Give and Take: Perceptions of a Conversational Coach Agent in Fitness Trackers

Citation for the original published paper (version of record):

Loerakker, M., Stefanidi, E., Niess, J. et al (2025). Give and Take: Perceptions of a Conversational Coach Agent in Fitness Trackers. Proceedings of the ACM on Human-Computer Interaction, 9(5). http://dx.doi.org/10.1145/3743718

(article starts on next page)

Give and Take: Perceptions of a Conversational Coach Agent in Fitness Trackers

MEAGAN B. LOERAKKER, TU Wien, Austria and Chalmers University of Technology, Sweden
EVROPI STEFANIDI, TU Wien, Austria
JASMIN NIESS, University of Oslo, Norway
THOMAS EẞMEYER, University of Bremen, Germany
PAWEŁ W. WOŹNIAK, TU Wien, Austria

Fig. 1. A visualisation of the WHOOP fitness tracker and its accompanying app used in this study. The tracker does not have a display, so all interactions take place in the mobile application. The main focus of the study is the participant's interaction with the so-called *WHOOP Coach*, a conversational agent embedded in the mobile app.

While Personal Informatics (PI) tools utilise data visualisations to communicate behaviour, users often struggle to make sense of their data and translate it to actionable insights. Conversational Agents (CAs) offer potential

Authors' Contact Information: Meagan B. Loerakker, TU Wien, Vienna, Austria and Chalmers University of Technology, Gothenburg, Sweden, meagan.loerakker@tuwien.ac.at; Evropi Stefanidi, TU Wien, Vienna, Austria, evropi.stefanidi@tuwien.ac.at; Jasmin Niess, University of Oslo, Oslo, Norway, jasminni@uio.no; Thomas Eßmeyer, University of Bremen, Bremen, Germany, mildner@uni-bremen.de; Paweł W. Woźniak, TU Wien, Vienna, Austria, pawel.wozniak@tuwien.ac.at.

ACM 2573-0142/2025/9-ARTMHCI014
https://doi.org/10.1145/3743718

for improved access to PI data, yet their role in PI tools remains under-explored. We conducted a two-week user study with journals, interviews and logging with $n = 36$ participants using a novel commercial fitness tracker with an embedded CA. We identified the *give and take* principle as essential for meaningful sensemaking with a CA—a dynamic resulting in more effective interactions given users' inputs (*give*) are met with prompts that are sufficiently specific and built upon prior data engagement (*take*). A critical point was how users perceived the CA during their initial interactions, with first impressions often determining further engagement. We contribute insights into how CAs can support or hinder the PI experience, offering implications for future PI system designs.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: Fitness tracker, conversational user interfaces, conversational agent, chatbot, personal informatics, sensemaking, metrics, give and take

ACM Reference Format:
Meagan B. Loerakker, Evropi Stefanidi, Jasmin Niess, Thomas Eßmeyer, and Paweł W. Woźniak. 2025. Give and Take: Perceptions of a Conversational Coach Agent in Fitness Trackers. *Proc. ACM Hum.-Comput. Interact.* 9, 5, Article MHCI014 (September 2025), 36 pages. https://doi.org/10.1145/3743718

1 Introduction

Personal Informatics (PI) systems seek to empower users to track and reflect on their personal data with the goal of fostering self-awareness and wellbeing [36]. However, making sense of this data and identifying actionable insights can be challenging, as reflection often does not occur automatically, but needs to be encouraged [124]. Moreover, it can be hard to interpret one's data without contextualising [125]. While the PI experience is often seen as solitary, a small body of PI work has explored PI as a social activity [1, 25]. Thus, one promising direction for interpreting data appears to be 'social sensemaking' [106], where sharing data with others facilitates meaningful dialogues, encourages question-asking, and promotes drawing comparisons.

While previous studies have focused on human-human interactions, it is increasingly important to examine PI systems that include human-agent interactions, given the proliferation of Conversational Agents (CAs) in commercial devices over the past decade. This has led to research starting to integrate CAs into PI tools [80]. CAs can offer more human-like, and thus more engaging interactions [66], introducing a new approach to personal data exploration. Therefore, as CAs continue to proliferate in mobile and wearable devices—with and without Large Language Models (LLMs)—it is increasingly important to study how users interpret and interact with these agents in applications that collect personal data, like PI tools. The question remains of how CAs can be effectively and meaningfully incorporated into these systems.

The integration of Artificial Intelligence (AI) into PI systems poses multiple challenges related to design, interaction, and sociotechnical factors [88]. For example, self-tracking tools rely on individual user datasets, which are often too small for typical ML models to generate personalised insights [39]. Furthermore, although data from larger groups of users could enable the CA to offer more insightful recommendations and suggestions, there is a high risk of offering biased and, therefore, flawed advice for people with more unique health conditions (e.g. medical conditions, disabilities) and minority groups [98]. Strömel et al. [127] found that textual data representations created by ChatGPT enriched the sensemaking process and guided users to focus, while reinforcing comparative mindsets. This presents a critical challenge of how PI-related CAs should be designed responsibly to enhance user experience while aligning capabilities with users' ethical values and beliefs [92]. Thus, it remains a challenge for Human-Computer Interaction (HCI) to explore how such technologies can be employed in an effective, ethical and usable manner.

This study was prompted by the market debut of the first commercial LLM-based PI tool, WHOOP Coach.[1] WHOOP, an established subscription-based fitness tracker vendor, enabled a chat feature in their application where users could, quoting WHOOP, *'choose what you want to focus on, or where you need more support, and WHOOP will coach you based on your body and your goals, providing individualised advice, insights, workout recommendations, nutrition coaching, fitness plans, and more.'* WHOOP Coach offered an opportunity to explore in what ways a CA could help users gain a better understanding of their fitness data. To the best of our knowledge, this is the first study that explores the continuous use of a CA system in PI. We explore how and why users employ a CA in personal tracking and their perceptions thereof. We present a central finding of our inquiry: effective PI sensemaking with a CA depends on what we term the *give and take* principle. This principle encapsulates a design requirement involving two key elements: 'data availability' (*give*) and 'prompt specificity' (*take*). When the user self-tracks a sufficient amount of personal data and articulates specific questions, the CA is more likely to generate personalised responses tailored to the user's needs.

To this end, this paper contributes: (1) a within-subject empirical study with $n = 36$ participants which describes the lived experiences of using a CA in a fitness tracking concept; (2) findings on how a CA can aid in the interpretation of tracker metrics and how it influences users' perceptions of their behaviours and routines; and (3) design considerations for the future development of CAs as interaction styles in PI tools.

2 Related Work

Here, we discuss HCI literature on conversational agents, how sensemaking has played a role in a user's PI journey and how users interact with metrics and data visualisations in PI tools in order to situate our work.

2.1 Conversational Agents for Wellbeing and Coaching Purposes

In HCI, different models have been developed with the goal of aligning the design of technologies with their users, such as Norman's action cycle [97] or Abowd and Beale's general interaction model [3]. However, as these models have been created with tangible or Graphical User Interfaces (GUIs) in mind, they often fall short in the context of users interacting with CAs. Per their nature, CAs have specific affordances that hinder the mapping of existing best practices from, for instance, GUIs [94] and require context-specific heuristics to inform usable design [75]. Building on existing HCI literature, models such as the CUI Expectation Cycle [92] and the Human-Agent Speech Interaction Model [144] have been developed to account for the unique presentational and perceptual effects that Conversational Agents (CAs) have on users.

Understanding their specific affordances is a crucial step to developing CAs that meet user expectations, which are often shaped by the overall increasing fidelity of available devices [135], further pushed by advancements in contemporary LLMs [71]. In successful cases, contemporary CA interactions can be promising aids in context- and domain-specific tasks, like healthcare [114, 140]. In this work, we follow recommendations of this existing body of work to explore the effects of CAs when coupled with fitness trackers.

HCI has long recognised wellbeing as a key design goal. With the rise of chatbots and CAs) over the last decade, a plethora of applications have emerged that employ CAs for wellbeing. These interfaces can help users navigate complex tasks by providing cognitive scaffolding [42, 109, 110, 139], and have been proven effective in guiding users through reflective practices to promote wellbeing and support mental health [18, 20, 70, 78, 79, 87, 105]. For instance, Kim et al. [67]

[1] https://www.whoop.com

illustrate the potential of CAs in addressing teenagers' emotional needs primarily by being good listeners, respecting their users' privacy by keeping secrets, and providing sufficient background when presenting information.

Depending on the setup, CAs leverage their interaction style to offer users novel and exciting ways to interact with technology. For instance, focusing on smart speakers, Reicherts et al. [111] showcased the appropriateness for CAs to expand their capabilities from prompt-based machines and become proactive conversational partners in social settings. By acting as fact-checkers, reminders, or health assistants, such as detecting recurring symptoms like coughing, this study highlights how benevolent CA interactions can positively affect users' wellbeing. Building on this work, Zargham et al. [145] illustrate the caveats of CA proactivity and offer guidance for the development of interactions that meet user expectations. Importantly, but not surprisingly, CA proactivity is mainly welcomed when expected and the context affords it, or if delivered information is of certain priority or urgency. In any case, users desire agency and prefer limited proactivity that announces itself first. Ruitenburg et al. [114] explored how CAs can assist couples in which one partner has dementia, experiencing reality disjunctions that lead to disagreements and possible conflicts. The work shows that CAs can help as fact-checkers, conflict defusers, or storytellers to maintain truth, comfort and connection between partners. Though these works consider neither physical activity nor health, they imply that a CA can indeed take different roles, such as that of a coach, as long as the interactions meet their users' expectations. Addressing this gap in this work, our aim is to connect these strands to explore a CA's capabilities as a coach in fitness trackers.

Although prior research has proposed that LLMs can take on the role of a 'coach' [13], and AI features are increasingly integrated in more devices, it becomes critical to deepen our understanding of how users interact with and perceive these advancements, especially when their health and wellbeing is concerned. For such AI systems to be effective, they must address longstanding challenges, including lack of explainability in AI-generated recommendations [72], and the presence of biases regarding sensitive attributes, including but not limited to, language [59, 113], age [63], and race [90, 133]. Concurrently, it has been shown that users can 'blindly' trust AI systems regardless of whether or not it is warranted [69]. Generally, overreliance on AI poses negative impacts on cognitive abilities, such as decision-making and critical thinking [146]. Transparency becomes especially meaningful given it is 'contextually appropriate' [26, 28]. This highlights a need for a greater understanding of how to design transparent CAs to minimise the risk of deception and to ensure that users can make informed decisions. In light of these risks, the limited real-world adoption of previous health-oriented CAs [2] substantiates the need for more knowledge on how we can employ these technologies in everyday contexts in both a safe and appropriate manner.

