented in primary studies are patients, relatives, caregivers, and physicians.

#### 2.2.2. Biases and disagreement resolution

To minimize biases and resolve disagreements during the feature classification process, reviewers responsible for method development and data elaboration were instructed to cross-examine each task. For the article selection phase, three reviewers collaborated to cross-validate the inclusion/exclusion criteria. The papers were divided into three sets, with each set assigned to two reviewers, who independently applied the criteria to ensure objective assessment. Their evaluations were kept confidential to prevent influencing each other. Once the review process was complete, the results were compared. Any disagreements, which could be either methodological (e.g., study design or quality) or ethical (e.g., ethical implications of persuasive technologies), were resolved by a third reviewer. The third reviewer played a key role in ensuring that disputes were addressed objectively, particularly in cases involving ethical concerns. These were evaluated against principles such as autonomy, consent, and fairness, based on established ethical frameworks from the literature. This process helped ensure that the final decisions reflected both methodological rigor and ethical soundness.

[1] http://ieeexplore.ieee.org
[2] http://www.sciencedirect.com/
[3] http://dl.acm.org/
[4] http://citeseerx.ist.psu.edu/index
[5] http://www.ncbi.nlm.nih.gov/pubmed

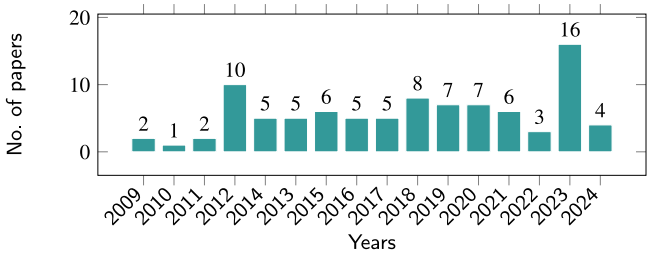
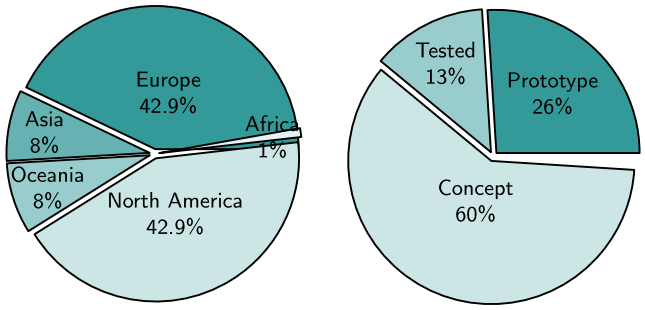



**Fig. 2.** Temporal distribution of the primary studies.




(a) Geographic distribution (b) Type of studies

**Fig. 3.** Geographic distribution and abstraction.

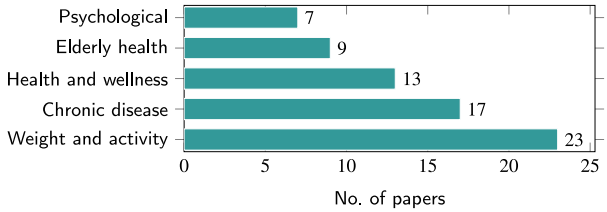



**Fig. 4.** Contributions per application scenario.

#### 2.2.3. Features and quality criteria

Evaluating and processing the quality of the primary studies are complex and crucial tasks in an SLR (Kitchenham & Charters, 2007). The approach presented by Galster et al. (2014) classifies primary studies by context, research justification, rational, critical examination, statement of findings, and the presence of biases and possible limitations of credibility. This interpretation and the analysis of the results was made complex because, in various articles, the information sought by the structured questions was not always explicitly reported. To facilitate the assessment of the quality of the information, the Y-P-N classification was used in this work, Y = information is explicitly defined/evaluated; P = information is implicit/stated; N = information is not inferable in accordance with the DARE criteria (Kitchenham et al., 2009).

## 3. Review results and analysis

Once the analysis of the inclusion criteria had been performed, only 92 articles were finally retained for this literature review. Below, we structure the results of the review according to the research questions defined in Section 2.1.

### *SRQ1 — demographics*

Figs. 2 and 3(a) show the temporal and geographical distribution of papers. The paper selection and analysis have been conducted in April 2024. Fig. 2 shows the demographic evolution of the selected primary studies over time. Considering that the need for CPT is not satisfied yet, the two peaks (2012 and 2023) can hint at cycles of technological advancements and (somewhat) limitations in the adoption. The primary studies are conducted in twelve countries. The results, reporting the number of publications per continent, showed that the highest number of publications was recorded in North America and Europe with almost half of the studies (Fig. 3(a)) followed by Asia. Oceania and Africa only represented less than 10% combined. This distribution could be due to the affordance of new technologies in the health field and the lack of dedicated funds. Note that the selected studies were not additionally filtered by a geographical criterion. We merely report the distribution of the studies that met the established inclusion criteria.

### *SRQ2 — abstraction*

The abstraction level of the majority of the studies (60%) is conceptual, highlighting the need to assess theoretical/small scales and still underdeveloped approaches and technologies. This analysis is also conveyed by Dominic, Hounkponou, Doh, Ansong, and Brighter (2013), who highlight the need for the developers to properly assess the complex users' socio-environmental context before approaching a prototype. Technical papers addressing *more advanced* prototypes and tested solutions are less frequent (13% – see Fig. 3(b)) and, on a few occurrences, fail to address the implications entangling persuasion theories and technological choices.

### *SRQ3 — application scenario*

In the healthcare domain, application scenarios where CP has been used are broad and disparate (see Fig. 4). Such heterogeneity has exacerbated the efficiency assessment and comparison of the elaborated studies. In doing so, the primary studies have been clustered in weight control and physical activity (33.3% - Weight and activity) – e.g., (Asbjørnsen et al., 2020; Wiafe & Nakata, 2010); physical/mental wellbeing (42% - Psychological & health and wellness & elderly health) – e.g., (Orji & Moffatt, 2016; Oyebode, Steeves, & Orji, 2024; Spanakis, Santana, Ben-David, Marias, & Tziraki, 2014); and chronic diseases (24.6% - Chronic disease) – e.g., (Almutairi, Vlahu-Gjorgievska, & Win, 2023; Bartlett, Webb, & Hawley, 2017; Samonte, Medina, San Juan, & Celestial, 2023). Some cross-cluster examples worth mentioning are diabetes (Jalil, 2013; Jalil & Orji, 2016; Kim et al., 2019), older individuals' health (Chatterjee et al., 2012; Srisawangwong & Kasemvilas, 2014), pulmonary disease (Bartlett et al., 2017), and rheumatoid arthritis (Srisawangwong & Kasemvilas, 2014). Most studies highlight how challenging cross- and domain-specific requirements are and how important it is to develop applications that suit the given patients/users. For example, when it comes to including new technologies in daily living, older individuals have more resilient needs and lower acceptability (Bartlett et al., 2017; Cabrita, Akker, Tabak, Hermens, & Vollenbroek-Hutten, 2018; Lee, Helal, Anton, Deugd, & Smith, 2012).

### *SRQ4 — intended users*

Users of computational persuasion solutions include direct users (patients, caregivers, etc.) and indirect users (family, relatives, application developers). In this analysis, the most represented user combination is the one of persuadees, i.e., users that are influenced by the system. At the same time, there are also numerous contributions where developers are also targeted. This could be explained by the large number of studies dealing with persuasion at a conceptual level, providing information for the development of these methods and the underlying technical details. This may indicate that computational persuasion technologies are still at an early implementation stage in the health sector (see Fig. 5).

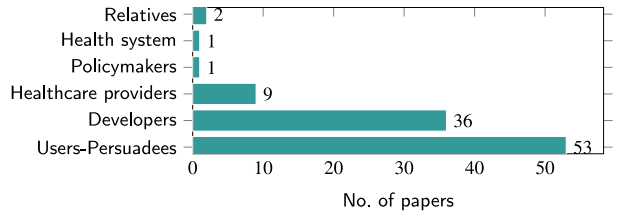



**Fig. 5.** Number of papers per type of end users.

*SRQ5 — requirements*

Table 1 shows the main requirements elicited from the primary studies. The most stated (or elicited) requirement is *"persuasion systems must be adapted to the users' needs, their environment, and social context"*. For example, on the one hand, Lentferink et al. (2017), Mylonopoulou (2018), Srisawangwong and Kasemvilas (2014) indicate that patients have to receive social support and attention, particularly when the professionals are impossibilitated (e.g., due to lack of time). This entails patients receiving motivational messages and being able to exchange them with other peers and stakeholders. On the other hand, caregivers benefit from platforms providing social assistance and guidance for their tasks (Premanandan, Ahmad, Cajander, Agerfalk, & Gemert-Pijnen, 2023). CPT (supporting, tracking, monitoring, etc.) are undoubtedly useful — yet they can apport a burden. Therefore, explaining and making the CPT's benefits evident while using the support systems is imperative. Wiafe and Nakata (2010) endeavored to emphasize the importance of being careful while designing the application to be able "to provide information on their performance'. Social support, as well as tracking and monitoring, showed particularly positive results when integrated into computational persuasion solutions if the selected persuasive components were also taken into account (Oyebode & Orji, 2023). Additional improvements could be obtained by shifting the persuasion strategy based on users' emotions (Oyebode et al., 2024). Schnall, Bakken, Rojas, Travers, and Carballo-Dieguez (2015) and Lee et al. (2012), among others, include "to ensure autonomy" as a key requirement to influence individuals' motivation to change a behavior in the well-being domain. Although it is difficult to compare results due to the studies' heterogeneity, the application design is common ground. Providing visual/graphical elements supports individuals' motivation dramatically. The application must also be able to promote health while remaining accessible in terms of price, usability, and confidentiality. This last aspect is particularly crucial due to the sensitive nature of health-related data.

*SRQ6 — objectives of the studies*

The objectives identified in the primary studies are reported to be positively attained in 36% of the cases ( Table 2). More than half (56%) reported only partially achieving their objectives, and 8% lack an explicit achievement assessment.

The general aim of almost all the studies was to contribute to the design of a fitness or health application that, through persuasion technologies, improves users' health (e.g., Cabrita et al., 2018). To this end, authors have adopted different approaches and starting points. However, only a few studies focus their research on assessing the given "type" of persuasive model. Thus, a concrete analysis discussion about their efficacy and effectiveness in changing user behavior is unattainable. Moreover, these studies primarily evaluate short-term metrics, often overlooking their long-term effects. Other studies tackle understanding the users' contextual needs (e.g., Bartlett et al., 2017). The cultural dimension is also considered crucial and necessary to be included in the application development process. Srisawangwong and Kasemvilas (2014) evaluated this element, assessing how to design the user interfaces and specify the persuasive domain.

**Table 1**
Requirements.

| Studies | Requirements |
|---|---|
| Henkemans, Paradies, Neerincx, Looije, and Pepijn Emepelen (2015), Lentferink et al. (2017), Oyebode and Orji (2023), Oyebode et al. (2024), Oyibo (2016), Srisawangwong and Kasemvilas (2014), Yoganathan and Sangaralingam (2015) | Adapt to users' context |
| Lentferink et al. (2017), Mylonopoulou (2018, 2018), Premanandan et al. (2023), Srisawangwong and Kasemvilas (2014), Wiafe and Nakata (2010) | Social support |
| Cabrita et al. (2018), Mylonopoulou (2018, 2018), Premanandan et al. (2023, 2023), Wiafe and Nakata (2010) | Tracking and monitoring |
| Erdeniz et al. (2023), Oyebode and Orji (2023), Schnall et al. (2015), Tsiakas, Barakova, Khan, and Markopoulos (2020) | Autonomy |
| Boontarig, Quirchmayr, Chutimasakul, and Papasratorn (2014), Buzcu et al. (2023), Cabrita et al. (2018, 2018), Duwaraka Yoganathan (2013), Henkemans et al. (2015), Lee et al. (2012), Schnall et al. (2015) | Persuasive technology |
| Duwaraka Yoganathan (2013), Henkemans et al. (2015), Orji and Moffatt (2016), Schnall et al. (2015), Wang, Wu, Lange, Fadhil, and Reiterer (2018), Wiafe and Nakata (2010) | Health and wellness promotion |
| Lehto (2012), Oinas-Kukkonen and Harjumaa (2009), Wiafe and Nakata (2010) | Accessibility |
| Ananthanarayan and Siek (2012), Jalil (2013), Lepri, Oliver, Letouzé, Pentland, and Vinck (2018), Wiafe and Nakata (2010) | Privacy |

*SRQ7 — functionalities realized*

The computational persuasion functionalities implemented are numerous and often follow or implement specific behavior change models studied in psychology. Table 3 reports the elicited mapping, showing that Persuasive System Design (PSD) (Oinas-Kukkonen & Harjumaa, 2009) and Fogg's behavioral models (Fogg, 2002) are the most employed. Moreover, it is worth mentioning the nudge theory. It is rooted in behavioral economics and psychology and proposes to influence human behavior through indirect cues and presentation of choices. This approach employs heuristics to steer patients toward beneficial options. By framing information or highlighting specific options in an easy-to-digest form, *nudges* aim to promote positive outcomes without altering a user's individual autonomy (Erdeniz et al., 2023; Tsiakas et al., 2020). Fogg's model has been designed with the idea that the users who have high ability but low motivation need to be prioritized so that they cross the behavior activation threshold (Fogg, 2002). The aim was to modify the habits or to improve the health of the individual by monitoring their behavior. Fogg's model showed in various studies a positive impact on behavior change and maintaining health and well-being (Boontarig