2.2 Goal Setting through Reflection on Data Representations

One recurring problem with PI tools is that 'assumptions' about the user's health goals and motivation to track are embedded within the device [120]. A commonly embedded assumption in PI devices is that the user always wants to improve their 'performance' and 'health' (e.g. [6]). A risk of performance-based foci in PI tools is that a user's behaviour may be framed negatively when they do not achieve their self-defined goals or the goals set by the device (e.g. [83]). Past work has also suggested that positive framing is overall a more desirable design technique than negative techniques, e.g. [23, 27, 37]. Proposed solutions include the option to track only the necessary data to achieve the corresponding goal, which is named *goal-directed self-tracking* [121], as well as to have default assumptions along the lines of 'maintaining' health. Although an excessive focus on goal-setting can become problematic, goal-setting can also be approached in different, arguably non-problematic, manners. For example, by giving users more control over the device through customisation features (e.g. [77]), though 'true' customisation is considered to remain challenging [126]. CAs could provide a promising way forward, providing means of receiving personalised recommendations with varying degrees of consideration for goal setting. However, they have to be explored in more depth to gain a deeper understanding of how those kinds of interactions with CAs would look like.

Agapie et al. [4] argued that reflective practices are one of the most effective ways to help people set goals, whereby a person may go through different stages of reflection, like those stages described by Fleck and Fitzpatrick [43]. Reflection refers to the introspective process in which individuals examine and evaluate their own thoughts, emotions, and behaviours [51]. In general, positive reflection can improve mood, increase the capacity to enjoy life, support the maintenance of relationships, help process past events, and contribute to the development of self-identity [86]. Still, even reflection on negative experiences can lead to health benefits [101]. However, it is important to distinguish between reflecting on negative experiences and rumination [131], which involves persistent cycles of negative thoughts and emotions that can stem from reflection [131] and often stand in contrast to the positive aspects of reflective practice [131].

Reflective practices are vital for acquiring a better understanding of one's personal data and, therefore, one's behaviour [14]. Reflection is a means towards inquiring about one's goals from a higher-level perspective [22], facilitated through perspective developments and changes [8, 43]. For example, users may forcibly adopt 'data interpretation approaches'—albeit flawed—when too few data points are provided, or when too much information is provided without contextualising them or showing their relations [52]. Herrewijnen et al. [52] refer to these phenomena as construction and deconstruction, respectively. Overall, the authors urge for detail in data representations while making the data points' connections explicit, rather than continuously adding more data points without situating them in the representation as a whole [52]. However, more research is needed to understand data sensemaking processes better. Research has demonstrated that technology-mediated reflection can enhance wellbeing [60]; thus, a plethora of HCI researchers have focused on exploring technology-mediated reflection and how to design for it (e.g. [9, 12, 60, 102, 115, 116, 119]). However, reflection can be a challenging activity and often does not occur automatically but needs to be encouraged [124]. In the PI community, it has not yet been explored whether CAs could be an effective means of instigating reflection and aiding sensemaking, as there is a lack of empirical contributions on these phenomena in the context of personal health and fitness data that has been self-tracked by the user. Although prior user studies (e.g. [67, 70, 78, 79]) have investigated how CA interaction plays a role in user wellbeing, it remains a challenge to study how users exhibit health-reflective behaviours with a CA through data sensemaking behaviours, like goal setting.

2.3 Sensemaking and Interpreting Metrics in Personal Informatics Systems

In the HCI community, PI research has focused on topics like how users set goals (e.g. [4]), how their self-tracking motivation changes over time (e.g. [38]), how to support people in transition periods after lapsing with technology usage (e.g. [37, 38, 96], how they adapt their health routines according to the data (e.g. [34]), and how they reflect on their data (e.g. [10, 12, 84]). However, in the literature, it has been reported that users are not always sure how to connect the quantitative data to their holistic real-life goals [96], and that the quantification of behaviour is not always considered appropriate [29]. Interpretation issues can also arise when there is a mismatch between how the data are measured and the user's lived experience of the metric which the tool aims to communicate (e.g. [32]). In addition, users may experience goal-setting with PI tools as overwhelming [35] for a variety of reasons.

In their work, Strömel et al. [127] incorporated ChatGPT into a web application by which participants were requested to import their personal data from applications like Fitbit, Garmin or Strava. Then, ChatGPT either created textual explanations of the participants' data or bar graphs, or both were presented [127]. Their objective was to examine whether unconventional representations of personal data can compensate the disadvantages of more traditional representations [127]. However, they argue that LLM models still have a tendency to hallucinate, attributed to lacking world knowledge [147]. This hallucination risk is especially relevant for LLM models applied in PI, considering that the usage of personal health data already poses similar issues [68], showing a clear gap in how technologies should 'interpret' and 'contextualise' metrics for more sensible communication of the personal data without 'making stuff up'. Quoting Goyal and Fussell [50, p. 288]: *'Sensemaking involves foraging for information pieces that could connect with each other, resulting in multiple initial hypotheses'*, the sensemaking process involves translating factors from the past (connecting pieces together) to informed expectations for the future. In other words, sensemaking is a means towards attaining a mental model that depicts the state of affairs and knowledge [104]. However, the construction of such mental models for the interpretation of PI data remains to be examined in great detail.

One promising area where users could potentially benefit from AI is in the data sensemaking process. Making sense of one's personal data takes a considerate amount of time and energy—not only to achieve goals and engage with data but also to develop the skills needed to interact effectively with PI tools [24, 25, 107]. Moreover, different types of metrics are perceived differently by users [64], perpetuating the assumption that we need more measures to support or even strengthen the sensemaking procedure. With the rise of complex metrics, like sleep, strain (energy used), and recovery (body battery), it is crucial to acquire more in-depth knowledge of whether users find these metrics intuitive and actionable within their health routines. Arguably, users may question the extent to which metrics are accurate and realistic representations of their lived experiences. Related to the hallucination phenomenon, users may question if the data is unveiling their objective experience of the self or about 'me'—the subjective experience of the self—which are both relevant to the sensemaking approach. In other words, *is there a potential (mis)match between the CA's interpretation of the user's data and the user's lived experience?* This work explores the consequences of using a CA as a part of PI experience.

2.4 Research Questions

Following the research gaps we identified in the HCI literature on AI applications, conversational agents, data sensemaking processes, and reflecting on complex metrics, we introduce four research questions to guide our research:

- **RQ1:** *What purposes do users have for using a CA in PI?*
- **RQ2:** *How are the interactions with a CA enacted?*
- **RQ3:** *How do users perceive and feel about interacting with their tracking data through a CA?*
- **RQ4:** *What are the design considerations for CAs as an interaction style in PI systems?*

3 Method

In this section, we describe how we examined the purposes for which participants interacted with the CA, the nature of their interactions, and how they perceived the CA's responses. To explore these aspects, participants interacted with a CA and their chosen metrics. Based on these findings, we developed design considerations to inform the design of future CAs that support the data sensemaking process. This study employed data source triangulation by collecting data from multiple sources—interviews, daily diaries, and users' interaction logs with the CA—to provide a comprehensive understanding of interaction phenomena (e.g. [73, 100]). The study began in July 2024 and concluded in February 2025. Since WHOOP operates on a subscription-based model, we provided participants with pre-made accounts. A total of ten individual user accounts were used in

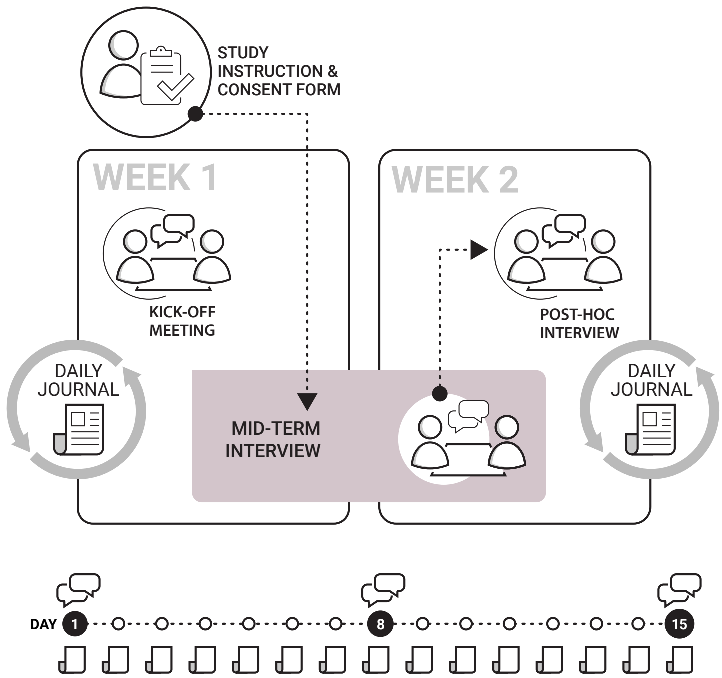


Fig. 2. An overview of the study structure, including when participants were given particular documents, when certain meetings were held, and when participants were interviewed about their experiences with the tracker. Besides those study elements, the participants could fill in a daily journal continuously throughout the entire experiment. The kick-off meeting was held on day 1 of the experiment, the mid-term interview on day 8, and the post-hoc interview on day 15.

the study. Notably, the WHOOP Coach was an entirely new product at the time of the study, thus it was impossible for any of the participants to have prior experience of using the system.

This user study was conducted in compliance with the ethical regulations of the first author's affiliation and local jurisdiction.

3.1 Study Structure

This study was conducted over two weeks, divided into two phases. Figure 2 provides a simplified overview of the study structure. To examine the impact of the WHOOP Coach feature, participants were randomly assigned to a condition where the feature was either enabled or disabled in the first week, then switched in the second week.

3.1.1 Study Preparation. Before the start of the study, participants received a study instructions document, which included a consent form, a link to the daily journal and a link to an instruction video. Upon signing the consent form, the researcher and the participant selected a day and time for the kick-off meeting.

3.1.2 Week 1. The first week started with a kick-off meeting between the researcher and the participant. The researcher asked if the participant had time to read the instruction document; if not, the author explained the study structure. Next, the first author handed the WHOOP fitness tracker over and helped them install the WHOOP app and log into the account assigned to that participant. Then, the participant was shown an instruction video on how to use the WHOOP tracker if they had not watched it upon receiving the study instructions document (see Appendix A for details on the video's content).

The instruction video did not contain any information on how or what they could use the tracker for, as we did not want to bias the participants using particular metrics. Instead, we wanted their decision-making and behaviour with the tracker to be as natural as possible.We wanted to let them choose and interact with the metrics of their desire. After the instruction video, the participant was asked if they had any further questions regarding the study and the WHOOP tracker.

During the kick-off meeting, depending on the participants' time availability, they were either interviewed on their health routines and preconceptions about artificial intelligence, or given a form with the same questions. The form allowed them to complete it in their free time. Participants were encouraged to complete the form as soon as possible, but could elect to provide it later. The questions pertained to long-term habits and prior technology usage and thus the completion time was not a factor.

After the kick-off meeting, the participants engaged with the WHOOP tracker at their own discretion, using it as frequently and in any manner they preferred. Participants were informed they could reach out with questions at any time during the study.