**Table 2**
Outcomes of primary studies.

| Studies | Outcome |
|---|---|
| Alahäivälä and Oinas-Kukkonen (2016), Almonani, Husain, San, Almomani, and Al-Betar (2014), Coorey et al. (2019), Fritz, Huang, Murphy, and Zimmermann (2014), Henkemans et al. (2015), Jalil (2013), Kelders, Kok, Ossebaard, and Gemert-Pijnen (2012), Lee et al. (2012), Orji and Moffatt (2016), Yoganathan and Sangaralingam (2015) | Positive. |
| Asbjørnsen et al. (2020), Bartlett et al. (2017), Cabrita et al. (2018), Chatterjee et al. (2012), Dominic et al. (2013), Gemert-Pijnen, Kelders, Jong, and Oinas-Kukkonen (2018), Jalil and Orji (2016), Lee et al. (2012), Lentferink et al. (2017), Matthews, Win, Oinas-Kukkonen, and Freeman (2016), Mylonopoulou (2018), Oyibo (2016), Oyibo and Vassileva (2020), Schnall et al. (2015), Srisawangwong and Kasemvilas (2014), Tian, Risha, Ahmed, Narayanan, and Biehl (2021), Tikka and Oinas-Kukkonen (2019), Wang et al. (2018), Wiafe and Nakata (2010) | Partially positive. |
| Duwaraka Yoganathan (2013), Lee et al. (2012), Tsvyatkova (2013) | Negative or not applicable. |

**Table 3**
Implemented/associated persuasion theories.

| Studies | Theories |
|---|---|
| Ainsworth (2012), Alahäivälä and Oinas-Kukkonen (2016), Ananthanarayan and Siek (2012), Asbjørnsen et al. (2020), Bartlett et al. (2017), Blom and Hänninen (2012), Cabrita et al. (2018), Coorey et al. (2019), Gemert-Pijnen et al. (2018), Henkemans et al. (2015), Jalil and Orji (2016), Kelders et al. (2012), Kim et al. (2019), Matthews et al. (2016), Mylonopoulou (2018), Oyibo (2016), Purpura et al. (2011), Tikka and Oinas-Kukkonen (2019), Wang et al. (2018) | The Persuasive Systems Design (PSD) |
| Ainsworth (2012), Almonani et al. (2014), Boontarig et al. (2014), Dominic et al. (2013), Duwaraka Yoganathan (2013), Fritz et al. (2014), Jalil (2013), Kueker, Koopman, McElroy, and Moore (2012), Lee et al. (2012, 2012), Lee, Kiesler, and Forlizzi (2011), Mylonopoulou (2018), Oyibo (2016), Schnall et al. (2015), Srisawangwong and Kasemvilas (2014), Tian et al. (2021), Tsvyatkova (2013), Wiafe and Nakata (2010), Yoganathan and Sangaralingam (2015) | Fogg's behavioral models |
| Duwaraka Yoganathan (2013), Yoganathan and Sangaralingam (2015) | Nudge theory |
| Erdeniz et al. (2023), Tsiakas et al. (2020) | Social cognitive theory (SCT) |
| Schnall et al. (2015) | Self-determination theory (SDT) |
| Tikka and Oinas-Kukkonen (2019) | Transformative learning theory |
| Tikka and Oinas-Kukkonen (2019) | Behavior changes support system (BCSS) |
| Kueker et al. (2012) | Transtheoretical behavior change model |
| Pinzon and Iyengar (2012) | Primary Persuasive Technology (PPT) |
| Boontarig et al. (2014) | ICT service design for senior citizen |
| Khalil and Abdallah (2013) | Theory of reasoned action (TRA) |
| Khalil and Abdallah (2013) | Theory of planned behavior (TPB) |
| Chatterjee et al. (2012) | Persuasive sensing |
| Oyibo and Vassileva (2020) | Persuasive Technology Acceptance Model |
| Alahäivälä and Oinas-Kukkonen (2016) | Health behavior change support systems |
| Asbjørnsen et al. (2020) | Behavior change techniques (BCTs) |
| Jalil and Orji (2016) | Not specified |

et al., 2014). PSD has been used in numerous studies to change user behavior in the health field, and the results have been conclusive. PSD was shown to improve participants' adherence to interventions and impact positively mental health outcomes (Gemert-Pijnen et al., 2018; Purpura, Schwanda, Williams, Stubler, & Sengers, 2011; Wang et al., 2018). However, some aspects of PSD appear to be more effective than others (e.g., social support, sharing, and comparison). Based on the selected studies and their analysis, it is difficult at this point to suggest which theory performs better than the other. Besides acknowledging the distribution of their adoption, the lack of pragmatic analysis does not allow us to conclude their effectiveness.

*SRQ8 — technology characterization*

The technological infrastructures identified in the computational persuasion domain are classified into seven areas (see Fig. 6). The top three technologies employed by these systems were mobiles-related (32.5%), Web (24.4%), and sensors or wearable-related (19.8%). The smartphone market has boomed and is also increasingly being adopted by older individuals. Smartphone and mobile OS producers engage in harsh competition, often due to their industrial philosophies being in stark contrast. Nevertheless, despite the diversity of these sophisticated technologies, they are easily accessible, (to a certain extent) affordable, and *easy to use*. Thus, several CPT leverage Mobile-related technologies (i.e., their sophisticated sensors and ease of sharing content over social media). However, Yoganathan and Sangaralingam (2015) related that the time spent with health and fitness apps is low with respect to the overall screen time, arguing the need for a deeper focus on improving user adherence. Fritz et al. (2014) reported in their results that participants had fully integrated the devices into their daily activities, only taking them off when they went to sleep and they noted a direct impact of the device on their activities. Other studies also reported that the devices enabled them to maintain this behavioral change over the long term (e.g., walking more). Using gaming/gamification as support to persuasion technologies is mainly reported in studies with scenarios based on wellness or children's health, as the "mobile game approach to preventing childhood obesity (Almonani et al., 2014). Moreover, the primary studies emphasize that the effectiveness of persuasion technologies varies according to the health area, the context, and the people targeted. Indeed, technologies and their intended use are widely heterogeneous and require different digital skills. Moreover, older individuals and children with chronic illnesses do not have the same capacity to use the technology as a healthy person who uses the application for wellness (Srisawangwong & Kasemvilas, 2014). Therefore, interfaces and procedures cannot be the same.

*SRQ9 — explanations in CPT*

Explanation generation is essential in CP systems to address ethical, social, and practical challenges. Transparency can be achieved through verbalization leveraging explanations in their many forms (e.g., visual, textual, hybrid, etc.). Explainable AI and recent human-center techniques (Anjomshoae, Najjar, Calvaresi, & Främling, 2019)

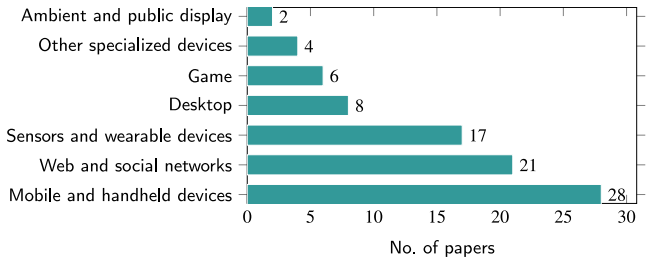



**Fig. 6.** Number of papers per technology type.

enable the users to gain insights into the decision-making processes, enhancing their understanding and fostering trust (Calvaresi et al., 2023; Calvaresi, Ciatto, et al., 2021; Graziani et al., 2023). The system facilitates autonomy as users become active listeners in a given setting, aligning advice with their preferences and values (Carli & Calvaresi, 2023). These techniques promote user acceptance and ethical deployment of computation persuasive systems in various domains. Cemiloglu, Arden-Close, Hodge, and Ali (2023) propose an explanation-based persuasive system for online gambling where they sought to increase the likeability of the system while allowing users to make more informed decisions via visual cards bearing texts, where cards would inform the gamblers of the techniques used by gambling sites. A theoretical test bed proposed by Tsiakas et al. (2020) utilize explanations in the form of encouraging words (e.g., *You did the given task perfectly before, that is why I think you can do it!* and correction nudges in *the right direction* to promote self-regulated learning for children assisted by AI, while the children complete goal-oriented tasks designed by the domain experts. Explanations can be in the visual form as well as the textual form. Toward this end, Sebastian, George, and Jackson (2023) utilize pre-defined visual explanations based on medicinal advertisements with varying types of texts. Similarly, Cemiloglu et al. (2023) exploit a similar style of visuals from real-life online gambling platforms to improve their textual explanations. Azazi, Richards, and Bilgin (2022) follow the visual explanation intuition by generating a 3D-modeled agent that resembles a human advisor. The virtual assistant utilizes explanations to build rapport with the users while trying to push them toward their goals, such as managing stress while the students are studying. Literature in this field sought to improve some areas of health care. For instance, a novel food meal planning system was proposed by Dragoni, Donadello, and Eccher (2020) to nudge users to guide them to follow a Mediterranean meal, which is assumed to be healthier generally. The explanations are textual and generated in three steps: first, a feedback (e.g., *Today you have drunk too much fruit juice*), then, an argument (e.g., *Do you know that sweet beverages contain a lot of sugars that can cause diabetes?*), then finally, a corrective suggestion (e.g., *Next time try with a fresh fruit*). Sassoon, Kökciyan, Sklar, and Parsons (2019) propose a wellness consultation framework via explanations. The framework is comprised of goal-oriented dialogue options within the domain of wellness consultation. For instance, a given user may consult the wellness agent to manage pain or lose weight, for which the agent should be able to answer why, give a counteroffer, and assert its own perspective. Explanations are also used to improve the convincibility of a nutrition virtual coach communicating with the user in the form of negotiation, where the system may provide explanations to explain its reasoning for a recommendation from a few template explanations retroactively (e.g., if a recipe has a good amount of protein, the system advocates for the recipe after the recommendation has been made: *I recommended you this recipe because it has got a high amount of protein*) (Buzcu et al., 2023). Additionally, textual explanations can be formulated according to the *nudge theory*, where the users are expected to make small changes in life toward a general improvement in their health. Practically, the explanations can follow a different form of the nudge principle (Erdeniz et al., 2023) while trying to persuade the user to follow the suggested decision support.

**Table 4**
Types of explanations.

| Type of Explanations | Studies |
|---|---|
| Text | Buzcu et al. (2023), Dragoni et al. (2020), Sassoon et al. (2019), Tsiakas et al. (2020) |
| Visual | Cemiloglu et al. (2023), Sebastian et al. (2023) |
| Virtual Avatar | Azazi et al. (2022) |

The evaluation of the generated explanations are generally personal, yet it may involve common dimensions among studies such as clarity, trust, relevance, and comprehensibility. The satisfaction with, and eventually the acceptance of, these explanations are significantly influenced by the perceived qualities (Hulstijn, Tchappi, Najjar, & Aydoğan, 2023).

Finally, on the one hand, from an ethical perspective, explanations assure users of the system's accountability and compliance with regulations while promising improvement toward a given user goal. On the other hand, there are potentially harmful effects should the trust instilled by the system in systems that exploit these explanation-based methods be misleading (e.g., improperly advising medical patients) either accidentally or intentionally (Carli et al., 2022; Schoenherr, 2021) (see Table 4).

*SRQ10 — strengths of the primary studies*

Table 5 summarizes the CPT's advantages stated by the primary studies. Among the first to catch our attention, it is worth mentioning the improvement of health and wellness (19 studies) and the enhancement of individuals' social engagement (mainly via sharing experiences and comparing results with other users — 20 studies). Studies such as Bartlett et al. (2017) have observed that participants living alone felt encouraged by the virtual coach and were able to enhance their mobility. Self-monitoring and self-efficacy assessment (when not overwhelming) have also been reported as empowering by 17 studies.

Some CPT have achieved positive results allowing the customization of the applications according to the user characteristics and needs and enabling the users to provide feedback (11 studies). It should be noticed, however, that feedback such as "punishment" was reported to have a negative effect on user motivation (Orji & Moffatt, 2016). Providing medically sound audio and video stimulation, possibly equipped by explanation (i.e., XAI) in the form of reminders and encouraging persuasive messages, has also contributed positively to raising the CPT credibility (Alahäivälä & Oinas-Kukkonen, 2016; Buzcu et al., 2023; Chatterjee et al., 2012; Schnall et al., 2015).

*SRQ11 — limitations of the primary studies*

Table 6 lists the limitations elicited from the primary studies. In some instances, results showed decreased effectiveness in the persuasion process. This is referred to as a direct consequence of the lack of customization, adaption, and personalization of the CTP to the users and their context. For instance, children followed the interventions, but adherence was strictly limited to the interactive games in a virtual environment (Dominic et al., 2013). Furthermore, such a study conveyed that the lack of people's awareness (crucial requirement) of PCT is a relevant co-factor, leading to a lack of motivation and, eventually, technological abandon. Another example is provided by Wang et al. (2018), which highlights the importance of involving users in the development (and setup) of the CPT so that they can pass on their experience. Several studies have been tested on small sample sizes, which did not allow generalizable recommendations. Eight studies have highlighted the difficulties of implementing computational persuasion methods,

**Table 5**
Strengths and advantages of primary studies.

| Studies | Advantages |
|---|---|
| Alahäivälä and Oinas-Kukkonen (2016), Almonani et al. (2014), Ananthanarayan and Siek (2012), Boontarig et al. (2014), Cabrita et al. (2018), Coorey et al. (2019), Dominic et al. (2013), Duwaraka Yoganathan (2013), Fritz et al. (2014), Gemert-Pijnen et al. (2018), Henkemans et al. (2015), Jalil (2013), Jalil and Orji (2016), Lee et al. (2011), Mylonopoulou (2018), Orji and Moffatt (2016), Schnall et al. (2015), Wang et al. (2018), Yoganathan and Sangaralingam (2015) | Health and wellness promotion |
| Ainsworth (2012), Almonani et al. (2014), Ananthanarayan and Siek (2012), Bartlett et al. (2017), Boontarig et al. (2014), Cabrita et al. (2018), Dominic et al. (2013), Duwaraka Yoganathan (2013), Fritz et al. (2014), Gemert-Pijnen et al. (2018), Kelders et al. (2012), Khalil and Abdallah (2013), Kim et al. (2019), Kueker et al. (2012), Lee et al. (2012), Mylonopoulou (2018), Orji and Moffatt (2016), Oyibo (2016), Schnall et al. (2015), Yoganathan and Sangaralingam (2015) | Social and health care support, sharing and comparison |
| Ananthanarayan and Siek (2012), Asbjørnsen et al. (2020), Bartlett et al. (2017), Boontarig et al. (2014), Chatterjee et al. (2012), Coorey et al. (2019), Duwaraka Yoganathan (2013), Fritz et al. (2014), Gemert-Pijnen et al. (2018), Jalil (2013), Matthews et al. (2016), Oyibo (2016), Oyibo and Vassileva (2020), Purpura et al. (2011), Schnall et al. (2015), Srisawangwong and Kasemvilas (2014), Yoganathan and Sangaralingam (2015) | Self-efficacy assessment, self-monitoring, perceived value |
| Asbjørnsen et al. (2020), Bartlett et al. (2017), Coorey et al. (2019), Dominic et al. (2013), Fritz et al. (2014), Jalil and Orji (2016), Khalil and Abdallah (2013), Kim et al. (2019), Lentferink et al. (2017), Matthews et al. (2016), Orji and Moffatt (2016), Purpura et al. (2011), Schnall et al. (2015), Wiafe and Nakata (2010), Yoganathan and Sangaralingam (2015) | Tracking and monitoring awareness |
| Ainsworth (2012), Asbjørnsen et al. (2020), Cabrita et al. (2018), Dominic et al. (2013), Duwaraka Yoganathan (2013), Fritz et al. (2014), Gemert-Pijnen et al. (2018), Jalil (2013), Jalil and Orji (2016), Kelders et al. (2012), Khalil and Abdallah (2013), Kim et al. (2019), Lentferink et al. (2017), Pinzon and Iyengar (2012), Srisawangwong and Kasemvilas (2014) | broad technological outreach (Mobile phone, computer, devices, smart home) |
| Asbjørnsen et al. (2020), Coorey et al. (2019), Henkemans et al. (2015), Jalil (2013), Jalil and Orji (2016), Kelders et al. (2012), Kim et al. (2019), Lee et al. (2012), Lentferink et al. (2017), Purpura et al. (2011), Tikka and Oinas-Kukkonen (2019) | Configuration tailoring and personalization |
| Alahäivälä and Oinas-Kukkonen (2016), Asbjørnsen et al. (2020), Chatterjee et al. (2012), Dominic et al. (2013), Fritz et al. (2014), Gemert-Pijnen et al. (2018), Jalil and Orji (2016), Lee et al. (2012), Matthews et al. (2016), Orji and Moffatt (2016), Yoganathan and Sangaralingam (2015) | Feedback |
| Almonani et al. (2014), Bartlett et al. (2017), Cabrita et al. (2018), Coorey et al. (2019), Dominic et al. (2013), Jalil (2013), Khalil and Abdallah (2013), Kueker et al. (2012), Orji and Moffatt (2016) | Visual, audio, simulation |
| Coorey et al. (2019), Gemert-Pijnen et al. (2018), Jalil (2013), Kelders et al. (2012), Lentferink et al. (2017), Matthews et al. (2016), Mylonopoulou (2018), Wang et al. (2018) | Persuasive messages, reminders |
| Ainsworth (2012), Boontarig et al. (2014), Gemert-Pijnen et al. (2018), Kueker et al. (2012), Mylonopoulou (2018), Yoganathan and Sangaralingam (2015) | Design, and environment |
| Boontarig et al. (2014), Gemert-Pijnen et al. (2018), Jalil and Orji (2016), Kim et al. (2019), Srisawangwong and Kasemvilas (2014), Wiafe and Nakata (2010) | Cost, accessibility, coverage |
| Asbjørnsen et al. (2020), Coorey et al. (2019), Fritz et al. (2014), Khalil and Abdallah (2013), Lentferink et al. (2017) | Clarity of goal and objectives |
| Alahäivälä and Oinas-Kukkonen (2016), Almonani et al. (2014), Ananthanarayan and Siek (2012), Dominic et al. (2013) | Gamification |
| Boontarig et al. (2014), Gemert-Pijnen et al. (2018), Matthews et al. (2016) | Credibility |
| Jalil (2013), Lee et al. (2011) | Tunneling |
| Jalil and Orji (2016) | Not specified |

while others reported limitations in terms of effective integration of behavior theories and practice in their design (Orji & Moffatt, 2016). Finally, user knowledge is an obstacle to effective use (5 studies), and the lack of privacy and transparency are other significant barriers (3 studies). Finally, the engagement of the user is promised to be boosted by explainable technologies. However, often the produced explanations are not satisfactory. Indeed, although correct, users have provided feedback such as "how this applies to me?", "I do not understand why", and "but what if ..".. This proves that the solutions provided so far do not match the human causal reasoning (Sloman & Fernbach, 2017).

*SRQ12 — proposed solutions*

Table 7 shows the solutions identified within the primary studies to overcome the identified limitations. Among the most pinpointed there is the "technology adaption to the user and his context". Along this line, Orji and Moffatt (2016) observed a need to target diverse demographics such as older adults and children. The recommendation of Dominic et al. (2013) is to know and handle preferences to make the technology attractive and engaging (especially in the long run), and to select/personalize the technologies and applications according to users' desires and objectives. Another suggestion is to integrate the users further while developing the application's design. The fourth proposal highlights the possibility for users to follow their activity and see their evolution about the efforts made. Finally, new technologies such as smartphones, the Web, and sensors can improve effectiveness but should never be overwhelming (just to satisfy data-eager scientists). Finally, to overcome the lack of personalization of the systems' doing (i.e., improving the explanations), more effort should be put into morphing/translating the systems' inner rules into clear and easy-to-understand statements (even less detailed if needed) (Buzcu et al., 2023; Contreras et al., 2022)

*SRQ13 — future challenges stated in the primary studies*

All the elaborated studies are aligned in the definition of the future challenges and envisioned future work. In particular, they suggest investigating how persuasion technologies can be used to engage and meet, possibly, the evolving needs of a given population and deeply understand the long-term effects of specific persuasive models for a target population. They also proposed to conduct a study identifying interactions between individual strategies and computational persuasion successful outcomes. Moreover, studies with larger populations and a direct assessment of user adherence are needed. Finally, recalling that

**Table 6**
Limitations and drawbacks.

| Studies | Drawbacks |
|---|---|
| Ainsworth (2012), Alahäivälä and Oinas-Kukkonen (2016), Blom and Hänninen (2012), Boontarig et al. (2014), Coorey et al. (2019), Dominic et al. (2013), Gemert-Pijnen et al. (2018), Henkemans et al. (2015), Kelders et al. (2012), Kim et al. (2019), Lentferink et al. (2017), Orji and Moffatt (2016), Oyibo (2016), Oyibo and Vassileva (2020), Srisawangwong and Kasemvilas (2014) | Lack of adaptation to user context |
| Alahäivälä and Oinas-Kukkonen (2016), Asbjørnsen et al. (2020), Chatterjee et al. (2012), Dominic et al. (2013), Gemert-Pijnen et al. (2018), Jalil and Orji (2016), Lee et al. (2011), Oyibo and Vassileva (2020), Tian et al. (2021) | Less effective, and not representative |
| Alahäivälä and Oinas-Kukkonen (2016), Cabrita et al. (2018), Chatterjee et al. (2012), Coorey et al. (2019), Fritz et al. (2014), Jalil (2013), Lentferink et al. (2017), Srisawangwong and Kasemvilas (2014) | Difficulties in PT implementation. |
| Ananthanarayan and Siek (2012), Matthews et al. (2016), Srisawangwong and Kasemvilas (2014), Tikka and Oinas-Kukkonen (2019), Wang et al. (2018) | Limited research |
| Ananthanarayan and Siek (2012), Chatterjee et al. (2012), Dominic et al. (2013), Lee et al. (2012), Oyibo (2016) | Lack of knowledge |
| Chatterjee et al. (2012), Coorey et al. (2019), Jalil (2013), Lentferink et al. (2017) | Lack of appeal/motivation |
| Kueker et al. (2012), Pinzon and Iyengar (2012), Tian et al. (2021), Wiafe and Nakata (2010) | Lack of privacy, transparency |
| Ananthanarayan and Siek (2012), Matthews et al. (2016), Mylonopoulou (2018) | Competition |
| Asbjørnsen et al. (2020), Fritz et al. (2014), Tian et al. (2021) | Lack of long-term goal settings |
| Almonani et al. (2014), Blom and Hänninen (2012), Lee et al. (2012) | User experience not considered |
| Henkemans et al. (2015), Tian et al. (2021) | Cost |
| Lentferink et al. (2017), Srisawangwong and Kasemvilas (2014) | Lack of social support |
| Ainsworth (2012), Lee et al. (2012) | Difficulty to understand the technology, and PT not desirable |
| Chatterjee et al. (2012), Khalil and Abdallah (2013) | Technology problems |
| Gemert-Pijnen et al. (2018) | Limitations of design |
| Lentferink et al. (2017) | Limited credibility |
| Jalil (2013) | Timing issues |
| Spanakis et al. (2014) | Insufficient multi/interdisciplinary interaction. |
| Wiafe and Nakata (2010) | Legal issues |
| Alahäivälä and Oinas-Kukkonen (2016) | Exposed to commercial messages |

**Table 7**
Proposed solutions of the primary studies.

| Studies | Solutions |
|---|---|
| Ainsworth (2012), Ananthanarayan and Siek (2012), Cabrita et al. (2018), Gemert-Pijnen et al. (2018), Kueker et al. (2012), Oyibo and Vassileva (2020), Pinzon and Iyengar (2012), Spanakis et al. (2014), Tian et al. (2021), Yoganathan and Sangaralingam (2015) | Adapt to the user's context |
| Chatterjee et al. (2012), Jalil and Orji (2016), Lee et al. (2012), Lentferink et al. (2017), Oyibo (2016) | Tailoring and personalization |
| Henkemans et al. (2015), Orji and Moffatt (2016), Purpura et al. (2011), Schnall et al. (2015), Wiafe and Nakata (2010) | Adopt a participatory design approach |
| Ananthanarayan and Siek (2012), Asbjørnsen et al. (2020), Khalil and Abdallah (2013) | Tracking and monitoring |
| Alahäivälä and Oinas-Kukkonen (2016), Dominic et al. (2013), Khalil and Abdallah (2013) | Modern technology, and simulation |
| Asbjørnsen et al. (2020), Matthews et al. (2016) | More appealing design |
| Bartlett et al. (2017), Kelders et al. (2012) | Dialogue support |
| Cabrita et al. (2018), Gemert-Pijnen et al. (2018) | Machine learning personalization |
| Asbjørnsen et al. (2020), Srisawangwong and Kasemvilas (2014) | Ability, and availability |
| Dominic et al. (2013), Khalil and Abdallah (2013) | Social support |
| Ananthanarayan and Siek (2012), Jalil (2013) | Privacy-preserving features |
| Coorey et al. (2019), Dominic et al. (2013) | Reminders |
| Kelders et al. (2012) | Tunneling |
| Oyibo and Vassileva (2020) | Persuasive value |
| Bartlett et al. (2017) | Primary task support |
| Dominic et al. (2013) | Gamification |

CPT comprise several components/modules, dedicated studies should target them singularly to understand their contribution and tuning. By doing so, it can pave the way to design and realize modular ecosystems well-suiting heterogeneous populations (Kelders et al., 2012). Human–machine interactions should also be improved to sound more natural and engaging. Communication with peers, family, and caregivers has proven to be a key element in the effectiveness of CPT and deserves more investigation and development. The security of users' data is also a challenge getting increasingly complex (let us consider the centralizing mechanisms of current ML predictors). XAI technologies have brought a plethora of techniques to foster trust and transparency (Anjomshoae et al., 2019; Graziani et al., 2023). Yet, their outcomes are still far from satisfying the users' backgrounds/knowledge and, as of today, still result overwhelming (Mualla et al., 2022).