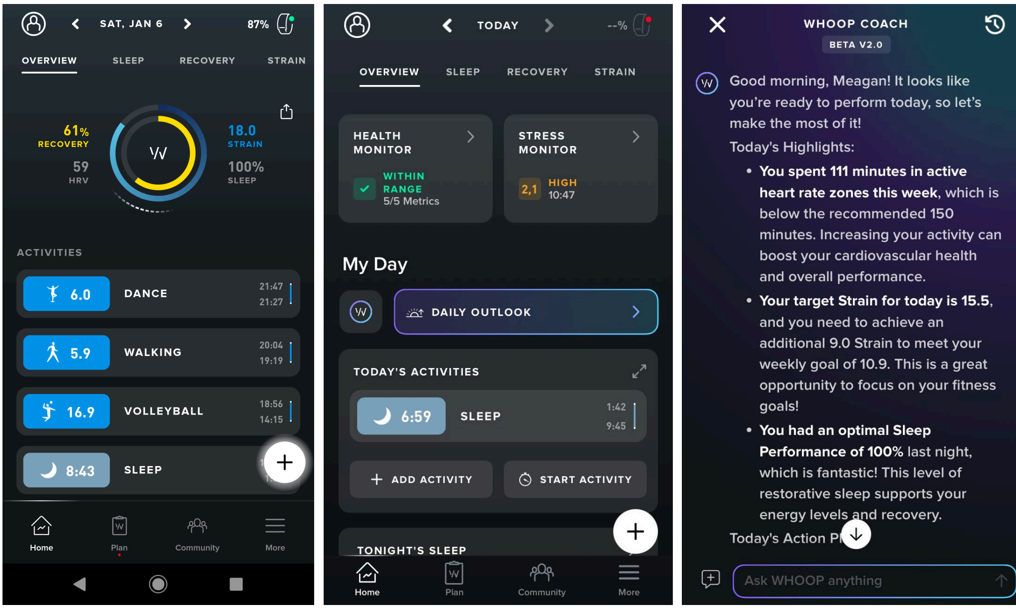
3.1.3 Week 2. After the first week, the researcher scheduled the mid-term interview with the participant, either online or in person. After the interview, the researcher changed the setting of the WHOOP Coach, turning the Coach feature either on or off, depending on the participant's randomly assigned condition order. If the participant participated in the study remotely, the researcher could log in to their account and change the coach's setting. Thereafter, the participants were again free to use the WHOOP tracker however they wanted. After the participant used the tracker for one more week, the researcher and the participant met one last time to have a final post-hoc interview about the participant's experiences with the tracker.

3.2 Apparatus: The WHOOP Fitness Tracker and the WHOOP Coach

The WHOOP fitness tracker is marketed as a PI tool designed for goal setting and achievement, activity planning and management, and personalised recommendations. The WHOOP Coach is powered by OpenAI's GPT-4 [138]. Before incorporating the device into the study, the first author used and experimented with it for approximately nine months. Based on these observations, the tracker provides a wide range of metrics, some of which are relatively unique. Two particularly noteworthy metrics are "strain" and "recovery," which are prominently displayed at the top of the WHOOP app's home screen. These metrics are visually represented as circles that fill up or deplete depending on the user's activity.

At the time of the study, the WHOOP Coach feature was embedded on the home screen, directly below these two primary metrics, functioning as a separate tool. As a result, users could easily ignore it if they chose to. When opened, the CA typically automatically provides a summary of the user's data and activities, such as a daily overview. Additionally, it offered a set of predefined questions that users could select, prompting automated responses. Users also had the option to type their own questions. All interactions with the CA were text-based. Illustrative screenshots of the WHOOP mobile application and its WHOOP Coach feature are presented in Figure 3.

Opting for a commercial fitness tracker rather than building our own system is an intentional decision as it increases the study's ecological validity (e.g. [16, 128, 134]). Our goal was to study CA interactions in real-world settings (see [112]), which necessitates the use of a robust prototype to



(a) (b) (c)

Fig. 3. Three screenshots of the WHOOP mobile application. Subfigure (a) shows the main homepage of the app, with the strain and recovery metrics at the top visualised in blue and yellow circles, respectively. Below the strain and recovery, an overview is provided of the activities the user did that day: dance, walking, volleyball, and sleep. Subfigure (b) visualises the health and stress monitor, and the positioning of the WHOOP Coach's and its *daily outlook*: directly below the *My Day* header. Below the WHOOP Coach, data on *Today's Activities* and *Tonight's Sleep* is given. Subfigure (c) is a screenshot of the beginning of a session with the WHOOP Coach. When opening the *daily outlook* window, the Coach starts the conversation by presenting a summary of the user's week and personalised recommendations. All three screenshots are taken at different points in time when the first author was exploring the WHOOP before the start of the official study.

ensure participant engagement throughout the study. The WHOOP tracker, being a commercially developed and robust product, was expected to facilitate a more reliable and ecologically valid user experience compared to a custom-built prototype. At the time of the study, the WHOOP Coach feature had only recently been released, offering a timely opportunity to employ it in our study with a low probability that participants were already familiar with or had previously used the system.

3.3 Interview and Diary Protocols

Each participant took part in at least two and at most three semi-structured interviews. The first interview, which was optional, took place before the study; participants could choose to complete a survey instead. The second interview was conducted midway through the study, one week after the start, and the final interview was held at the end. The mid-study and post-study interviews followed a similar protocol to examine how participants' opinions evolved over time.

3.3.1 Daily Diary. Participants were given a link to a questionnaire—which functioned as a daily diary—and were asked to use this same link for the whole duration of the two-week study whenever they made any special observations of the WHOOP app, or had any particular opinions to express.

Table 1. Overview of the participants' demographics information, including their participant ID (PID), gender, age, what types of tracking devices they have used (including devices they are currently using), what brand of tracking devices, and their total tracking experience in years (yr), months (mo) and weeks (wk).

PID	Gender	Age	Type(s) of device(s) used	Tracking device(s) used	Total experience
P1	Female	29	Fitness tracker	Fitbit	2 - 5 yr
P2	Female	24	*NA*	*NA*	*NA*
P3	Female	22	*NA*	*NA*	*NA*
P4	Male	57	Fitness tracker	Garmin Instinct	1 yr
P5	Male	26	Smartwatch; Smart ring; Smart scale	Apple Watch; Oura Ring; Withings Scale	> 5 yr
P6	Female	27	Fitness tracker	Apple Watch	2 - 5 yr
P7	Female	29	*NA*	*NA*	*NA*
P8	Female	32	*NA*	*NA*	*NA*
P9	Male	29	*NA*	*NA*	*NA*
P10	Female	42	Fitness tracker	Fitbit	2 - 5 yr
P11	Male	25	Fitness tracker	Xiaomi Mi Smart Band	Few mo
P12	Female	27	Smartwatch	Apple Watch	2 - 5 yr
P13	Female	26	Smartwatch	Apple Watch	2 - 5 yr
P14	Female	61	Smartwatch	Apple Watch	2 - 5 yr
P15	Male	27	Smartwatch	Apple Watch	2 - 5 yr
P16	Female	26	Fitness tracker	Garmin Forerunner	1 mo
P17	Female	29	Fitness tracker	Garmin Vivoactive	2 - 5 yr
P18	Female	32	Smartwatch	Garmin Smartwatch	Few mo
P19	Male	32	*NA*	*NA*	*NA*
P20	Male	26	Fitness tracker	Garmin Forerunner	2 - 5 yr
P21	Female	36	Smartwatch	Apple Watch	2 - 5 yr
P22	Female	30	Smartwatch	Apple Watch	1 yr
P23	Female	23	*NA*	*NA*	*NA*
P24	Female	57	*NA*	*NA*	*NA*
P25	Male	28	Smartwatch	Apple Watch	> 5 yr
P26	Male	28	Fitness tracker; Smart scale	Garmin Forerunner; Garmin Index S2	> 5 yr
P27	Female	28	Smart ring; Fitness tracker; Smartwatch	Oura; Garmin; Fitbit; Apple Watch	2 - 5 yr
P28	Female	23	Fitness tracker	Knauermann Neo 2024	Few mo
P29	Male	31	*NA*	*NA*	*NA*
P30	Male	32	Self-tracking app	Strava	Few mo
P31	Female	26	Smartwatch	Apple Watch	1 yr
P32	Male	28	Smartwatch	Withings ScanWatch 2	2 - 5 yr
P33	Male	26	Self-tracking app	Teamfit; Food-tracking app	Few mo
P34	Female	25	Self-tracking app	Teamfit	< 1 mo
P35	Male	26	*NA*	*NA*	*NA*
P36	Male	25	Fitness tracker; Self-tracking app	Garmin Forerunner; Strava	2 - 5 yr

This diary was optional to fill in. Three questions were included in the diary to stimulate the participants to reflect on their tracker use:

(1) *What observations did you make regarding the WHOOP tracker and its mobile application?*
(2) *What are your thoughts and opinions regarding those observations?*
(3) *How did you use the WHOOP tracker and its mobile application today (or the last couple of days)?*

Participants were encouraged to include any remarks and opinions in the diary.

3.3.2 Pre-Study Interview. The goal of the short pre-study interview during the kick-off meeting was to gain insights into the participants' activity levels and health habits prior to the study, as well as into what kind of prior experience with tracking technologies they had. If they had used tracking technologies before, we inquired about the types of technologies they had used, in what types of situations they utilised the technology, and what expectations they had regarding the technology's functionalities. Additionally, we asked whether or not they had concrete fitness or health goals in mind, regardless of whether they used tracking technology.

Finally, we interviewed them on potential preconceptions about tracking different types of health metrics and artificial intelligence. We asked them if they had used AI before, and if so, in what kinds of situations and applications. An additional goal was to get the participants' opinions on AI, with the expectation of observing a change in their opinions over the course of the study.

The pre-study interview duration ranged from 9 minutes, 44 seconds–59 minutes, 21 seconds (M = 19.40, SD = 16.41). The pre-study interview protocol is provided in the auxiliary material.

3.3.3 Mid- and Post-Study Interviews. The mid- and post-study interviews consisted of two parts. Initially, the participants were asked to elaborate on their thoughts about WHOOP (compared to previously used tracking technologies, if applicable), what kinds of functionalities they were interested in most, as well as in what kinds of contexts they were inclined to use WHOOP. If the participants had the CA functionality turned on that week, the researcher would ask them about their experiences with the CA. These included what kinds of questions the participant would ask the CA, what suggestions the CA would typically provide, and whether and how the participant incorporated the CA's insights into their routine.

In the second part of the interview, the participants were asked about their usage of the WHOOP tracker in more detail to gain more insight into their sensemaking processes in relation to their metrics of choice. Some other questions include how their routine was influenced, how they interpreted those metrics and to what extent these reflected their reality. Both interviews were concluded by asking the participants if the WHOOP tracker offered all the functionalities that they would want and expect from a fitness tracker, and the participants were given the opportunity to make any final remarks.