Overall, developers of CPT must consider the characteristics of users who want accessible and easy-to-use technologies that Oinas-Kukkonen and Harjumaa (2009), didactic, allow them to visualize their efforts

**Table 8**
Future challenges in primary studies.

| Studies | Features envisioned and future challenges |
|---|---|
| Alahäivälä and Oinas-Kukkonen (2016), Ananthanarayan and Siek (2012), Coorey et al. (2019), Gemert-Pijnen et al. (2018), Jalil (2013), Kueker et al. (2012), Lee et al. (2012), Orji and Moffatt (2016), Oyibo (2016), Oyibo and Vassileva (2020), Spanakis et al. (2014), Srisawangwong and Kasemvilas (2014), Wiafe and Nakata (2010) | Dynamic adaptation to user's context. |
| Bartlett et al. (2017), Coorey et al. (2019), Dominic et al. (2013), Jalil and Orji (2016), Khalil and Abdallah (2013), Mylonopoulou (2018) | Dialogue, social support, and sharing and comparison |
| Chatterjee et al. (2012), Jalil and Orji (2016), Lee et al. (2012), Lentferink et al. (2017), Oyibo (2016) | Tailoring and personalization |
| Coorey et al. (2019), Henkemans et al. (2015), Wiafe and Nakata (2010) | Privacy, security, and transparency |
| Dominic et al. (2013), Lee et al. (2012), Wang et al. (2018) | Modern technology, easy-to-use functionalities |
| Alahäivälä and Oinas-Kukkonen (2016), Coorey et al. (2019), Henkemans et al. (2015) | Slicker appealing design |
| Ananthanarayan and Siek (2012), Oyibo (2016) | Adopt a participatory design approach |
| Mylonopoulou (2018) | Tracking and monitoring |
| Chatterjee et al. (2012) | Machine learning and data-driven personalization |
| Purpura et al. (2011) | Ethics-compliance |
| Khalil and Abdallah (2013) | Feedback-loops. |
| Bartlett et al. (2017) | Primary task support. |

and use modern media that allow them to communicate and share their performance while respecting ethics and privacy (Calvaresi et al., 2023) (see Table 8).

*SRQ14 - (possible) adverse effects of persuasion*

The element of persuasion in AI encompasses several possible risks for the end user. Such risks are more peculiar (or intensify their severity) for applications operating in the healthcare domain. The main risk linked to persuasion is the difficulty that still exists in the literature in drawing a clear line between persuasive, manipulative, and coercive techniques (Carli et al., 2022). Overall, we can briefly identify: (i) persuasion as the dynamic that expands the basket of choices available to the subject, supporting some as preferable and providing a motivation (Rudinow, 1978); (ii) coercion as the reverse mechanism, which restricts the basket of choices by identifying some of them as not practicable or forbidden (Leonard, Thaler, & Sunstein, 2008); (iii) manipulation, i.e., the dynamic that leads to a distortion of the subjects' perception, their decision-making powers, and their needs, in a way that goes beyond rationality and is therefore difficult to identify (Susser, Roessler, & Nissenbaum, 2019). The healthcare field, then, offers more peculiar critical profiles. Very often these applications are used in place – or sometimes even against the advice – of a human specialist. Therefore, their functionality is not monitored concerning the safety of the provided recommendations, nor about the appropriateness of the purpose set by the user. This also depends on the fact that, often, the data required by the AI system to pursue its original purpose are not exhaustive or at least sufficient to cover the complexity of some pathological situations that often affect several aspects simultaneously. Then, if we consider the possibility of addiction on the use of the application or the development of pathologies that affect the psyche, the situation becomes even more challenging to manage. An example could be a person who develops a dependence on fitness (beyond the limits of health) or who develops an eating disorder (including orthorexia). The chances for the system to detect similar – initial or occurred – pathological situations are very limited. At the same time, the individuals have no way to protect themselves or to disclose their condition. Indeed, they are often in the first place and very long unaware of needing a balanced and physiological approach to health/weight/appearance.

*SRQ15 — legal implications*

From a legal perspective, CPTs present a two-folded problematic profile. The first one depends on the difficulty that legal and technical experts have in clearly identifying the nature of the object of analysis (Galanos, 2018). The second one concerns applications that can appeal to the psychological and cognitive sphere of the user, giving rise to immaterial and, more specifically, psychological damages, which are still difficult to prove and unambiguously addressed by the law (Echeburúa, Corral, & Amor, 2003). This scenario could be further decomposed into the following subcategories of challenges:

*The problem of (general) definition*

The term CPT covers various AI applications with different purposes, characteristics, and technical elements (Fogg, 2009). The fact that this discussion focuses on those AI systems that exploit persuasion in the healthcare domain does not, in itself, make it possible to narrow down the field of application particularly. This uncertainty on the definition (Nordström, 2022) stems from a broader problem, which originates from the question of what is meant by AI, more generally (Roberge, Senneville, & Morin, 2020; Wang, 2020). This aspect is crucial from a legal point of view, especially from the perspective of regulating new technologies. Indeed, the law needs an unambiguous identification of its subject matter to operate correctly and be binding (Scherer, 2015). It has been argued that this is actually a false issue, for the regulatory process would be able to capture, handle, and somehow solve the vagueness in itself (Danaher, 2021). Nevertheless, the substantial problem that this defining and descriptive imprecision creates is related not so much – or at least not only – to the nature of AI or CPTs *per se*. The biggest side effect is an ambiguity in the software goals and means (Elish & Boyd, 2018).

*(Potential) variety of applications*

Regulators have apparent difficulties fully circumscribing the nature, extent, and target population and any adverse effects of CPT due to a latent lack of a clear scope. Being able to distinctly identify these aspects is essential not only for regulating the dynamic itself but also for provisioning mitigation tools and evaluating the relevance/limitation of AI tools in a given context. Moreover, the difficulty of unambiguously identifying the device under analysis and its specific field of application makes it difficult to understand whether there is a legal framework within which it can be traced, to what extent, with what precautions, and with what expected effects. As a consequence, intercepting existing regulatory gaps is challenging. Thus, we could face a double risk: under-regulating—thinking that there are no legal categories to which the case in question can be traced—or over-regulating—causing an unnecessary and counterproductive multiplication of normative instruments.

*Variety of health conditions*

Where it is difficult to clearly define the object of regulation and its possible applications, the law inevitably faces an additional problem: risk (Nair & Howlett, 2017). This is true in terms of (i) identification, (ii) circumscription, (iii) assessment, and (iv) prevention of risks associated with a given system (Tarling & Burrows, 2004). This basic legal assumption appears even more valid in the case of technologies that employ persuasion to change a health-related human behavior or habit. In fact, a general objective of improving one's athletic performance or fitness cannot be put on the same level as the need to lose weight due to a specific medical condition or under the prescription of a specialist. At the same time, the desire to lose weight from a state of semi-healthiness does not have the same implications and impact – both organically and emotionally – that losing weight would have for a person in a state of severe obesity. Furthermore, using an application to pursue a single health goal is not comparable to doing so to address a chronic disease or syndrome (which, by definition, brings together concomitant diseases or possible symptoms). This generates two main issues: the wide range and diversity of risks that may be produced, and the difficulty in predicting them in detail.

Both of these challenges relate to the fact that the risks that we can foresee are just those that have already occurred in similar or comparable circumstances or those that can be reconstructed from direct or indirect knowledge that has been consolidated in the past. This does not exclude the possibility that new technologies confront us with consequences and challenges that could not have been contemplated before (Rosenberg, 1995). This argument is in line with the very well-known "black box phenomenon". Stressing it here does not mean corroborating those theories which identify in it the rising of sentient/intelligent machines. However, it helps to underline that AI systems introduce a level of "randomness and uncertainty" (Renda et al., 2019) that should be considered peculiar. Some consider this issue easy to solve with an ex-post mechanism based on explainability and the field of Explainable AI (henceforth XAI) (Biran & Cotton, 2017; Lepri et al., 2018). Nevertheless, many doubts still persist in the literature as to whether AI can be explained and, even more so, whether providing an explanation can be an effective and efficient harm reduction/containment tool (Carli et al., 2022). It is beyond the scope of the present discussion to go into these issues in depth, but even assuming the benefit of explanation in the sense described above, the problem of the foreseeability of the impact would remain. In fact, if the explainability would work, it would work as an *ex-post* checking mechanism, not as an *ex-ante* tool to address safety.

*Classes of users*

As discussed above, even if we restrict the scope of analysis to AI systems used in the health domain, the variety of applications to be considered is vast. Consequently, the individuals exposed to them belong to potentially very different classes of users. Indeed, the range of age, initial health condition, familiarity with the technology, sensitivity to persuasion, and propensity to addiction is vast. Then, all these characteristics may intertwine and, at times, overlap. The fact that a user is elderly, for example, does not exclude that their health condition is generally stable and that they are endowed with a good level of resilience. This is possible even if, as a rough approximation, elderly persons tend to be identified as fragile subjects by default. Conversely, subjects in their thirties, with a severe physical disability and lacking social support figures, may appear more at risk in terms of manipulative effects, regardless of their technological literacy. These differences are not neutral for the legal system. They are expressions of how our inherent human vulnerability can manifest itself. As known, the liberal tradition gives some of these ways a different legal qualification. The figure of the legally incapacitated person, for instance, is subjected to different protections – and corresponding limitations – than the minor and still different than the ill/older adult who is not interdicted or incapacitated. Nonetheless, the law does not distinguish between an adult, considered legally capable, with good and stable health conditions, and one with a terminal disease, even if the psychological fragility could be highly different. At the same time, the law regulates in different ways (i) an adult, considered legally capable, (ii) an old user, (iii) a minor, even if all of them have good and stable health conditions. Depending on the nature and scope of the system, it could be not sensitive to such peculiar classifications- or the lack of them.

*Data protection and privacy issues*

The collection, use, and sometimes profiling of personal data is a central factor for persuasive systems, especially those involved in health-related behavioral change mechanisms. This is necessary to generate recommendations that are in line with the goals set by the user but also aligned with the user's general characteristics. In fact, if an individual has chronic conditions or current illnesses, they will receive recommendations calibrated to this background that will be different than those provided to someone who has the same goal but dissimilar baseline health conditions. Therefore, not only the persuasive system cannot disregard the user's personal data to perform the function for which it was designed, but also to ensure the very safety of the individual involved. However, the issue of data collection and management by means of AI systems, and consequently privacy, are among the most debated issues in doctrine and jurisprudence. This section is not intended to report a detailed overview of that debate, but only to highlight some of the controversial issues most relevant to the topic addressed in this discussion. The World Health Organization Director General has recently shared concern for the possible unethical data collection, cybersecurity threats, data biases, and consequence misinformation ((WHO)).

*Direct and indirect users*

Another important distinction among the user classes is the one between (i) direct users and (ii) indirect users. Those who have decided to use the application belong to the first category. Therefore, they have a personal interest in such usage, are looking for a personal benefit, have entered their data, and have consented to its analysis and processing. The second category is broader and includes those potentially impacted by some of the effects/implications/data disclosures related to the use of the system by the previous group. Nevertheless, they may not be aware of it – in whole or in part –, they may be aware of it but may not have seen or accepted any information on data processing or operation of the application; they may fall into categories that the law subjects to different protection from that of the direct user. An illustrative example of this second group can be found in the context of recommender systems that are required to collect data on the health status or susceptibility to certain diseases of users. In this case, the individuals who directly use the system and who have consented to the processing of their data will also be required to provide data regarding third parties to the interaction, even if the latter are not aware of this disclosure. Moreover, if the inquiry pertains to genetic diseases or familiarity with the development of certain conditions, whether in a general sense or more specifically within the female or male branch of the family, by combining this data with the user's first and last name and age, it may be possible not only to collect information about other people but also potentially to trace their presumed identity.

*SRQ16 — ethical implications*

There is limited literature on the ethics (ethics principles, ethical aspects, and ethical implications and concerns) related to CPT in eHealth. Indeed, while a great deal of ethics scholarship has focused on the phenomena of computational manipulation and coercion in general, less work has been developed on the ethics of CPT as solutions eliciting a voluntary change of behavior and attitude, following the definition provided by Fogg (2002), in the domain of healthcare. Even in the ethics scholarship on CPT in general, scholars mainly focus on how

CPT might deploy manipulative or coercive algorithmic techniques that undermine individuals, rather than on CPT as persuasive tools *per se*. While such concerns are relevant, we defer their consideration in the debate on the ethics of CPT to the Discussion Section (Section 4). This section narrows to specific ethical aspects and implications that can arise from CPT designed with the intentional benevolent goal of improving users' health and well-being according to their stated goal. Therefore, we set apart definitions of CPT provided by some scholars such as Berdichevsky and Neuenschwander (1999), Kampik, Nieves, and Lindgren (2018), who broad CPT definition as "an information system that proactively affects human behavior in or against the interests of its users" (p. 5), insofar as easily blurring the distinction between beneficial (or ethical) persuasion and harmful persuasion, deceit, or manipulation. Thus, we only consider studies on CPT intended as systems designed to induce people's voluntary health and behavioral change through both rational means (e.g., reasoning and argument) and non-rational means (e.g., peer pressure, restructuring of choice architecture, etc.) but always *in accordance with* the interests, goals, and expressed preferences on means of the subjects.

Our research identified three papers (Jacobs, 2020; Kip, Jong, Gemert-Pijnen, Sanderman, & Kelders, 2018; Rughiniș, Rughiniș, & Matei, 2015) that focus substantially or at least treat consistently the ethics of CPT in eHealth. In their analysis of CPT in eHealth, Gemert-Pijnen et al. (2018) devote a specific space to the ethical aspects of *persuasive eHealth technology*. Indeed, they stress ethics as a key component of CPT in eHealth, outlining manipulation and coercion as the dark side of persuasion and out of the scope in the positive-oriented approach endorsed by persuasive eHealth technology. Particularly, they outline at least four ethical issues that should be considered in the design of CPT in eHealth (Kip et al., 2018). The first ethical issue concerns the *responsibility* for the well-being of people using persuasive eHealth technology. On the one hand, CPT designers might hold a certain degree of responsibility for people's health self-management as they develop solutions that trigger specific health behavior changes toward goals they set (in accordance with users). On the other hand, especially in the case of CPT in eHealth, people choose to use them voluntarily. Hence, users might be considered fully responsible for their health when using eHealth CPT. However, especially in the health domain, not everyone might be able to deal with such a responsibility (especially vulnerable people). The second ethical issue the authors identify to be considered by CPT design is the impact of such systems on people's *autonomy*, that is, on their capacity and right to make their own choices based on their own values. Even without manipulating or coercing, CPT may limit individual autonomy, as they set the desired behavior, for instance, based on social norms and the related steps people should necessarily follow to achieve a certain goal. While people can agree with them, such steps might diverge from their deep and/or genuine desires, values, and interests, requiring a tradeoff in the light of a possible health benefit. Linked to this ethical issue, the third ethical implication related to CPT in eHealth outlined in Kip et al. (2018) concerns people's *self-control*. The authors outline how relying on CPT might make people increasingly dependent on persuasive technology, undermining, in the long run, their capacity to adopt a specific healthier behavior without the assistance of a certain CPT. Finally, the authors ask to consider *equity* by design and use of CPT in eHealth. Indeed, such solutions can make healthcare more accessible in many ways, but they can also hinder equity. In this regard, these technologies might reach only those individuals who already share those ideas and norms about the desired behavior such solutions are designed to promote. Furthermore, people with lower literacy skills may lack the ability to weigh arguments and, therefore, to truly provide consent to be persuaded toward a certain goal, resulting in being more vulnerable to CPT. The latter aspect opens a set of specific ethical considerations concerning the design of persuasive eHealth solutions for vulnerable people currently missing in this study. Jacobs (2020) addresses this issue with a first analysis of the ethical concerns related to CPT for health behavior change with specific attention to their design for *vulnerable people*. In this study (Jacobs, 2020), instances of persuasion induced by CPT imply the user's reasonable consent both to (i) the ends of persuasion (i.e., the target behavioral change) and to the (ii) means of persuasion (i.e., means deployed to achieve the target goal). Drawing on standard ethical theory, Jacobs underlines that persuasion is often evaluated as *ethical* if it is aligned with a person's individual goals, interests, and needs and does not thwart a person's autonomy. However, such issues deserve particular attention when it comes to designing CPT in a suitable way for vulnerable people. Indeed, CPT designers develop solutions often considering an idealized person or based on their own needs, interests, and experiences, therefore overlooking real-life contexts and oppressed agency conditions of those who differ from this conception, resulting in being inadequate or oppressive for such users. Furthermore, when vulnerability is considered, a common mistake to avoid is that of labeling and treating vulnerable people as all belonging to the same category of people who are "fragile and susceptible to wounding" (Jacobs, 2020): this results in stereotyping vulnerable individuals and problematically obscuring the diverse context-specific risks of specific groups at risk. The view of vulnerable people as needing extra attention and care can also be problematic for the adequate design of CPT, as it can boost unwanted and sometimes unfair paternalistic measures. In this context, a helpful conceptualization of vulnerability for the design of CPT is identified by the author in that provided by Mackenzie, Rogers, and Dodds (2014), as it detects some key dimensions of vulnerability that should be considered in the design of CPT. Such dimensions are: a) the wrong or harm a person is vulnerable to, which can be *dispositional* or *occurrent*; (b) the source of the vulnerability, which can be *inherent* (i.e., intrinsic to a person's psychophysical condition) and *situational* (context-sensitive, that is, caused and/or exacerbated by the social, political, economic, or environmental context); and (c) the safeguards that are needed in response.

A subset of situational sources of vulnerability is also evidenced, called *pathogenic vulnerabilities*, to refer to those situations where the solution for a specific harm paradoxically ends up worsening the harm itself or creating a new one. The author considers as an example MyFitnessPal: a CPT for calorie-counting and food tracking that is based on both peer pressure and goal achievement rewards. While the app is designed to help users maintain healthy weight goals, it has been shown to be largely used by the eating disorder population and to contribute to or exacerbate situational and inherent sources of harm by eliciting eating disorder triggers such as peer competition by comparison. The design of such a CPT is unsuitable for vulnerable people because it mainly considers an idealized user or is not sufficiently or adequately informed by its prospective users' diversified experiences, interests, and needs. To avoid such potential harm, a key ethical criterion for designing CPT for vulnerable people's health behavioral change is ensuring that the interests and needs of (vulnerable) users are properly taken into account. To do so, Jacobs suggests CPT designers involve and elicit the needs of their prospective users during the design phase while (a) taking users' real-life contexts into account and (b) providing adequate tools to support communication between stakeholders and designers on the values, needs, and interests important to the users (Pommeranz, Detweiler, Wiggers, & Jonker, 2012). Nevertheless, elicitation tools supporting a shared understanding of interests between stakeholders and designers are poorly explored. Another obstacle is that there is no consensus on what method works best for users to express their needs and interests – a particularly severe problem in vulnerable people who are less able to express or safeguard their needs and interests. Another key ethical consideration for designing CPT for vulnerable people concerns securing their autonomy, that is, their capacity to make choices and actions based on their values and beliefs. Jacobs suggests that a valid consent procedure for vulnerable people is needed to respect the autonomy of vulnerable people and protect them from instances of manipulation and coercion. The author distinguishes four

aspects of CPT to which the users should consent to ensure a valid consent procedure: (1) the *set goals* and targeted *behavioral outcomes* of a CPT; (2) *persuasive tools and strategies* to trigger a behavioral change; (3) the types of *user-CPT interaction* (e.g., messages, lights, sounds, etc.); (4) the use of *users' data* by the CPT company provider. Also, specific design requirements should be considered to make it easy for the user to give consent on all four aspects: (i) providing a limited amount of accurate and relevant information (to avoid overwhelming users with information); (ii) providing user-friendly ways to extend this amount of information; and (iii) easy ways of rescinding consent once given, ensuring that the person is not coerced, insofar as (a) expected outcomes of CPT are often difficult to foresee, and (b) a person can change over time (O'neill, 2017).

Finally, it is essential that the consent procedure is intelligible by the users and therefore consider by design various sources of vulnerabilities of diverse prospective users that can hinder such comprehension. Rughiniş et al. (2015) analyze a specific type of CPT in the field of eHealth from the ethical perspective: smoking cessation apps. The authors describe smoking cessation apps as instances of CPT issuing eloquent voices, which rely, for instance, on quantification and text advice to guide users in this difficult phase of their lives. In particular, their study focuses on a set of eHealth apps (currently available in the market) harnessing quantitative indicators of health and finance and/or a 'coach' offering advice for moments of craving a cigarette. Their ethical inquiry highlights a set of specific ethical issues related to such type of eHealth CPT. First, the most common concern highlighted is a more general one and concerns privacy and the use of personal information: to be used, indeed, smoking cessation apps require users' consent to various levels of access to their information, leading to possible data misuse as well as the risk of unwanted access through security breaches (Rughiniş et al., 2015). Second, to date, there is no process of authorization for health-related apps, which are treated similarly to other apps for entertainment, according to criteria such as proven impact or adherence to medical procedures, making it difficult for the user to find where there is medical expertise behind their design (Rughiniş et al., 2015). In this regard, the authors conclude that for apps for smoking cessations (similarly for those for panic disorders, alcohol-control, diabetes self-management, weight loss, and fitness), there is low compliance with clinical guidelines — with the exceptions for the European Commission (EC) apps.

Another issue concerns the transparency of commercial interests in smoking cessation apps (Rughiniş et al., 2015). Indeed, many of the free apps include unmarked advertisements, such as offering "E-cig coupons" without presenting this as a commercial interest, accompanied by encouragement [ad]vice, undermining transparency (e.g., "Having an electronic cigarette can help you during those tough times when you really want to smoke. (...) Just take a few pulls from the electronic cigarette to hold yourself off until the next scheduled smoke time"). Moreover, Rughiniş et al. (2015) devote space to personal autonomy, defined as one of the key ethical considerations in CPT in eHealth. In this regard, the authors identify layers of the app to work on to support users' autonomy. The first layer relates to the *control and actual involvement of users* in the app intervention and interaction (i.e., what degree of control do users have?). For instance, this entails whether users can choose when they want to see app-related information and advice. The second layer refers to *personalization*, namely, the extent to which users can communicate their preferences to customize the treatment; this would allow the users to direct their own behavior change as a form of self-persuasion (Spahn, 2012) The third layer concerns the extent to which such systems enhance the *knowledge* and information base on which users can base their decisions. For instance, smoking cessation apps provide much information concerning smoking risks and their evolution after cessation, on nicotine addiction, withdrawal symptoms, and indications for managing cravings. However, the accuracy (or truthfulness) of other information, such as numerical projections about risks decrease after smoking and other quantified estimates of health improvements, are still problematic, considering such apps cannot access individualized or fine-grained information of health data, raising the risk of deceitful persuasion (Guttman & Salmon, 2004). In this regard, Rughiniş et al. (2015) propose a few recommendations: apps should (i) inform the users about the sources of information that is presented to them; (ii) communicate that estimates are mostly at the aggregate level and include a degree of approximation; and (iii) remind users that messages refer not to "you" but, rather, to "people like you", where such likeness is determined based on input information. The fourth and fifth layers refer to *enhanced self-understanding and self-direction*: they refer to the messages that users receive concerning their agency, the understanding that shapes their actions, and the imputation of responsibility for various outcomes. It is worth highlighting that the implicit model of human action mainly coded into persuasive strategies and messages of such apps is that of 'mind over body', glorifying individual control over bodily reactions, with little if any external support. Such an individualistic model can result in relapses often framed as a user's failure. Here, the ethical issue outlined concerns how to support users by encouraging self-efficacy and avoiding moral recrimination for instances of relapse (Rughiniş et al., 2015). Linked to this, it is worth outlining the widespread moral, cognitive, and aesthetic portrayal of smoking that ground persuasive strategies in CPT and public health campaigns, often resulting in the stigmatization of smokers, and if this is morally acceptable. According to Rughiniş et al. (2015), some apps for smoking cessation use stigmatizing apps in their persuasion strategy. Examples of negative messages define smokers as wrongdoers (e.g., "Quitting smoking means: You will no longer hurt yourself and others"), aesthetic smoking as disgusting ("You've taken the first steps toward busting this disgusting habit"; "Smoking is a disgusting and stinking habit"), focus on bodily disfigurement ("Are you worried about your sex appeal? Studies have shown a clear link between smoking and impotence and reduced sexual pleasure. Fancy a cigarette?") or stress smoking as a stupid behavior ("When you haven't smoked for a month or more you will realize how stupid it was to spend all that money on an addiction that was literally killing you! Never again!").

Overall, most apps rely on framing smoking as a useless behavior, a sign of lack of will, irrationality, or disease; those promoted by EC tend instead to use messages that elicit self-observation and introspection (e.g., "It's normal to feel panic from time to time. Are you afraid you'll lose part of your identity? Rest assured, you won't!"; "Stay calm! Is that anger you're feeling? Or is it fear? Don't walk away from your feelings. Observe them. Then they're easier to let go"). From an ethical standpoint, it is crucial to understand how persuasive messages with stigmatizing content might lead the user to devalue their past self, while they can also promote a sharper self-understanding. The last level instead relates to *moral deliberation*, that is, the moral values highlighted by the app and instantiated in the actions and lifestyle it recommends (i.e., what values are explicitly or implicitly promoted by smoking-cessation app-based interventions?). In this regard, it is highlighted how such apps tend to promote medicalization as an implicit moral orientation while making other sources of value in lifeless visible; indeed, smokers are usually encouraged to primarily consider the present and future state of their bodies, their aesthetics, and their savings (Rughiniş et al., 2015). Additionally, most apps are also focused on the individual smoker, with little representation of the adverse effects that second-hand smoke has on others (humans, animals, and the environment broadly).

## 4. Discussion

Persuasion aims at influencing individuals' attitudes, beliefs, or behaviors and is rooted in rhetoric, psychology, and sociology. Over time, persuasion techniques and models have evolved, leveraging eloquent speeches, compelling narratives, or logical arguments. In recent decades, traditional methods of persuasion have ingrained technologies

into their core (CPT), reshaping the landscape of influence. CPT closely follow the technological market: their applications target primarily mobile phones and wearables (e.g., smart watches) as hardware and cloud-based (over the internet) applications and services. Access to technologies and services can be considered uneven worldwide. Nevertheless, the scientific interest in CPT is spread, with the primary studies' institutes located in more than twenty countries. There is notable concentration in Europe (42.9%) and North America (42.9%) for the CPT research, which could be attributed to several interrelated factors. These regions benefit from substantial research funding, advanced technological infrastructure, and established academic communities that foster interdisciplinary studies between computer science, healthcare, and behavioral science. Additionally, the prevalence of English as the primary language for scientific publication, coupled with the strong presence of leading journals and conferences in these regions, further amplify the research output. Europe and North America also prioritize digital health innovation, supported by policies and incentives that encourage the integration of digital solutions into healthcare systems, thereby driving more research in areas like computational persuasion. Meanwhile, the lower representation from Asia (8%), Africa (1%), and Oceania (8%) may reflect disparities in funding, healthcare infrastructure, and regional priorities, as well as fewer established networks for conducting and disseminating research.