The duration of the mid-term interviews ranged from a minimum of 10 minutes, 47 seconds and maximum of 91 minutes, 52 seconds (M = 34.04, SD = 16.99), whereas post-study interviews ranged from a minimum of 8 minutes and a maximum of 60 minutes, 53 seconds (M = 36.77, SD = 14.33). We provide the interview guide followed for both the mid- and post-study interview in the auxiliary material.

3.4 Data Analysis

In this section, we describe how we analysed the data collected from the interviews, the daily diaries and the history chat logs. From the pre-study interview, we acquired an overview of the participants' opinions and uses of AI.

3.4.1 Sample Overview: Pre-Study Interview Analysis. To acquire a general overview of the sample's perceptions of AI's added value and their reasoning for using AI, the first author first open-coded participants' answers. Then, the second author coded with those pre-defined categories and added new ones when necessary. These codes were discussed among the first, second and last authors, and revised accordingly. The first author then used affinity diagramming to group the codes. Our results showed that the participants had mainly PRAGMATIC and PRODUCTIVE grounds for using AI. The most prominent PRAGMATIC uses included *writing*, *programming* and *data retrieval*. Regarding PRODUCTIVE reasoning, we found *cognitive offloading*, *efficiency*, and *comfort and accessibility* as the main rationales for using AI. Many participants also expressed that AI had NO ADDED VALUE for them.

We also analysed participants' prior experience with AI. First, the first and second authors coded the data with the categories 'low', 'moderate' and 'high' experience. Cohen's Kappa was calculated to determine the inter-rater reliability (IRR) between the two authors: $\kappa = 0.79$, which suggests satisfactory agreement [44, 74]. The overall percentage agreement between the two authors was 86.11%. The sample had mixed previous experiences with AI. In the auxiliary materials, we provide an overview of how the first two authors coded the data and our calculation of Cohen's Kappa.

3.4.2 Qualitative Analysis of WHOOP Coach Experiences. Qualitative data was collected through the mid- and post-study interviews, the daily diary, and the participant's history logs with the CA. The audio-recorded interviews were transcribed verbatim. Screenshots of all history logs with the CA were taken, also transcribed verbatim.

The interviews were coded with *ATLAS.ti* [48] software. Two authors first coded around 10% of the full dataset to establish a coding tree. Therefore, six mid- and post-study interview transcriptions (three of each), and two daily diary entries, were open coded by the first and second author separately following Blandford et al.'s guidelines [15]. The codes that emerged in the first coding round were discussed among the first, second and last authors. From this discussion, a coding tree was established. The second author then coded five additional interview transcripts and three CA histories. The first author coded the remaining data. Then, the first author conducted thematic analysis to identify the main patterns and themes in the participants' perceptions of the WHOOP Coach and metric usage [15]. Lastly, the first, second and last authors iteratively discussed the results of the thematic analysis to reach consensus on our understanding of the data.

The interviews, daily diaries and WHOOP Coach history logs were all analysed in one batch, composing a shared coding tree for all data sources. This is in line with guidelines on the data source triangulation process, typically referred to as the merging step [30].

3.5 Positionality Statement

The authors of this work have lived, gained education, and worked in Europe for most of their lives, with backgrounds shaped by WEIRD (**W**estern, **E**ducated, **I**ndustrialised, **R**ich, and **D**emocratic) contexts. All authors have academic expertise in HCI, with some rooted in the academic tradition of Computer Science and others in Social Science. They share a general interest in physical activity and approach technology design from interdisciplinary perspectives.

We acknowledge that our positionality has influenced the framing of this research, including our assumptions about user interactions with technology. While we have sought to incorporate diverse perspectives, our lived experiences as able-bodied researchers from European countries shape our interpretations. Recognising the role of cultural and disciplinary backgrounds in shaping research, we encourage further studies that integrate perspectives from regions and communities with different relationships to technology, physical activity, and digital infrastructures.

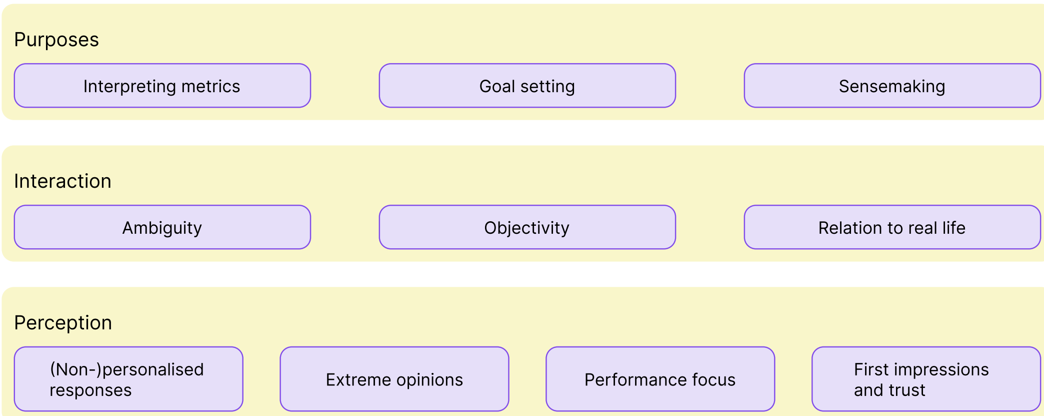


Fig. 4. A schematic overview of the themes and sub-themes constructed based on our qualitative inquiry. As shown in the visualisation, `Purposes` encompasses Interpreting Metrics, Goal setting and Sensemaking. `Interaction` encompasses Ambiguity, Objectivity, and Relation to Real Life. (Non-)Personalised Responses, Extreme Opinions, Performance Focus and First Impressions and Trust are part of the `Perception` theme.

3.6 Participants

In Table 1, we provide an overview of the participants of this study. We recruited $n = 36$ participants. We reached data saturation with our current sample, whereupon we stopped additional participant recruitment, as our qualitative analysis stopped yielding new insights after analysing the collected data from 33 participants. The sample consisted of 21 females (58.33%), and 15 males (41.67%), with an age range of 22–61 ($M = 30.53$, $SD = 9.22$). We recruited participants from eight different European countries, through convenience and snowball sampling. They were either interviewed in the first author's native language or English. The sample consisted of people with a wide variety of backgrounds, life contexts, occupations, health, and sports routines. The sample was also relatively diverse in terms of prior experiences with personal informatics tools, with some having used wearables for several years and others not having used them at all.

4 Findings

In this section, we present the findings of our qualitative analysis based on the daily journal entries, the history logs of the conversations with the chatbot Coach, and interviews. Based on our qualitative inquiry, we constructed three themes: `Purposes`, `Interaction`, and `Perception`, and ten sub-themes: Goal setting, Sensemaking, Interpreting metrics, Ambiguity, Objectivity, Relation to real life, (Non-)Personalised responses, Extreme opinions, Performance focus, and First impressions and trust. The three main themes map to **RQ1**, **RQ2**, and **RQ3**, respectively. The structure of the themes and sub-themes is illustrated in Figure 4. In the following, we describe these themes and their corresponding sub-themes with excerpts from our data sources.

4.1 RQ1: Purposes

Our analysis showed that participants exhibited specific kinds of usage patterns with the Coach. This theme describes how users interacted with metrics, and how this was facilitated through the CA as an interaction style.

4.1.1 Interpreting Metrics. Participants employed three main actions—how they used the metrics—through the CA. Namely, participants queried it through: *Interpreting Data*, referring to what the metric entails, what it is supposed to represent and how to interpret its values; *Contextualising Metrics*, referring to how how to contextualise the data in the context of one's life circumstances and routines, including how measured metrics are translated to the corresponding metric(s) and how to connect these metrics together; and *Interpreting Behaviour*, referring to how the metric is influenced by one's behaviours, and consequently resulting in gaining insight into one's behavioural patterns and progression. We present each of these below.

Interpreting Data. Participants' willingness to consult the CA on how to interpret the metrics typically stemmed from a source of doubt and uncertainty. They were typically unsure about the meaning of the metrics, how to interpret their values, and even how their data was generated. This is evident in P17's uncertainty about how their sleep performance was calculated:

> 'So I felt that like the app was already offering quite a lot of information that I could read. And I also read them more than once. So like when I was not 100% sure how, I don't know, my sleep performance was measured, I was reading the information text about it again and how this is like calculated.
> That's my performance across the day, for example. And I was just hoping that the chatbot would give me more sources to read up on this, which go beyond the app, but it did not offer. So I was just like, "yeah, sure."' (P17, post-hoc interview)

Sometimes, participants would simply ask how a number, score or percentage was calculated:

> 'So I was trying to get more ideas about how to improve my sleep, for example. And I was like, do you have other ideas? Can you tell me a little bit more? I was also using it to get more data into how something is calculated.' (P17, post-hoc interview)

Contextualising Metrics. In order to make sense of their personal data, participants frequently relied on contextualisation. This involved linking different metrics to understand their interconnections and placing them within the broader context of their daily lives. By identifying relationships between variables, they aimed to uncover how one factor influenced another. For instance, P17 used the CA to explore the relationship between their sleep performance and the tracker's estimation of the day's strain goal:

> 'In both instances [when I looked at the app, in the morning after waking up and at night before going to bed], I used the chatbot at the beginning to just make sure that I understood the data correctly, that I can relate them a little bit more. So like, how does my REM sleep phase relate to my sleep performance, to my sleep debt? How does it relate to the app's interpretation of how good my performance will be this day? So I was really trying to understand all these different instances and how they relate to each other. And for this purpose, I was trying to use the chatbot as well to make sense of the data. But it didn't really help that much.' (P17, post-hoc interview)

During one interview, P24 described the Coach as 'enlightening' in multiple ways. They noted, for example, that it explained how the recovery score was partially influenced by the manual journal entries. The following excerpt in P24's history logs illustrates this interaction:

> '**P24:** What behaviors contributed to my Green Recovery?
> **WHOOP Coach:** Hey P24! Here are the behaviors that have positively impacted your Green Recovery over the last 90 days:
> - Magnesium Supplement: Logged 21 times as a "yes" with a significant positive impact of 9 on your recovery.

> These behaviors have been key in boosting your recovery. Keep them up for continued success! Any other questions about your recovery or fitness goals?' (P24, Coach history)

Interpreting Behaviour. Participants often sought to understand which behaviours influenced their metric values, especially when uncertain. In their daily diary, P22 reported that they used the Coach to gain insights into how to achieve their strain goal, by asking about the impact of specific workouts and the reasoning behind it.

> 'Now it seems more doable. I also played around with their AI chatbot which was quite useful to understand the impact of different workouts on my strain goals (e.g., the chatbot broke down which workouts I could include today to achieve my strain goal with workout type and duration; this was really helpful).' (P22, daily diary)

The following example illustrates how participants used the CA to understand their behaviour by identifying specific factors influencing their sleep data:

> '**P24:** What factors affect my deep sleep?
> **WHOOP Coach:** Here are some factors impacting your deep sleep over the last 90 days: [Coach provides some details on the participant's metrics.] Minimizing alcohol and maintaining activities that boost recovery can help improve your deep sleep. How's your routine been lately?
> **P24:** How does magnesium affect my recovery?
> **WHOOP Coach:** The data for magnesium's impact on your recovery wasn't found. Make sure to log it consistently in your WHOOP app to get insights. Are you considering adding magnesium to your routine?' (P24, Coach history)

4.1.2 Goal Setting. One of the most common reasons participants used the Coach was to plan and manage their daily routines, such as sleep and workouts. We observed that participants' planning and management behaviours were supported by both implicit and explicit goal-setting strategies. Progress toward these goals was often facilitated by seeking advice and suggestions on how to achieve them. Within this theme, we observed that participants asked the CA for guidance on establishing more efficient and effective routines to reach their objectives.

At the start of the study, some participants had pre-defined health goals and maintained daily routines to achieve them. Additionally, participants sometimes aimed to meet the goals set by the tracker. For example, in the excerpt below, P22 describes asking the Coach for suggestions on setting up a workout plan in order to reach WHOOP's strain goal.

> 'Now that I added my actual weightlifting routine, I feel like the strain goals become a lot more achievable. In the past days, WHOOP strain measurements made me question my activity level and how much I would need to workout to achieve them. Now it seems more doable. I also played around with their AI chatbot which was quite useful to understand the impact of different workouts on my strain goals (e.g., the chatbot broke down which workouts I could include today to achieve my strain goal with workout type and duration; this was really helpful).' (P22, daily diary)

Similarly, P11 highlighted having used the Coach to perfect their bicycle routine. While P22 pointed towards asking the Coach for advice on achieving the goals suggested by the tracker for the user, P11 had a self-made workout routine and asked the Coach for advice to perfect it. For P11, the Coach was a way to get easy and efficient access to health and sports advice, applied to their current routine.

> 'I remember that I used [the Coach] to plan a one-hour bicycle routine at the gym, which was nice. Then I just made some small adjustments but it was pretty nice. [...]

> And because I had one doubt about the metrics I used, so I asked the WHOOP Coach about this.
> It was something like "give me a one hour bicycle routine in the gym to either burn fat or for cardio," or something like that. To know how to have a routine with the bicycle, so how much time for a certain resistance to be able to do a certain amount of cardio. So effectively planning my usage of time on the bicycle. The WHOOP Coach is faster than Googling it, that would have taken me more time than googling it. I also wanted an accurate response, which is more accurate than just clicking on the first link in Google. Because I also trusted the Coach because it is already tailored to sports and physical activity. Also, I was in the gym when I asked the Coach about a bicycle routine, so that meant the Coach was faster than Googling it.' (P11, post-hoc interview)

In this example, P11 specifically mentioned how they assumed the WHOOP Coach would be more efficient in specifying a cycling routine, considering they were already in the gym, and did not want to risk having to waste time on Googling.

4.1.3 Sensemaking. Our analysis revealed that participants engaged in sensemaking behaviours when interacting with the Coach. A common pattern emerged in these practices, involving two key aspects: 1) identifying patterns in personal data by interpreting individual data elements and their connections, often through past records; and 2) using these patterns to plan and manage routines and individual behaviours to achieve specific outcomes. In the following excerpt, P11 inquired about how the values of the recovery metric, aiming to gain an understanding how the metric gets influenced. This implies that the participant was aiming to get a sense of how to behave in the future to prevent low recovery scores.

> '**WHOOP Coach:** Your HRV is 2% below average, resulting in a yellow Recovery. If you can, spend some time on recovery activities to boost your health. What part of your health data interests you today, P11?
> **P11:** why my recovery yesterday was so low if I slept almost 8 hours?
> **WHOOP Coach:** Hey P11! Actually, your Recovery yesterday was fantastic at 95% (green)! Here are some details: [Coach provides some details on the participant's metrics.]
> You nailed it with 56% of your sleep in restorative stages (REM and Slow Wave Sleep). Keep up the great work!' (P11, Coach history)

Participants also occasionally 'experimented' with the CA to better understand how to interpret specific metrics. For example, P11 tried to assess how to interpret the strain and recovery metrics by responding to the CA's remark on that it needed data on his Bodypump sessions for accurate interpretation.

> '**P11:** If I go to bodypump on a Saturday morning and I end at 11.00, is it enough resting time if I go again on a Sunday at 17.00?
> **WHOOP Coach:** It looks like you don't have any Bodypump activities logged for the last Saturday or Sunday. To give you the best advice, could you log your Bodypump sessions in the app? This way, WHOOP can analyze your Strain and Recovery data more accurately.
> **P11:** Does it depend on the strain I've achieved and the recovery I've met?
> **WHOOP Coach:** Absolutely! Your ability to handle another Bodypump session depends on both your Strain and Recovery. [...]

> With a yellow recovery on Saturday and green on Sunday, you seem well-prepared for another session by Sunday at 17.00. Just listen to your body and monitor your Recovery and Strain!' (P11, Coach history)

4.2 RQ2: Interaction

In this section, we describe the patterns in participants' interactions with the CA, accompanied by a recurring issue with the CA: a lack of personalised advice and an abundance of generic responses from the CA.

4.2.1 Ambiguity. When phrasing a question, participants benefitted from removing ambiguity from their prompts by asking direct and specific questions. This helped eliminate double meanings and led to clearer responses. For example, questions like *'What are my best recovery strategies?'* and *'How does magnesium affect my recovery?'* (P24, Coach history) are examples of concise and straightforward prompts.

When asked about their CA usage, P28 noted that the CA's responses were somewhat generic. They implicitly mention how this could have been attributed to the casual nature of their inquiries, such as asking about their day or possible areas for improvement. However, this also highlights that participants may have different expectations of specificity, as some participants may find such details sufficiently specific.

> 'Like, it says, "oh, your recovery was", I don't know, "brilliant" or something. Today is your level... Your aim for strain is this and this number, so... And I'm like, "yeah, hmm, I'll try it."
> But it wasn't that I set it as my aim to be exactly like the thing it said. But it was interesting to see what I can achieve in a day. [...]
> Most definitely I ask how the day was, what I could have done better. Yeah, that's basically... The things I ask. [...] Like, they weren't that specific. But like, overall, the overview was... That was good.' (P28, mid-term interview)

Remarkably, P24 was one of the few participants who leveraged the Coach's default question options, eliminating the need to generate their own questions. They actively engaged with the CA, applying specificity in their responses, which, in turn, led to more tailored (and therefore, 'specific') default question options aligned with their interests.

> 'You automatically arrive in the chatbot programme and then you get to see follow-up questions, so [questions like] "what is resting heart rate?" I would then sometimes ask. Then it would of course automatically come up with things that is in such a database, where they search for certain text [pieces] in the question, and then you can go over all those follow-up questions.
> Sometimes in the text, things are explained, and then based on that you get more follow-up questions. I found the Coach quite enlightening. There were also videos and tutorials, so then you have all these things that you can look at, and you can also surprisingly read articles.' (P24, mid-term interview)

4.2.2 Objectivity. Unlike P24, P26's experience with the CA's default question options was different, as they engaged less with the system. Without frequent follow-up questions or shared personal information, the CA's responses remained more general.

In the following excerpt, P26 reflects on why they did not feel the need to respond to the CA's daily overviews:

> 'I never asked the chatbot anything, I always [only] looked into like, the recommendations.

> I think sometimes at the overview... I think the first notification that it sends you is "get my outlook." And then you get like the summary thingy. [...] But I looked into it, but then I never felt like... I felt it was like so superficial or so basic, that I never felt the impulse to ask follow-up questions because I felt like it's not gonna help me much.' (P26, mid-term interview)

In some cases, participants engaged minimally with the Coach from the start. For instance, P29 mentioned that they generally rarely used the app as it was 'too crowded', which suggests that an overload of data might have been a limiting factor. Although they noticed the default question options provided, they did not find them compelling enough to sustain ongoing interaction.

> 'The only thing that I missed, I think it's like this beta [functionality] asking anything. Probably I would use it more if I notice that there is something interesting.
> Now, when I go into the app, I see that there are actually some [default] questions that you can ask. So, it's actually on my part that I should go into [it], but I didn't go into the details of this. So, yeah, I think... Yeah, again, the app is too crowded, too complicated.' (P29, post-hoc interview)

This finding suggests that the CA has the potential to help users identify behaviours that positively contribute to their routines and health, but its effectiveness depends on regular and consistent engagement with the tracker—and the Coach itself.

4.2.3 Relation to Real Life. Interestingly, having a consistent workout routine did not determine whether participants liked or disliked the CA. Sporty and active individuals expressed both positive and negative opinions, often shaped by other factors. In the case of a negative or neutral opinion, they typically argued something along the lines of: *'Nobody knows my body better than I do. I already possess the knowledge (from years of training) on what to do to improve my performance or health; I don't need a chatbot to tell me that'*, quoting P4. They perceived the CA as offering little added value to their well-established routines. Below, we quote P26, a triathlete, on this perspective:

> 'I did [look at the CA] and sometimes I looked into them, but I felt mostly that the information was sort of too basic. Nothing like super added value. In the sense of, like, when it analyzed my sleep and then it told me "today is going to be a hard day, you should hydrate more". And it's like "okay thanks, I know that already". Plus I also already used the... I was testing, I think I wrote it in a journal. I think Strava also integrated some sort of like AI feature to analyze the training.
> And that's sort of what I was looking into and I feel like it fit more into my workflow of like actually looking [into an analysis of my training].' (P26, mid-term interview)

Notably, P26 compared their experience with the WHOOP Coach to Strava, which they used to track their training. In the same interview, P26 explained how Strava's AI feature offered more abstract interpretations of their runs, providing session summaries through comparisons with their own baselines.

Concurrently, several physically active and health-focused individuals recognised the benefits of the Coach. For example, P11 and P22, who have a consistent workout routine consisting of weightlifting, bodypump, tennis, padel, running, and volleyball, found the CA useful for setting up training schedules (P11).

When asked why they used the CA less frequently, participants with less physically active lifestyles often felt 'uninspired' about what questions to ask. Many remarked that they struggled to come up with questions for the CA, as they did not have concrete health or exercise goals. In the words of P12 (post-hoc interview): *'I don't know. I didn't feel the need to use it, let's say.'* Several participants also specifically stated that the chatbot could be useful for users with more 'specific questions and goals' instead (e.g. P1, P5).

4.3 RQ3: Perception

This theme discusses participants' perceptions of the CA, focusing on its ability to provide personal or impersonal responses and its performance focus. We also present users' most common reactions, including extreme opinions, first impressions, and the CA's impact on their trust.

4.3.1 (Non-)Personalised Responses. A common concern among participants was that the Coach would provide 'generic', 'unhelpful' and 'impersonalised' responses to their prompts. Many emphasised that the Coach provides 'unclear' (health) suggestions or that it could not answer questions when asked for specific details.