The dominance of mobile devices in computational persuasion technologies (CPT) for e-health is largely due to their widespread accessibility, convenience, and versatility, allowing for tailored, real-time health interventions that may engage the system users through heterogeneous modalities, such as text, voice, video, and interactive applications (i.e., in compliance with the Nudge Theory). However, several challenges arise from their use, including privacy and security concerns related to the collection of sensitive health data, necessitating robust data protection measures and transparent privacy policies. Additionally, disparities in digital literacy can create a digital divide, particularly among older adults or those less familiar with technology, limiting the reach and effectiveness of these interventions. Addressing such challenges require a holistic approach that combines user-friendly design, strong ethical and security frameworks, content personalization, and social efforts to improve digital literacy and engagement across diverse populations.

Moreover, the elaborated primary studies promote information systems as enabler for user behavioral change. However, studies present computational persuasion techniques at conceptual levels, and only a few prototypes have been tested on a large scale — yet, for a short time. Unfortunately, this limits the relevance of the results — given that behavioral change is measured/observed mostly in the long run. Moreover, although research advancements pass through proof of concepts, designers and developers of CPT must consider the crucial implications of the healthcare domain (e.g., interfaces, accessibility (Oinas-Kukkonen & Harjumaa, 2009), data visibility/use, and psychological effects (Tikka & Oinas-Kukkonen, 2019)) and the importance of assessing efficacy and effectiveness (Almutairi et al., 2023) (overseen by the vast majority of the elaborated studies) making questionable the actual achievement of the selected objectives. Moreover, further research analyzing and testing whether benchmark theoretical frameworks for CPT that can address or fail ethical concerns raised by CPT is also highly needed. This entails that there is still a long way to go for CPT in healthcare. Furthermore, it is imperative to bridge the gap between the technologies produced and the end users (whose capabilities, as of today, are still not adequately considered when designing/producing current solutions). This entails a more careful consideration of the user classification. Indeed, within the same cluster of end users, their needs can be deeply diverse and require different approaches and overall solutions. Furthermore, the boost of acceptability and adherence (in both the short and long term) seems to be critical.

To this end, some studies have identified possible answers in explainable XAI. Supporting the systems recommendations/instructions (overall, decisions) with explanations seems to have moved a step further users acceptance and trust in the CPT. However, explanations are (too often) provided solely in textual form. While sometimes the provided explanation seems to "hit the right spot" being timely, concise, and accurate, oftentimes the user expressed alienation, not understanding how/why a given explanation was meant for them. For example, some explanations have been targeted as too generic (raising reactions like "buff, you tell this to everybody") or too articulated and complex (raising reactions like "I'm not a doctor nor an engineer, what is this?"). Such reactions have occurred in nutritional coaching, wellness, and medical follow-up procedures.

An interesting step further within explanation generation and communication has been moved by a few studies that leveraged negotiation techniques to handle the level of detail parsimoniously and steer the generation future recommendations and relative explanation. Finally, further considerations are moving toward analyzing how far banal manipulation (and its implementation) could go, introducing a subtle transparency trade-off — telling too much (even if correct and sound) might be counterproductive from a psychological perspective. Interestingly, users appreciate this trade-off, valuing both clear explanations and persuasive support to achieve their health goals (which is what they want overall).

Applying Banal manipulation and trade-off AI may be rather straightforward from a technological perspective. However, from a legal perspective AI has been defined as an 'umbrella term' to which very different applications are attributed — yet, having in common only some similar technological traits. This has led legislators to provide guidelines – rather than proper definitions – that would identify as AI all those systems in which certain programming rules are respected (Commission, 2021; HLEG, 2019). This evidently leads to a tautology based only on the technical knowledge we currently have. The undesired effect is flattening the differences between existing classes of applications, thus making it difficult not only to produce an effective regulation (Waldron, 1994), but also one that is flexible enough to be able to project its effectiveness on future technological developments, which are notoriously rapid and constantly evolving.

The idea of horizontal regulation of new technologies, based on the principle of technological neutrality, should therefore be abandoned. At the same time, the ambiguity of the object of the analysis, the too wide circle of possible uses, of users, and of the various conditions of which they are carriers – as proved in this analysis – are claimed as possible obstacles for a legal regulation (Scherer, 2015). This has led to the idea that AI should not be regulated until this problem is solved. However, it is important to underline that the law is already used to regulate what is not very well defined (Danaher, 2021). In those circumstances vagueness is absorbed and to some extent solved within the regulation process. Compared to those examples, AI presents some peculiar difficulties. Firstly, the one related to the fact that the ambiguity about the object can also create ambiguity regarding the *ex-ante* evaluation of its means and the goals it has to pursue (Elish & Boyd, 2018). This is even more true if we consider that we cannot be sure of the impact of a narrow AI implementation as well, for many systems are developed so as to be creative and autonomous in their general operation. Therefore, it could be useful for legal scholars and policymakers to identify legal frameworks that can be more flexibly applicable — yet binding and enforceable, to ensure the protection of individuals. For this reason, a bottom-up approach to regulations is to be preferred to a top-down one. Indeed, the first one would have the merit of starting from the understanding of the technology in question, of its particularities, potentialities, and limitations, while the second one imposes flat normative procedures to AI as a general, not better-clarified entity. At the same time, it would be important to abandon the perspective of "technological neutrality" in regulation, which is lacking in understanding not only the specificities of the different classes of applications but, above all, the radical difference in the impact that they can have on the single user and the whole society.

One consequence of this is the emergence of regulatory approaches that, while proving promising and absolutely necessary, are not yet able to cover the dynamics and potential risks emerging from the interaction with PT. One example is the Digital Service Act (DSA), which addresses recommender systems, for which it seeks to regulate and limit mainly profiling mechanisms and target advertising. However, no reference is made to persuasion systems or, more generally, to all those cases in which a tailor-made interaction with the user turns out to be essential for the achievement of the goal set by this same individual and for his or her own benefit. However, one of the undisputed merits of this regulation is certainly the attempt to protect users of online platforms from the so-called dark patterns (i.e., those design features of the user interface that can manipulate individuals), inducing behavior that they would not have conducted in the absence of those specific patterns. This includes not only adhering to more permissive privacy policies or agreeing to provide much more personal data than one would be comfortable sharing but also the very fact of continuing to use or increasing the time spent using a given application. This attempt to regulate manipulative design certainly has the limitation of the lack of a clear distinction between manipulation and persuasion. Still, it lays the foundations for regulating what may or may not be considered permissible in terms of interface design. Ethical design guidelines had already made a similar attempt, but these had the major limitation of not being legally enforceable and binding and of lacking systematicity. In fact, each industrial sector could adapt such guidelines not only to the type of system developed, but also to the purpose of use of that system, or even to the prototype users to which they refer, thus generating a pluralism that makes effective protection of the individuals involved challenging.

Similarly, the Artificial Intelligence Act (AIA) represents the first attempt to regulate AI systems and adopts an approach based on four different levels of risk, to lay the foundations for legislation that is as comprehensive as possible. Once again, the lack of solid theoretical foundations that can distinguish the persuasive phenomenon from the manipulative one, makes it difficult to understand how PT are intended to be regulated under this Act. Indeed, despite the objective of this regulation to provide clarity on the matter, the absence of unidirectional interpretation within the legal doctrine pertaining to certain expressions utilized by the European legislature indicates that some articles may necessitate further investigation to ensure their accurate implementation and to achieve the intended concrete efficacy. For instance, Article 5 regulates the so-called prohibited practices, which are those capable of producing significant harm by being based on subliminal techniques and deceptive design features that can manipulate the end user. However, no reference is made in the text as to what ‘‘significant harm’’ really means, nor does it specify the criteria to be used in determining the extent of such harm. Consequently, the present wording of the regulation may permit a considerable degree of discretion with regard to the technologies that are to be proscribed in accordance with this article. Moreover, considering that any design device aims to influence – even positively – the user by appealing to his or her most subconscious sphere, the lack of definition of what is to be understood by subliminal practices leaves such a wide margin of interpretation that it is not particularly significant.

AI systems that are used in healthcare, then, are a priori identifiable as high risk. In this regard, the PTs examined in this article could be considered to fall into this category. However, the regulation of systems that may represent a high risk for the psycho-physical integrity and fundamental rights of users are in fact subject to a system of certifications that originate from declarations made by the manufacturers themselves. Thus, the fight against so-called self-made standards, which a part of European legal doctrine had mooted, is to some extent reintroduced through the provisions of Articles 6 and 9 of the AIA. Apart from this, transparency plays a fundamental role in preventing AI systems from being considered prohibited in this Act. In fact, high- and medium-risk technologies are required to meet very high standards of transparency, to be calibrated to the nature of the system in question, and its purpose or context of use. However, this presupposes a view of the human being still based on the polarization between average individuals and vulnerable individuals, according to which an increase in the amount of information provided to a subject – and net of particular cognitive or evolutionary impediments – corresponds to an increase in their awareness and ability to pursue their interests efficiently. Many studies from behavioral psychology, behavioral economics, and even consumer protection disciplines, however, show that this is not the case. Individuals, especially when interacting with AI systems, may be fully aware of the artificial nature of the application, the computational mechanisms behind certain recommendations, and be subjected to the same level of influence – and potential manipulation – as if this information had not been conferred. Moreover, especially in the health care environment, it has been proved that a certain degree of opaqueness can actually ensure accuracy much more that an excessive push toward transparency (Ebers & Navas, 2020; Kiseleva, Kotzinos, & De Hert, 2022).

The Ethical dimension crucially intersects CPT with several angles. Nevertheless, as reported in the results of this SLR (see Section 3 - SRQ16), ethical investigations and studies on CPT within the domain of eHealth are still in their infancy, both on the theoretical and applied levels. Indeed, our inquiry shows a dearth of systematic studies focusing on testing eHealth CPT and analyzing related implications from the ethics perspective.

Kip et al. (2018) pave the way for such an effort, but the ethical themes proposed are explored narrowly. For example, the issue of responsibility in eHealth CPT requires an in-depth inquiry from both an ethical and legal perspective. Indeed, understanding who is responsible for the changes and actions CPT induces to the users (even if in accordance with them) is a matter of both moral responsibility (who is to blame if something goes wrong) and legal responsibility (who should legally respond for that harm). This issue is a central one in the debate on healthcare AI and is generally addressed through distributed responsibility approaches (Morley et al., 2020). Hence, further research should explore, define, and provide case studies of paradigms on distributed responsibility in multi-agent contexts in the specific domain of health CPT. Here, on the one side, those who accept to use them are often charged with responsibility (agreeing to be persuaded and how); however, on the other side, they are often unprovided with the right literature or skills to understand what implications such use choices arise from both an ethical and legal standpoint. Differently from other AI-based applications in healthcare where there is the presence of a physician or a healthcare provider (e.g., diagnostic algorithm-based technology in clinical contexts), the users being alone risk being overcharged with responsibility for consequences they cannot adequately foresee.

The issue of autonomy in health CPT is also poorly explored. The analysis proposed by Rughiniş et al. (2015) provides an insightful context-sensitive inquiry into autonomy but is confined to smoking cessation apps. What the respect and promotion of autonomy entails for the design of CPT in healthcare remained undefined in the scholarship analyzed. For example, there is no clarity on what conception of autonomy and related dimensions is considered when designing health CPT (Tiribelli et al., 2023) and whether diverse cultural interpretations of autonomy are included when developing CPT used at a transnational scale (see Mhlambi & Tiribelli, 2023 on Western and non-western conception of autonomy in AI design). As shown in Tiribelli and Calveresi (2024), the design of PT based on health recommender systems to empower diverse users’ autonomy (e.g., the elderly or the most vulnerable) needs to consider diverse dimensions of autonomy (i.e.: physical, cognitive, epistemic, socio-relational and moral) to provide effective and autonomy-preserving recommendations and advice. Individualized or ethnocentric conceptions of autonomy translate into what values eHealth CPT promotes while persuading users toward health behavior goals, what moral issues are negotiated or sacrificed for such goals, and

the autonomy of whom is truly promoted by such systems (Mhlambi & Tiribelli, 2023). Such issues are central also in the debate in healthcare AI (value-sensitive design), but become even more critical in cases of eHealth technology where the interaction between the user and the app is constant and the users might have more space to personalize how they are persuaded according to their preferences, needs, and values.

The issue of fairness is also under-examined in the few studies considered. Fairness is mainly considered in terms of accessibility, that is, how CPT can improve people's accessibility to health, considering digital skills and the technology divide. However, further fairness-related implications deserve specific exploration in the domain of health CPT. A topic to be specifically explored is that of bias in data and model design of CPT. If health CPT apps are deployed on users who are different from the target sample used to train them (i.e., population target bias) they might produce inaccurate persuasion strategies and advice. These outputs become harmful especially for people with health or psychological conditions (e.g., rare pathologies) not represented in the data input and training and the system design broadly. Such biases would lead to recommended actions for health behavioral change that can be inappropriate and detrimental from an ethical and health perspective, endangering people's safety. Stereotyped representation of protected groups or vulnerable people (e.g., ageism) and other kinds of unfair correlations or biases (gender, ethnicity, etc.) are some of the prominent sources of unfairness and discrimination in AI-based technology in healthcare (Giovanola & Tiribelli, 2023). Therefore, in-depth analyses especially on real applications and case studies are needed to understand whether health CPT intentionally or accidentally embeds, perpetuates, and exacerbates cultural bias leading to unfair and harmful persuasive actions; as well as, whether they target cognitive biases possibly unfairly undermining autonomy.