> 'I have to say that the chatbot was also rather superficial. So, not really helpful. I mean, interacting with it was like with any other chatbots as well. But I felt that just by Googling and by asking ChatGPT, I would have had better results that were maybe also more diverse because I feel that the chatbot is based so much on the data that is already gathered in the app that it does not really provide more ideas or that they offer like other links where you can click on and where you say like, oh, now I have like this, I don't know, this podcast or this review where some of your internal workings from your body would get explained to you more. So I was just missing that if I really want to dive really deep into one topic, that it would not provide me the means to actually do so. And yeah, in the end, I mean, I even did Google it, how to improve my sleep pattern.' (P17, post-hoc interview)

In P10's Coach history, we found that they asked the coach how to improve their current sleeping routine (*'How can I improve my sleep quality?'*). The Coach initially responded with a list of suggestions in bullet points. Next, P10 followed up on one of the points (consistent bed times), prompting the CA to provide another bullet-pointed response elaborating on consistency:

> '**P10:** Am I sticking to consistency?
> **WHOOP Coach:** Great choice! A consistent sleep routine can do wonders. Here are some tips to stick with it: [CA provides some details on the participant's metrics.]
> Your goal: Get to bed by 10:30 PM and wake up by 8:00 AM. Let me know how it goes or if you need further support!' (P10, Coach history, translated from P10's native language to English)

4.3.2 "Extreme" Opinions. Overall, participants had mixed perceptions of the CA Coach. Some ignored it, while others found it useful. In our sample, opinions tended to be polarised, but the reasoning behind them varied. The ignorance of the CA could potentially suggest that it is not clearly designed in communicating its functionality and added value. Many users would only respond once to the CA's initial 'daily overview' message, sometimes not initiating any conversation at all (e.g. P24, P26). For instance, P26 mentioned not feeling the need to respond to the CA's daily overviews. Notably, many participants actively ignored the CA despite having noticed it on the home screen. Their reasons varied, but the most common were: 1) a general dislike of AI or CAs, 2) making premature or incorrect assumptions about the CA, and 3) uncertainty about what to ask or how to engage with it.

Many participants expressed a lack of interest in using the WHOOP Coach due to scepticism towards AI and CAs (P5, P6), particularly in relation to health data (P5). P29 voiced feeling confused about what to ask the CA and how to interpret the default question options.

> 'I didn't know what to ask or what to talk about with the application for monitoring. And then, yeah. So, I didn't know... how should I interact with the chatbot? [...]
> Usually, if I want to use, for example, ChatGPT or something, usually I have a question or inquiry that I want to know more about. So, if it had like some kind of... "ask me these, and then I can answer them", then I would be more motivated. Maybe one of those questions about what I want to ask.
> But I saw it, but I didn't know what this is about. So, I just discarded it.' (P29, post-hoc interview)

Some participants reacted strongly to the implicit assumptions embedded in the CA's functionalities, such as its focus on increasing performance, as illustrated in the next excerpt. This emphasis on performance is further explored in the following section (subsubsection 4.3.3).

> 'So I was asking the chatbot, I was trying to read all the stuff, how it gets calculated and so on. So I really spent a lot of time engaging with the data at the beginning.
> "Oh, yeah, but today you are not going to be on top performance and today you have a sleep debt of like 15 hours." Yeah, but it's not helping me to reduce this. So it just made me mad.
> It made me furious. It made me sad. It made me just like doubting or like whatever I have learned so far and whatever my body is doing.' (P17, mid-term interview)

4.3.3 Performance Focus. Some participants felt that the CA adopted a performance-based perspective, prompting a range of reactions. P17, for instance, attributed WHOOP's negative framing to its emphasis on continuous (improved) performance. Other participants also voiced this opinion, like P1.

> 'And I just felt that the whole interface and the whole thing that it told me like, "oh, yeah, you're sleep debt and your body battery is running low". And "oh, yeah, you are not at top performance"... it just initiate a feeling that I am doing bad, right? And I think this is the whole thing: how it is presented. Because sleep debt is something very negative, right? And just the name of it is also very negative. And then when it tells me like, "oh, yeah, you will not be in top performance", it reduces my motivation because I'm looking at the data in the morning.' (P17, mid-term interview)

Some participants noted that WHOOP may evoke negative feelings when indicating that a user is not 'performing well' according to their metrics (e.g. P6), or potentially not reaching their goals.

Conversely, some participants found the CA's functionalities helpful (e.g. P11, P22, P24). For example, P22 commented that the CA helped them gain useful insights into achieving specific health goals and described it as convenient.

> 'It wasn't a pop-up, but it was there. It's just telling you, "hey, this is your strength goal for today". And then I had the chatbot [...], that was super nice, because then I could also just ask, "okay, this is my strength goal, what would be workouts that would be helpful for me to achieve this?" And then it would just tell me, "hey, if you go for a 40 minute run, you would already be at a strain level of 9.8 out of the 12 that you want to have." So that would be something super easy.' (P22, mid-term interview)

Notably, P22 was also one of the few participants who explicitly stated in their daily diary that they noticed the CA's absence in their post-hoc interview: *'I don't have the chatbot anymore which frustrates me.'*

4.3.4 First Impressions and Trust. Several participants were unsure about the CA's added value and functionalities, often making assumptions about its capabilities. This would often discourage participants from using it or even trying it. P1 chose not to engage with the CA, assuming it would not be useful if they did not have a specific health goal.

> 'I was just curious what it would tell me. Like, I opened it once and then it was like "oh yeah, something something with my HRV and whatever"... And then I was like "okay, anyways, that value might be skewed so..." and yeah, and then I didn't continue chatting, but I thought it would be quite cool if it was really very personalized somehow.' (P27, post-hoc interview)

The quote above illustrates how the success (or failure) of the first interaction with the CA is crucial to the user's overall impression. P27, for instance, expressed disinterest, citing a lack of personalisation (subsubsection 4.3.1).

In line with this, P9 assumed that AI tends to be more 'positive' by default, influencing their perception of the Coach's reliability.

> 'And asking too much when you have, like, other struggles... and I tried to do that a little bit with the WHOOP Coach but it did not feel attached enough to my individual perspective. And because it's like AI, I am never really sure if it's accurate enough. Because AI always tends to give it a more positive spin than it actually is, because it doesn't want to demotivate you. So it never really feels super reliable to me.' (P9, mid-term interview)

Additionally, P17 specifically mentioned how a lack of interpretation and contextualisation of the data (see subsubsection 4.1.1) resulted in them distrusting the CA, believing that it would not be able to provide valuable insights.

> 'But if the information does not help me from before and it just does not fit to my own personal felt experience, that is like where I mistrust everything related to WHOOP, which is the data gathered a little bit, the chatbot, which only refers back to this data and this information that have already read. And this is why I needed to go outside of this bubble to trust the results more.' (P17, post-hoc interview)

5 Discussion

Our two-week user study with participants using the WHOOP fitness tracker and its integrated CA allowed us to gain insights into how users interact with the CA, and specifically, the purposes they had for interacting with data (**RQ1**) and the nature of these interactions in practice (**RQ2**). Additionally, we gained insights into our participants' impressions of the CA in light of these purposes and interactions (**RQ3**). Finally, we derived concrete design considerations for CAs in PI tools to enhance their effectiveness and usability as an interaction style (**RQ4**).

5.1 'Give and Take': The Trade-Offs of CAs in PI Sensemaking

Our study on health recommendations provided by a CA raises questions about the usefulness of the abstraction of raw data in PI systems. As previously discussed, WHOOP utilises a range of metrics, encompassing both objectively measured and derived metrics, following the definitions of Bentvelzen et al. [11]. This diversity underscores the need for a theoretical framework for designing such metrics in alignment with users' life contexts and prior experiences.

The CA explored in this study functioned as an alternative option for interpreting and interacting with PI data. Our findings highlight how participants used the CA to engage with specific metrics, summarised with the `Purposes` theme, and how the CA supported interpretation (subsubsection 4.1.2), goal setting (subsubsection 4.1.3), and sensemaking (subsubsection 4.1.1), ultimately supporting users in planning and managing their health routines. However, the ways in which users approached sensemaking, goal setting, and data interpretation varied substantially. The sub-themes within Purposes are not exhaustive; rather, they were often intertwined rather than isolated, influencing each other dynamically. For example, in P11's case, an activity categorised as 'interpreting data' also exhibited elements of 'interpreting behaviour', demonstrating the complexity of the sensemaking process (**RQ1**).

Our findings indicate that CAs have the potential to serve as an efficient and effective sensemaking tool, with some participants finding the WHOOP Coach convenient, insightful, and personalised. However, unlocking this potential was not straightforward. In reality, most participants were underwhelmed by the CA's performance. The primary issue was its generic and impersonal responses, as described in the Interaction theme (**RQ2**). Specifically, the Ambiguity theme (subsubsection 4.2.1) highlights how vague and ambiguous response prompts often resulted in uninformative and unclear outputs. Our findings demonstrated that users faced a dilemma: if the CA made too many assumptions about the user, it risked generating 'hallucinations', erroneous inferences based on incomplete understanding, as observed in LLMs [53]. In the WHOOP Coach's case, it sometimes made unwarranted assumptions about a user's routine planning, even when they had not selected a specific training or recovery plan. Conversely, if the CA made too few assumptions, it provided overly generic and impersonal responses, lacking meaningful insights. This balance was explored in the Objectivity theme (subsubsection 4.2.2). Additionally, Relation to real life (subsubsection 4.2.3) captured users' ability to contextualise and integrate data into their daily routines, which is an essential aspect of successful sensemaking (**RQ2**). When the CA failed to support these interactions effectively, users perceived it negatively, as detailed in Perception. We discuss this aspect in more detail below (subsection 5.2).

Importantly, the challenge of balancing personalisation with objectivity is a well-documented issue in CA design [76, 148], and our findings reinforce these concerns. At the same time, excessive anthropomorphism can create feelings of unease and undermine trust [33, 92], while insufficient personalisation leads to disengagement, as was demonstrated by our findings. Based on the analysis of our collected data, we propose what we call the *give and take* principle: we argue that CAs are particularly prone to negative user perceptions when they fail to balance user engagement with the CA's responsiveness. The 'giving' aspect refers to the user actively engaging with their data and providing meaningful prompts, while the 'taking' aspect refers to extracting valuable insights from the interaction. Our findings suggest two key factors affecting this balance: 1) a lack of specificity when framing the prompts, and 2) insufficient prior engagement with the personal data (**RQ2**). This aligns with previous research on information retrieval in CAs, where Zhang et al. [148] found that users of news-oriented CAs reported usability limitations, including inefficiency and lack of effectiveness. Similarly, Mildner et al. [92] emphasised the importance of managing users' expectations to facilitate successful interactions. Our findings suggest that user expectations must be aligned with prompt quality – essentially, *what you give is what you get* (**RQ1, RQ2**). Ultimately, this tension reveals a clash between participants' purposes for using certain metrics (*why* they use them), requiring a degree of specificity and input, and the CA's inability to realise those purposes and failing to meet these expectations due to overly generic responses, limiting its usefulness for meaningful sensemaking (**RQ1, RQ2**).

5.2 Contrasting Views and Long-Term CA Engagement

Our study demonstrated that the purposes and interaction styles of the users influence their eventual perception of the CA. To this end, the theme Perception (subsection 4.3) describes how the overall experiences with the CA is dependent on the context of their sensemaking process. In general, we found that CAs in a PI tool do not make positive or memorable impressions. Remarkably, users' perceptions could be broadly categorised as the following: very positive or very negative, with little middle ground (subsubsection 4.3.2). Our data suggests that these extreme opinions were often shaped by users' initial impressions, which directly influenced their trust in the CA (see First impressions and trust, subsubsection 4.3.4). A key factor behind these polarised reactions was users' difficulty in understanding the CA's functionality and limitations, leading to incorrect assumptions about its capabilities (**RQ3**).

In traditional PI tools, personal data is typically represented through visual formats such as graphs and numerical summaries. However, alternative interaction styles with data have also been explored, including ambient light-based feedback (e.g. Crimson Wave [45] and Ambient Cycle [56], for menstrual cycle tracking), data physicalisation (e.g. LOOP [118] and SweatAtoms [65]), and even tangible objects representing bodily functions(e.g. the Curious Cycle [21]). While traditional data representations are generally easier to interpret, new ways to interact with personal data, like interacting with the WHOOP Coach, introduce unfamiliarity and abstraction, requiring users to adapt to new ways of making sense of their data. In addition, they may adopt different kinds of interaction styles. We argue that a CA represents yet another unique form of data representation—one that is narrative-driven, relying on multidirectional textual interactions. This shift from visual to textual modality may explain users' confusion and initial uncertainty, similar to how users respond to other unconventional PI data representations (**RQ3**).

Further complicating user engagement, participants frequently described the CA as too focused on performance (Performance focus, subsubsection 4.3.3) and not personalised enough ((Non-)personalised responses, subsubsection 4.3.1). This aligns with prior research in the PI community, where users have noted that self-tracking devices often embed implicit assumptions about their goals and behaviours (e.g. [120]). Not unfamiliar to PI research either is that users perceived that the model tended to provide suggestions related to the improvement of performance and health. This could suggest that WHOOP's LLM model is likely a non-PI model, meaning that the model has not been trained on PI data. This presents a clear research gap in CA research in PI contexts, also previously pointed out by Strömel et al. [127]. Developing a dedicated PI LLM model could potentially offer more relevant interactions for PI system users.

On a similar note, we observed that the CA's suitability varied based on user characteristics. Generally, physically active users were more likely to find value in the CA, yet this user characteristic was not a guaranteed predictor of positive engagement. As discussed in subsection 5.1, the CA's effectiveness still depended on factors such as how well users framed their prompts and engaged with their data. This finding is fairly unsurprising, considering participants observed a bias in the model towards performance, which is arguably more aligned to the needs and goals of users who already prioritise fitness (**RQ3**). We speculate that when this alignment occurred, users experienced an 'illusion' of personalisation, where the CA's responses appeared tailored to their goals, even when they were actually generic.

These findings further reinforce the relevance of the *give and take* principle. Users who engaged more actively with their data, by providing clearer prompts and interacting with the CA more consistently, were better able to extract useful insights. Likewise, physically active users inherently generated more data for the CA to interpret, which in turn improved the relevance of its responses. Similarly, users with well-defined goals were able to structure their queries more effectively. This suggests that while CAs in PI tools may not be universally beneficial, they do hold potential for specific user groups who can leverage their functionality more effectively (**RQ3**).

5.3 Design Considerations for CAs in PI Systems

Based on our analysis, we address **RQ4** in this section. We present considerations for design and research that guide further exploration into how CAs can help drive the PI experience to benefit user wellbeing.

Design Consideration 1: Navigating Giving and Taking—Sharing Data with the CA Benefits Long-term Sensemaking. Firstly, as discussed in subsection 5.1, we introduce the concept of a *give and take* interaction with a CA in PI, suggesting that users can only benefit from accessing their data through a CA if they provide the CA with additional data. At the same time, users may not always use the opportunity to extract useful information from the CA—the taking. This may be due to biases on the user's end, bad first impressions, or biases in the CA model.

Although some sensitive attributes are increasingly taken into consideration in Machine Learning (ML) systems for mobile and wearable computing, like gender, age and physiology, other biases related to race, nationality and language may still persist within these systems [141]. Yfantidou et al. [142, p. 17] reports that in PI, there is still 'significant underrepresentation of minority groups across all protected attributes and measured behavioral differences—not necessarily realistic—for users with diabetes, joint issues, unhealthy BMI, non-white users, and females.' Such biases in PI systems can have detrimental consequences, including delayed diagnoses or even misdiagnoses [98], reinforcement of stereotypes [85], perpetuation of health disparities [89], and lack of healthcare access and resources [123]. It has been reported that GPT-4 displays racial, cultural and gender biases (e.g. [46, 129, 143]), showing potential for such biases to manifest in the WHOOP Coach system behaviour as well. Given that it is unknown what data ChatGPT—and by extension GPT-4—is trained on, it is hard to uncover where potential biases could originate from [117].

Given the prominent occurrence of negative opinions about the CA, there is a need to gain a better understanding of the power of first impressions and whether more long-term engagement with PI data changes users' pre-formulated notions of the new interaction style. This is particularly pertinent given that participants were often discouraged from using the CA due to pre-defined assumptions, thus refraining from interaction. This suggests that there is a need for future systems to clearly communicate potential benefits and ways of interacting with CAs so that the benefit of long-term sensemaking with the CA becomes apparent to users. Users should also be made aware of potential biases the LLM may have, especially in the case of long-term usage. Therefore, future CA designs should consider informing the user both of the benefits of increased and longitudinal engagement, and potential risks.

Design Consideration 2: CA Design Should Explicitly Focus on the Interpretation of Derived Metrics. One of the CA's core strengths is that it offers another approach to interpreting complex metrics (like derived metrics [11]). CAs have the potential to offer human-like support. More so, a CA's response can be 'tailored' according to the user's needs and expectations, and therefore offering a gateway towards customised support. Agapie et al. [4] remark that receiving support from others can be considered a form of 'customised support', given the empathic support [19], the opportunity to get to see things from a different perspective [93, 99], and the opportunity to get to know ourselves better through others [77]. However, comparisons can yield benefits or drawbacks to the PI journey. For instance, excessive comparing can result in negative thought cycles [95], an overly critical mindset [82], and extreme competitiveness [95, 108]. At the same time, it presents opportunities for positive social interactions [103, 130] and aiding the sensemaking process [58]. This highlights that CA design should focus on providing accurate, contextualised explanations of tracker metrics, while preventing the user from overly comparing themselves to suggestions the CA makes.

Design Consideration 3: Managing Ambiguity as a Key Design Dimension. In subsection 5.1, we argue that ambiguity in prompts is not merely a challenge but a crucial design dimension for future CA in PI. Our analysis highlights that achieving an appropriate level of ambiguity is essential for balancing specialisation and generalisation in CA-generated responses. Rather than assuming a fixed optimal level, designers must account for how ambiguity shapes interaction, influencing whether responses are tailored or broadly applicable.

In related literature, Gaver et al. [47, p. 233] propose that ambiguity can compel users to 'interpret situations for themselves, it encourages them to start grappling conceptually with systems and their contexts, and thus to establish deeper and more personal relations with the meanings offered by those systems.' More concretely, Turmo Vidal and Duval [132] identify three interactive design qualities to foster ambiguity as a resource to cater for the plurality of bodies: 1) *versatility*; 2) *appropriation*; and 3) *adaptability*. Respectively, these qualities engender 1) meaningful usage for multiple purposes; 2) user co-construction and participatory design practices of the interactive experience [57]; and 3) dynamic use and sensemaking processes according to varying contextual factors [136, 137]. These recommended design elements are in line with the results of our study, as participants related their CA interactions with real life and made sense of their data through the CA in a variety of manners.

Future CA designs should integrate mechanisms that help designers manage ambiguity in prompts. This includes offering structured ways to shape prompt formulation while maintaining transparency about how different levels of ambiguity affect outputs. Managing ambiguity effectively in design aligns with findings in psychology, particularly construal level theory [81], which suggests that varying levels of abstraction influence interpretation and decision-making.

Design Consideration 4: Managing the CA Relationship Over Time. While this study spanned two weeks, the question remains how interactions with a CA evolve over extended periods of use. Sustained engagement and the capacity to support long-term user needs present a critical design challenge. Participants were concerned about whether CAs can effectively support ongoing, evolving relationships rather than isolated interactions. For instance, the Coach discouraged participants from engaging if it either made too many or too few assumptions. We observed a lack of faith in the CA's ability to be flexible and change its responses according to a given input, suggesting it may present itself as overconfident and persistent to the user. Arguably, AI's power is in its ability to connect pieces of information together, upon which it can derive 'tacit' knowledge [7]. Although confident AI can result in increased human performance (e.g. [5]), this seems to bear less significance for health applications where there is no one-size-fits-all answer.

Our findings suggest that many participants were unable to fully leverage the CA's capabilities over time. To address this, future CA designs should incorporate mechanisms that actively scaffold long-term engagement. This may involve periodic prompts encouraging continued interaction, delivered through notifications or other attention-directing design elements. Additionally, instead of relying solely on a fixed set of default prompt suggestions, CAs could display its capacity to be flexible by dynamically adjusting the types of recommended interactions, fostering sustained engagement and adaptation to evolving user needs.

Design Consideration 5: Addressing the Embedded Performance Focus in Commercial LLMs. Current commercial CA implementations, such as WHOOP's use of OpenAI models (though the specific model is not disclosed), appear to embed a strong emphasis on performance. This inclination towards performance-framed responses can be problematic, as it risks steering individuals towards behaviour that prioritises optimisation over their actual intentions. While CAs integrated into PI tools often frame interactions around improvement, this does not mean they should exclusively promote performance-driven goals. De-Arteaga et al. [31] pose that algorithmic aversion and automation bias are two extreme phenomena, whereby algorithmic aversion degrades the effectiveness and quality of decision-making through overreliance on algorithmic recommendations, whereas automation bias is the principle of excessive overriding. On the one hand, the Coach's lack of personalisation poses particular challenges for users who over-comply, especially given this embedded performance focus. On the other hand, participants' complete non-compliance was often a result of their taking notice of this shortcoming. *How do we design CAs that endorse a healthy degree of trust?*

To ensure that CAs in PI support broader aspects of wellbeing, custom or fine-tuned models are needed. Rather than imposing an implicit focus on performance, future CA systems should dynamically align with the goals users disclose and adapt to their behavioural patterns. This would help prevent misalignment between user expectations and the system's responses, fostering a more supportive and personalised interaction. PI systems have the potential to encourage self-compassion rather than solely reinforcing improvement-focused narratives, as highlighted by prior work on designing for self-compassion in PI [84].