As mentioned previously, most of today's scholarship is concerned with the use of CPT in general as a possible manipulative tool, stressing it as a threat to individuals' autonomy. However, there is a dearth of research on health CPT as intentionally beneficial persuasive tools and on conditions should be respected to not overcome the fine line between beneficial persuasion and harmful persuasion (or hyper-persuasion as a soft form of manipulation) in the design of eHealth CPT (Christiano, 2022; Ienca, 2023; Klenk, 2024; Rosenberg, 2023; Tiribelli, 2024). Scholars identify specific ethical conditions and criteria that should be respected to ensure persuasion does not infringe autonomy: how such conditions are helpful for the design of eHealth CPT is worth systematic exploration.

Overall, further ethical scholarship is encouraged in the field, adopting both a top-down approach, that is, showing how key ethics principles in bioethics and AI ethics are helpful for the ethical design of CPT in eHealth (see, for example (Tiribelli, Monnot, et al., 2023)), and especially a bottom-up approach, as in Rughiniş et al. (2015), that is, extrapolating ethical considerations based on the ethical assessment of context-sensitive and purpose — specific eHealth CPT systems, to provide actionable moral compasses to engineers called to their design as well as to decision-makers for their trustworthy approval.

## 5. Conclusions

This study systematically reviewed the current literature on adopting computational persuasion technology in the eHealth domain from a technological, ethical and legal perspective, revealing several critical insights.

From a technological and theoretical perspective, the insights indicate that most reviewed studies present CPT at a conceptual level, with a significant portion of prototypes remaining untested (on a large scale and for long periods). This limitation affects the relevance of the results, as behavioral changes are often observed over the long term. The dominant models implemented in CPT include the Persuasive Systems Design (PSD) and Fogg's behavioral models, with notable mentions of nudge theory. However, determining the most effective model remains challenging due to a lack of comparative, pragmatic analysis. Clinical studies, especially targeting mid and long-term effects will be needed in order to assess the actual effectiveness of applying these CPT models in concrete healthcare scenarios.

The application and user context emphasize the critical need to bridge the gap between technological solutions and the diverse needs of end-users. Studies indicate that even within a single user cluster, needs can vary significantly, necessitating tailored approaches. Enhancing user acceptability and adherence to CPT in both the short and long term is paramount. Incorporating XAI techniques into CPT has shown promise, though current implementations often fall short, with explanations perceived as either too generic or overly complex.

Ethical and legal considerations are under-explored in the context of CPT in eHealth. Issues such as responsibility, autonomy, and the potential for manipulation require deeper investigation to ensure that CPT applications are both morally and legally sound. Current ethical frameworks and AI regulations are often too broad, failing to address the specificities of different applications. A bottom-up regulatory approach, starting from the understanding of particular technologies, is recommended to create more effective and flexible regulations. Similarly, further research on ethical frameworks for CPT in healthcare needs to be developed to ensure their ethical design.

Further research and development are essential to fully realize CPT's potential in eHealth. Future studies should focus on large-scale, long-term testing of CPT prototypes and comparative analyses of different persuasion models to determine their relative effectiveness. Developing more nuanced user classifications to create tailored CPT solutions is also crucial. Additionally, exploring ethical implications and establishing clear regulatory guidelines that balance transparency, user autonomy, and protection against manipulation are vital steps forward.

In conclusion, while significant advancements have been made in the field of computational persuasion technology, especially in eHealth, considerable work remains to address current limitations and enhance the effectiveness, acceptability, and ethical grounding of these technologies.

## CRediT authorship contribution statement

**Davide Calvaresi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Rachele Carli:** Writing – original draft, Visualization, Resources, Investigation, Conceptualization. **Simona Tiribelli:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Conceptualization. **Berk Buzcu:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Reyhan Aydogan:** Writing – review & editing, Validation. **Andrea Di Vincenzo:** Writing – original draft, Formal analysis. **Yazan Mualla:** Writing – review & editing, Validation. **Michael Schumacher:** Writing – review & editing, Project administration, Funding acquisition. **Jean-Paul Calbimonte:** Writing – review & editing, Supervision, Data curation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This study has been partially supported by the CHIST-ERA project EXPECTATION (EU grant CHIST-ERA-19-XAI-005 and national grant 20CH21_195530).

## Data availability

No data was used for the research described in the article.

## References


Adaji, I., & Adisa, M. (2022). A review of the use of persuasive technologies to influence sustainable behaviour. In *Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization* (pp. 317–325).

Ainsworth, T. (2012). Improving therapeutic exercise devices for people with rheumatoid arthritis: A research method combining cultural probes and persuasive design theory. In *Persuasive technology: design for health and safety; the 7th international conference on persuasive technology; PERSUASIVE 2012; linköping; Sweden; June 6-8; adjunct proceedings* (pp. 1–4).

Alahäivälä, T., & Oinas-Kukkonen, H. (2016). Understanding persuasion contexts in health gamification: A systematic analysis of gamified health behavior change support systems literature. *International Journal of Medical Informatics*, *96*, 62–70.

Almonani, E., Husain, W., San, O., Almomani, A., & Al-Betar, M. (2014). Mobile game approach to prevent childhood obesity using persuasive technology. In *2014 international conference on computer and information sciences* (pp. 1–5).

Almutairi, N., Vlahu-Gjorgievska, E., & Win, K. (2023). Persuasive features for patient engagement through mhealth applications in managing chronic conditions: A systematic literature review and meta-analysis. *Informatics for Health and Social Care*, *48*, 267–291. http://dx.doi.org/10.1080/17538157.2023.2165083, Publisher: Taylor & Francis eprint: DOI: http://dx.doi.org/10.1080/17538157.2023.2165083.

Ananthanarayan, S., & Siek, K. (2012). Persuasive wearable technology design for health and wellness. In *Proceedings of the 6th international conference on pervasive computing technologies for healthcare* (pp. 236–240).

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: results from a systematic literature review. In *18th International conference on autonomous agents and multiagent systems* (pp. 1078–1088).

Asbjørnsen, R., Wentzel, J., Smedsrød, M., Hjelmesæth, J., Clark, M., Nes, L., et al. (2020). Identifying persuasive design principles and behavior change techniques supporting end user values and needs in ehealth interventions for long-term weight loss maintenance: Qualitative study. *Journal of Medical Internet Research*, *22*, Article e22598.

Azazi, A., Richards, D., & Bilgin, A. (2022). Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. *Autonomous Agents and Multi-Agent Systems*, *36*.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science. American Association for the Advancement of Science*, *373*(6552), 284–286.

Bartlett, Y., Webb, T., & Hawley, M. (2017). Using persuasive technology to increase physical activity in people with chronic obstructive pulmonary disease by encouraging regular walking: A mixed-methods study exploring opinions and preferences. *Journal of Medical Internet Research*, *19*, Article e124.

Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. *Communications of the ACM*, *42*, 51–58. http://dx.doi.org/10.1145/301353.301410.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI. vol. 8* (pp. 8–13).

Blom, J., & Hänninen, R. (2012). Air pollution in everyday life: toward design of persuasive urban air quality services. In *Persuasive technology: design for health and safety; the 7th international conference on persuasive technology; PERSUASIVE 2012; linköping; Sweden; June 6-8; adjunct proceedings* (pp. 5–8).

Boontarig, W., Quirchmayr, G., Chutimasakul, W., & Papasratorn, B. (2014). An evaluation model for analysing persuasive systems in mobile healthcare. In *2014 international conference on computer, information and telecommunication systems* (pp. 1–5).

Bouderhem, R. (2024). Shaping the future of AI in healthcare through ethics and governance. In *Humanities and Social Sciences Communications*: *vol. 11*, (1), (pp. 1–12). Palgrave.

Buzcu, B., Varadhajaran, V., Tchappi, I., Najjar, A., Calvaresi, D., & Aydoğan, R. (2023). R explanation-based negotiation protocol for nutrition virtual coaching. In *PRIMA 2022: principles and practice of multi-agent systems* (pp. 20–36).

Cabrita, M., Akker, H., Tabak, M., Hermens, H., & Vollenbroek-Hutten, M. (2018). Persuasive technology to support active and healthy ageing: An exploration of past, present, and future. *Journal of Biomedical Informatics*, *84*, 17–30.

Calvaresi, D., Carli, R., Piguet, J., Contreras, V., Luzzani, G., Najjar, A., et al. (2023). Ethical and legal considerations for nutrition virtual coaches. *AI and Ethics*, *3*, 1313–1340.

Calvaresi, D., Ciatto, G., Najjar, A., Aydoğan, R., Torre, L., Omicini, A., et al. (2021). Expectation: personalized explainable artificial intelligence for decentralized agents with heterogeneous knowledge. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 331–343).

Carli, R., & Calvaresi, D. (2023). Reinterpreting vulnerability to tackle deception in principles-based XAI for human-computer interaction. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 249–269).

Carli, R., Najjar, A., & Calvaresi, D. (2022). Risk and exposure of XAI in persuasion and argumentation: The case of manipulation. *Explainable and Transparent AI and Multi-Agent Systems*, 204–220.

Cemiloglu, D., Arden-Close, E., Hodge, S., & Ali, R. (2023). Explainable persuasion for interactive design: The case of online gambling. *Journal of Systems and Software*, *195*, Article 111517, https://www.sciencedirect.com/science/article/pii/S0164121222001935.

Chatterjee, S., Byun, J., Pottathil, A., Moore, M., Dutta, K., & Xie, H. (2012). Persuasive sensing: A novel in-home monitoring technology to assist elderly adult diabetic patients. *Persuasive Technology. Design for Health and Safety*, 31–42.

Christiano, T. (2022). Algorithms, manipulation, and democracy. *Canadian Journal of Philosophy*, *52*, 109–124.

Commission, E. (2021). *Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. COM/2021/206 final. COM/2021/206 final*. European Commission.

Contreras, V., Marini, N., Fanda, L., Manzo, G., Mualla, Y., Calbimonte, J., et al. (2022). A dexire for extracting propositional rules from neural networks via binarization. *Electronics*, *11*, 4171.

Coorey, G., Peiris, D., Usherwood, T., Neubeck, L., Mulley, J., & Redfern, J. (2019). Persuasive design features within a consumer-focused ehealth intervention integrated with the electronic health record: A mixed methods study of effectiveness and acceptability. *PLOS ONE*, *14*, Article e0218447.

Danaher, J. (2021). Is effective regulation of AI possible? Eight potential regulatory problems.

Dominic, D., Hounkponou, F., Doh, R., Ansong, E., & Brighter, A. (2013). Promoting physical activity through persuasive technology. *International Journal of Inventive Engineering and Sciences*, *2*, 16–22.

Dragoni, M., Donadello, I., & Eccher, C. (2020). Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, *105*, Article 101840, https://www.sciencedirect.com/science/article/pii/S0933365719310140.

Duwaraka Yoganathan, S. (2013). Persuasive technology for smartphone fitness apps. In *Pacific Asia conference on information systems PACIS 2013 proceedings* (p. 185).

Ebers, M., & Navas, S. (2020). *Algorithms and law*. Cambridge University Press.

Echeburúa, E., Corral, P., & Amor, P. (2003). Evaluation of psychological harm in the victims of violent crime. *Psychology in Spain*, *7*, 10–18.

Elish, M., & Boyd, D. (2018). Situating methods in the magic of big data and AI. *Communication Monographs*, *85*, 57–80.

Erdeniz, S., Trang Tran, T., Felfernig, A., Lubos, S., Schrempf, M., Kramer, D., et al. (2023). Employing nudge theory and persuasive principles with explainable AI in clinical decision support. In *2023 IEEE international conference on bioinformatics and biomedicine* (pp. 2983–2989). [ISSN: 2156-1133] https://ieeexplore.ieee.org/document/10385315.

Fogg, B. (2002). Persuasive technology: using computers to change what we think and do. *Ubiquity*, *2002*, 2.

Fogg, B. (2009). Creating persuasive technologies: an eight-step design process. In *Proceedings of the 4th international conference on persuasive technology* (pp. 1–6).

Fosch-Villaronga, E., Chokoshvili, D., Vallevik, V. B., Ienca, M., & Pierce, R. L. (2021). Implementing AI in healthcare: An ethical and legal analysis based on case studies. In *Data protection and privacy, Volume 13: Data protection and artificial intelligence*: *vol. 13*, (p. 187). Bloomsbury Publishing.

Fritz, T., Huang, E., Murphy, G., & Zimmermann, T. (2014). Persuasive technology in the real world. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 487–496).

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* (pp. 359–373).

Galster, M., Weyns, D., Tofan, D., Michalik, B., & Avgeriou, P. (2014). Variability in software systems-a systematic literature review. *IEEE Transactions on Software Engineering*, *40*, 282–306.

Gemert-Pijnen, L., Kelders, S., Jong, N., & Oinas-Kukkonen, H. (2018). Persuasive health technology. *eHealth Research, Theory and Development*.

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–336). Elsevier.

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society*, *38*, 549–563.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J., Yordanova, K., Vered, M., et al. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, *56*, 3473–3504.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*, 44–58.

Guttman, N., & Salmon, C. (2004). Guilt, fear, stigma and knowledge gaps: ethical issues in public health communication interventions. *Bioethics*, *18*, 531–552.

Henkemans, O., Paradies, G., Neerincx, M., Looije, R., & Pepijn Emepelen, V. (2015). V lost in persuasion A multidisciplinary approach for developing usable, effective, and reproducible persuasive technology for health promotion. In *Proceedings of the 9th international conference on pervasive computing technologies for healthcare*.

HLEG, A. (2019). A definition of artificial intelligence: main capabilities and scientific disciplines. In *High-level expert group on artificial intelligence (AI HLEG)*. 2019 Apr.

Hulstijn, J., Tchappi, I., Najjar, A., & Aydoğan, R. (2023). Metrics for evaluating explainable recommender systems. In *Explainable and transparent AI and multi-agent systems* (pp. 212–230). Cham: Springer Nature Switzerland.

Hunter, A. (2018). Towards a framework for computational persuasion with applications in behaviour change. *Argument & Computation*, *9*, 15–40.

Ienca, M. (2023). On artificial intelligence and manipulation. *Topoi*, *42*, 833–842.

Jacobs, N. (2020). Two ethical concerns about the use of persuasive technology for vulnerable people. *Bioethics*, *34*, 519–526.

Jalil, S. (2013). Persuasion for in-home technology intervened healthcare of chronic disease: Case of diabetes type 2. In *Adjunct proceedings of the 8th international conference on persuasive technology, Vol-973*.

Jalil, S., & Orji, R. (2016). Integrating persuasive technology to telemedical applications for type 2 diabetes. In *Proceedings of the international workshop on personalization in persuasive technologyco-located with the 11th international conference on persuasive technology* (pp. 92–100).

Kampik, T., Nieves, J., & Lindgren, H. (2018). Implementing argumentation-enabled empathic agents. In *European conference on multi-agent systems* (pp. 140–155).

Kelders, S., Kok, R., Ossebaard, H., & Gemert-Pijnen, J. (2012). Persuasive system design does matter: a systematic review of adherence to web-based interventions. *Journal of Medical Internet Research*, *14*, Article e152.

Khalil, A., & Abdallah, S. (2013). Harnessing social dynamics through persuasive technology to promote healthier lifestyle. *Computers in Human Behavior*, *29*, 2674–2681.

Kim, M., Kim, K., Nguyen, T., Ko, J., Zabora, J., Jacobs, E., et al. (2019). Motivating people to sustain healthy lifestyles using persuasive technology: A pilot study of Korean Americans with prediabetes and type 2 diabetes. *Patient Education and Counseling*, *102*, 709–717.

Kip, H., Jong, N., Gemert-Pijnen, L., Sanderman, R., & Kelders, S. (2018). *eHealth research, theory and development: a multi-disciplinary approach*. Routledge.

Kiseleva, A., Kotzinos, D., & De Hert, P. (2022). Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. *Frontiers in Artificial Intelligence. Frontiers Media SA*, *5*, Article 879603.

Kitchenham, B., Brereton, P., Turner, M., Niazi, M., Linkman, S., Pretorius, R., et al. (2010). Refining the systematic literature review process-two participant-observer case studies. *Empirical Software Engineering*, *15*, 618–653.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software enginering*. School of Computer Science.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*, 7–15.

Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. *Ethics and Information Technology*, *26*, 9.

Kueker, D., Koopman, R., McElroy, J., & Moore, J. (2012). Evaluation of persuasive design features in a prototype of a tobacco cessation website. In *Persuasive technology: design for health and safety; the 7th international conference on persuasive technology; PERSUASIVE 2012; linköping; Sweden; June 6-8; adjunct proceedings* (pp. 17–20).

Lee, D., Helal, S., Anton, S., Deugd, S., & Smith, A. (2012). Participatory and persuasive telehealth. *Gerontology*, *58*, 269–281.

Lee, M., Kiesler, S., & Forlizzi, J. (2011). Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 325–334).

Lehto, T. (2012). Designing persuasive health behavior change interventions. In *Critical issues for the development of sustainable e-health solutions* (pp. 163–181). http://dx.doi.org/10.1007/978-1-4614-1536-7_11.

Lentferink, A., Oldenhuis, H., Groot, M., Polstra, L., Velthuijsen, H., & Gemert-Pijnen, J. (2017). Key components in ehealth interventions combining self-tracking and persuasive ecoaching to promote a healthier lifestyle: A scoping review. *Journal of Medical Internet Research*, *19*, Article e277.

Leonard, T., Thaler, Richard H., & Sunstein, Cass R. (2008). *Nudge: improving decisions about health, wealth, and happiness: Yale university press, New Haven, CT, 2008* (p. 293). Springer.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*, 611–627.

Liao, S. M. (2023). *Ethics of AI and health care: towards a substantive human rights framework*: *vol. 42*, (3), (pp. 857–866). Topoi. Springer.

Mackenzie, C., Rogers, W., & Dodds, S. (2014). Introduction: What is vulnerability and why does it matter for moral theory. In *Vulnerability: new essays in ethics and feminist philosophy* (pp. 1–29).

Matthews, J., Win, K., Oinas-Kukkonen, H., & Freeman, M. (2016). Persuasive technology in mobile applications promoting physical activity: a systematic review. *Journal of Medical Systems*, *40*.

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). *Ethical and regulatory challenges of AI technologies in healthcare: a narrative review*. Heliyon: Elsevier.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. *Topoi*, *42*, 867–880.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., et al. (2020). The ethics of AI in health care: a mapping review. In *Social science & medicine*: *vol. 260*, Elsevier, Article 113172.

Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., et al. (2022). The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial Intelligence*, *302*, Article 103573.

Mylonopoulou, V. (2018). Design for health behavior change supportive technology. In *Proceedings of the 10th nordic conference on human-computer interaction*.

Nair, S., & Howlett, M. (2017). Policy myopia as a source of policy failure: Adaptation and policy learning under deep uncertainty. *Policy & Politics*, *45*, 103–118.

Nordström, M. (2022). AI under great uncertainty: implications and decision strategies for public policy. *AI & Society*, *37*, 1703–1714.

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, *24*, 28.

O'neill, O. (2017). Some limits of informed consent. *The Elderly*, 103–106.

Orji, R., & Moffatt, K. (2016). Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Informatics Journal*, *24*, 66–91.

Oyebode, O., & Orji, R. (2023). Persuasive strategy implementation choices and their effectiveness: towards personalised persuasive systems. *Behaviour & Information Technology*, *42*, 2176–2209. http://dx.doi.org/10.1080/0144929X.2022.2112612, Publisher: Taylor & Francis eprint: DOI: http://dx.doi.org/10.1080/0144929X.2022.2112612.

Oyebode, O., Steeves, D., & Orji, R. (2024). Persuasive strategies and emotional states: towards designing personalized and emotion-adaptive persuasive systems. *User Modeling and User-Adapted Interaction*, http://dx.doi.org/10.1007/s11257-023-09390-x.

Oyibo, K. (2016). Designing culture-based persuasive technology to promote physical activity among university students. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*.

Oyibo, K., & Vassileva, J. (2020). HOMEX: Persuasive technology acceptance model and the moderating effect of culture. *Frontiers in Computer Science*, *2*.

Pinzon, O., & Iyengar, M. (2012). Persuasive technology and mobile health: A systematic review. In *Persuasive technology: design for health and safety; the 7th international conference on persuasive technology; PERSUASIVE 2012; linköping; Sweden; June 6-8; adjunct proceedings* (pp. 45–48).

Pommeranz, A., Detweiler, C., Wiggers, P., & Jonker, C. (2012). Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, *14*, 285–303.

Premanandan, S., Ahmad, A., Cajander, A., Agerfalk, P., & Gemert-Pijnen, L. (2023). Design suggestions for a persuasive e-coaching application: A study on informal caregivers' needs. *Digital Health*, *9*, Article 20552076231177129. http://dx.doi.org/10.1177/20552076231177129, Publisher: SAGE Publications Ltd.

Purpura, S., Schwanda, V., Williams, K., Stubler, W., & Sengers, P. (2011). Fit4life. In *Proceedings of the SIGCHI conference on human factors in computing systems*.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. In *Journal of the American medical informatics association*: *vol. 27*, (3), (pp. 491–497). Oxford University Press.

Renda, A., et al. (2019). *Artificial intelligence. Ethics, governance and policy challenges*. CEPS Centre for European Policy Studies.

Roberge, J., Senneville, M., & Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, *7*, Article 2053951720919968.

Rosenberg, N. (1995). Why technology forecasts often fail. *The Futurist*, *29*, 16.

Rosenberg, L. (2023). The manipulation problem: conversational AI as a threat to epistemic agency. arXiv preprint arXiv:2306.11748.

Rudinow, J. (1978). Manipulation. *Ethics*, *88*, 338–347.

Rughiniş, C., Rughiniş, R., & Matei, Ş. (2015). A touching app voice thinking about ethics of persuasive technology through an analysis of mobile smoking-cessation apps. *Ethics and Information Technology*, *17*, 295–309.

Samonte, M., Medina, E., San Juan, K., & Celestial, A. (2023). A chronic pain management e-health system through persuasive strategies. In *Proceedings of the 2023 6th international conference on information science and systems* (pp. 262–272). https://dl.acm.org/doi/10.1145/3625156.3625194.

Sara, A., & Mostafa, H. (2019). Study and analysis the effectiveness of persuasive systems for behavioral change. In *Proceedings of the 4th international conference on big data and internet of things* (pp. 1–6).

Sassoon, I., Kökciyan, N., Sklar, E., & Parsons, S. (2019). Explainable argumentation for wellness consultation. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 186–202.

Scherer, M. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, *29*, 353.

Schnall, R., Bakken, S., Rojas, M., Travers, J., & Carballo-Dieguez, A. (2015). mHealth technology as a persuasive tool for treatment, care and management of persons living with HIV. *AIDS and Behavior*, *19*, 81–89.

Schoenherr, J. (2021). Trust and explainability in a/IS-mediated healthcare: Operationalizing the therapeutic alliance in a distributed system. In *2021 IEEE international symposium on technology and society* (pp. 1–8).

Sebastian, G., George, A., & Jackson, G. (2023). Persuading patients using rhetoric to improve artificial intelligence adoption: Experimental study. *Journal of Medical Internet Research*, *25*, Article e41430, http://www.ncbi.nlm.nih.gov/pubmed/36912869.

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: The myth of individual thought and the power of collective wisdom*. Pan Macmillan.

Spahn, A. (2012). And lead us (not) into persuasion…? Persuasive technology and the ethics of communication. *Science and Engineering Ethics*, *18*, 633–650.

Spanakis, E., Santana, S., Ben-David, B., Marias, K., & Tziraki, C. (2014). Persuasive technology for healthy aging and wellbeing. In *2014 4th international conference on wireless mobile communication and healthcare-transforming healthcare through innovations in mobile and wireless technologies* (pp. 23–23).

Srisawangwong, P., & Kasemvilas, S. (2014). Mobile persuasive technology: a review on thai elders health service opportunity. In *2014 14th international symposium on communications and information technologies*.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*.

Taj, F., Klein, M. C. A., & van Halteren, A. (2019). Digital health behavior change technology: bibliometric and scoping review of two decades of research. *JMIR mHealth and uHealth*, *7*(12), Article e13311.

Tarling, R., & Burrows, J. (2004). The nature and outcome of going missing: the challenge of developing effective risk assessment procedures. *International Journal of Police Science & Management*, *6*, 16–26.

Thaler, R., & Sunstein, C. (2009). *Nudge: improving decisions about health, wealth, and happiness*. Penguin.

Tian, X., Risha, Z., Ahmed, I., Narayanan, A., & Biehl, J. (2021). Let's talk it out. In *Proceedings of the ACM on human-computer interaction, vol. 5* (pp. 1–32).

Tikka, P., & Oinas-Kukkonen, H. (2019). Tailoring persuasive technology: A systematic review of literature of self-schema theory and transformative learning theory in persuasive technology context. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *13*.

Tiribelli, S. (2024). Who decides what online and beyond: freedom of choice in predictive machine-learning algorithms. *Ethics in Online AI-based Systems*, 299–321.

Tiribelli, S., & Calveresi, D. (2024). Rethinking health recommender systems for active aging: An autonomy-based ethical analysis. In *Science and engineering ethics*: *vol. 30*, (p. 22). Springer.

Tiribelli, S., Monnot, A., Shah, S., Arora, A., Toong, P., & Kong, S. (2023). Ethics principles for artificial intelligence–based telemedicine for public health. *American Journal of Public Health*, *113*, 577–584.

Tiribelli, S., et al. (2023). The AI ethics principle of autonomy in health recommender systems. *Argumenta*, *16*, 1–18.

Tsiakas, K., Barakova, E., Khan, J., & Markopoulos, P. (2020). BrainHood: Towards an explainable recommendation system for self-regulated cognitive training in children. In *Proceedings of the 13th ACM international conference on pErvasive technologies related to assistive environments*. http://dx.doi.org/10.1145/3389189.3398004.

Tsvyatkova, D. (2013). Persuasive technology in healthcare: Designing for children with diabetes. In *HCI 2013 the 27th international british computer society human computer interaction conference: the internet of things*.

Waldron, J. (1994). Vagueness in law and language: Some philosophical issues. *California Law Review*, *82*, 509.

Wang, P. (2020). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *11*, 73–86.

Wang, Y., Wu, L., Lange, J., Fadhil, A., & Reiterer, H. (2018). Persuasive technology in reducing prolonged sedentary behavior at work: A systematic review. *Smart Health*, *7–8*, 19–30.

(WHO) *W WHO outlines considerations for regulation of artificial intelligence for health*. World Health Organisation, News Release.

Wiafe, I., & Nakata, K. (2010). A semiotic analysis of persuasive technology: an application to obesity management. In *12th international conference on informatics and semiotics in organisations* (pp. 157–164). http://centaur.reading.ac.uk/18178/.

Yoganathan, D., & Sangaralingam, K. (2015). Designing fitness apps using persuasive technology: A text mining approach. In *Pacific Asia conference on information systems PACIS*.