5.4 Limitations and Ways Forward

This study comes with several limitations, which we acknowledge here in this section. First, we used convenience and snowball sampling for this experiment, resulting in a non-randomised sample in which participants may share certain commonalities. The research was conducted in the European Union and included participants from various countries and cultural backgrounds. However, the sample primarily consists of adults in their twenties and thirties.

To mitigate potential biases, we aimed to include participants with diverse gender identities, educational backgrounds, levels of experience with tracking technologies, and degrees of physical activity in their routines. Despite these efforts, we acknowledge that our sampling methods may have influenced the results. Cultural differences shape how users perceive and interact with technology, and regional variations in technological advancement affect familiarity with digital tools. In line with this, we asked participants about their general opinion of AI and their prior experience with it; but overall, a large percentage of our sample was familiar with the technology or had incorporated it into their routines, which we attribute to the recruitment of mainly young adults with relatively high technological affinity. We recognise that previous use of CAs and AI may have influenced our sample's expectations and interactions with the Coach. While we sought to account for factors such as gender, education, occupation, and upbringing, the study does not fully represent perspectives from regions with significantly different technological infrastructures or cultural attitudes toward AI and CAs. Future research should include a more diverse participant pool to improve the generalisability of these findings.

Furthermore, the within-subjects design of the experiment—whereby participants were exposed to the CA in one week whereas the CA was concealed in another—may have primed participants to notice the difference between the two weeks, hence influencing their responses in the interviews. Moreover, due to the exploratory nature of this study by making the purposeful decision to not prone users towards using the CA, but rather let them explore the tracker naturally, may have resulted in fewer participants getting the 'chance' to interact with the CA. On the one hand, some participants may have started using the CA eventually if given a longer use period. On the other hand, we argue that a lack of interest to use the CA is an indication of their long-term needs and tracking behaviour, resulting in valuable insights nonetheless. Consequently, this was also reflected in the participants' reasoning for not using the CA, with participants providing a multitude of arguments and opinions.

WHOOP did not support the deletion of personal data from the account; in other words, to "start from scratch" by collecting new data without history of the old data, without having to delete the entire account and subscription. This proved to be an obstacle as participants reused the trackers in the study. As such, participants could not start their two-week experiment without the WHOOP being 'influenced' by prior data collected by other participants. Considering the WHOOP adapts, i.e., the recovery and strain metrics according to history data, the participants' metrics will have been construed according to others' data, and hence, their perceptions of their data may have been influenced accordingly. This may, in turn, influence the generalisability of our findings, as the conditions in which the study took place were not perfectly identical for each participant. Some participants were given a tracker that had not been used as much by others before, whereas others did. However, although a few participants noticed this, the overwhelming majority of participants were either not aware of this or were not bothered by it. In addition, the overall goal of the study was to get users' perceptions and experiences with a CA, which may be partially influenced by the 'accuracy' of the data, but not entirely. Future PI research may also consider recruiting larger samples and conducting studies of longer duration. Nevertheless, it is not uncommon for empirical PI research to have smaller sample sizes than ours (see [49, 54, 55, 62, 84]).

Third, the two-week period may not have been a sufficient amount of time for the participants to get to know all the features of the WHOOP app. Although we aimed for participants to have an as natural experience as possible, as we did not nudge them towards using particular functionalities, a two-week period cannot capture the full picture of what long-term wearable interactions and reflections look like in a variety of contexts. Overall, a short period of time cannot grasp all lived experiences with long-term CA interactions in combination with derived metrics. Yet, we decided on a two-week study and not longer, as it has been argued that shorter diary studies result in more active engagement (e.g. [40, 41]). In the past, it has been reported that the median duration of (health) diary studies is 17 days [61], with one to a few weeks being the most typically chosen duration for a diary study [122], showing that two weeks is a reasonable middle ground. Nonetheless, the PI community needs longitudinal in-the-wild studies of longer duration to assess how CAs as interaction styles influence PI behaviour and data perception over time. Notwithstanding, we collected data until we reached saturation, showing that the findings of our study bear some degree of generalisability (e.g. [17])—or at least *transferability* of knowledge [91]—for future CA designs in PI systems.

Lastly, the WHOOP tracker is a commercial product that regularly gets software updates. Since the study was done over an extended period of time, the app was updated several times in terms of interface design and (accessibility) of functionalities. We recommended participants to allow the updates to ensure the app would continue running efficaciously throughout the study. For instance, at the start of the study, the step count metric was not readily available on the homepage, while it was offered as a beta version in October 2024. This introduction of the metric is also reflected in participants' answers, as some who participated early on in the study were confused, or even annoyed, by the absence of step count, whereas participants recruited at a later time would not voice this opinion. Although this introduces some differences in opinions and observations of the app's functionalities, the main goal of the study was to get a broader understanding of participants' overall perceptions of the app's derived metrics and CA functionality. Hence, we infer that these updates will likely have had little effect on participants' experiences.

6 Conclusion

In this paper, we studied the user experience of a text-based Conversational Agent (CA) in the WHOOP fitness tracker, called the *WHOOP Coach*. We conducted a qualitative study with $n = 36$ participants to capture user experiences with the integrated CA, paying special attention to how the users' understanding of fitness tracker metrics was affected by the CA. The analysis revealed that receiving generic, impersonal responses was a common issue. We attributed this to the *give and take* principle, affirming that participants were a lot more likely to receive informative suggestions and tips from the CA with increased engagement, and that specific question framing associated to one's behaviours were more likely to produce useful responses. The *give and take* principle showcased that CAs for PI sensemaking should aim to strike a balance in their design: to promote users to sufficiently engage with the tracker (*give*), and to intuitively inform users on how to effectively use a CA as an interaction style to be able to acquire meaningful insights from it (*take*). We provide design considerations and critical insights on the integration of CAs in Personal Informatics tools. This work contributes to our understanding of the sensemaking process of metrics in self-tracking tools and guides future potential use of CAs in PI experiences.

Acknowledgments

We would like to thank several funding bodies that supported our research endeavours described in this paper. Specifically, this work was supported by the Swedish Research Council with award number 2022-03196, and Paweł W. Woźniak is supported by an endowment from TU Wien. Further, this research was supported by the Leibniz ScienceCampus Bremen Digital Public Health, which is jointly funded by the Leibniz Association (W72/2022), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology – BIPS.

We would like to express our gratitude to all the participants, for their effort and enthusiasm to take part in our research. We want to thank Tim Johansson for his help with the instruction video employed in the study. Also, we appreciate the help from Ayça Ünlüer and Tora Jarsve with the design of the figures.

References

[1] Rikke Aarhus, Stinne Aaløkke Ballegaard, and Thomas Riisgaard Hansen. 2009. The eDiary: Bridging home and hospital through healthcare technology. In *Proceedings of the 11th European Conference on Computer Supported Cooperative Work* (Vienna, Austria) *(ECSCW 2009)*. Springer, London, UK, 63–83. doi:10.1007/978-1-84882-854-4_4

[2] Alaa A. Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: Scoping review. *Journal of Medical Internet Research* 23, 1 (2021), e17828. doi:10.2196/17828

[3] Gregory D. Abowd and Russell Beale. 1991. *Users, systems and interfaces: A unifying framework for interaction.* Vol. 91. Cambridge University Press, Cambridge, UK. 73–87 pages.

[4] Elena Agapie, Patricia A. Areán, Gary Hsieh, and Sean A. Munson. 2022. A Longitudinal Goal Setting Model for addressing complex personal problems in mental health. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2, Article 270 (Nov. 2022), 28 pages. doi:10.1145/3555160

[5] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining human expertise with artificial intelligence: Experimental evidence from radiology.* Technical Report w31422. National Bureau of Economic Research. doi:10.3386/w31422

[6] Felwah Alqahtani, Ghazayil Al Khalifah, Oladapo Oyebode, and Rita Orji. 2019. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. *Frontiers in Artificial Intelligence* 2 (2019), 30. doi:10.3389/frai.2019.00030

[7] David Autor. 2024. AI could actually help rebuild the middle class. Retrieved March 21, 2025 from https://www.noemamag.com/how-ai-could-help-rebuild-the-middle-class/

[8] Eric P.S. Baumer. 2015. Reflective informatics: Conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI 2015)*. Association for Computing Machinery, New York, NY, USA, 585–594. doi:10.1145/2702123.2702234

[9] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The development and validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 366, 8 pages. doi:10.1145/3411764.3445673

[10] Marit Bentvelzen, Jasmin Niess, and Paweł W. Woźniak. 2021. The Technology-Mediated Reflection Model: Barriers and assistance in data-driven reflection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 246, 12 pages. doi:10.1145/3411764.3445505

[11] Marit Bentvelzen, Jasmin Niess, and Paweł W. Woźniak. 2023. Designing reflective derived metrics for fitness trackers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4, Article 158 (Jan. 2023), 19 pages. doi:10.1145/3569475

[12] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in HCI: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27. doi:10.1145/3517233

[13] Robbert Jan Beun, Siska Fitrianie, Fiemke Griffioen-Both, Sandor Spruit, Corine Horsch, Jaap Lancee, and Willem-Paul Brinkman. 2017. Talk and Tools: The best of both worlds in mobile user interfaces for e-coaching. *Personal and Ubiquitous Computing* 21 (2017), 661–674. doi:10.1007/s00779-017-1021-5

[14] Arpita Bhattacharya, Samantha Kolovson, Yi-Chen Sung, Mike Eacker, Michael Chen, Sean A Munson, and Julie A Kientz. 2018. Understanding pivotal experiences in behavior change for the design of technologies for personal wellbeing. *Journal of Biomedical Informatics* 79 (2018), 129–142. doi:10.1016/j.jbi.2018.01.002

[15] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI research: Going behind the scenes. *Synthesis Lectures on Human-Centered Informatics* 9, 1 (April 2016), 115 pages. doi:10.2200/S00706ED1V01Y201602HCI034

[16] Jeffrey R. Blum, Pascal E. Fortin, Feras Al Taha, Parisa Alirezaee, Marc Demers, Antoine Weill-Duflos, and Jeremy R. Cooperstock. 2019. Getting your hands dirty outside the lab: A practical primer for conducting wearable vibrotactile haptics research. *IEEE Transactions on Haptics* 12, 3 (2019), 232–246. doi:10.1109/TOH.2019.2930608

[17] Clive Roland Boddy. 2016. Sample size for qualitative research. *Qualitative Market Research: An International Journal* 19, 4 (2016), 426–432. doi:10.1108/QMR-06-2016-0053

[18] Petter B. Brandtzaeg, Marita B. Skjuve, Kim K. Dysthe, and Asbjørn Følstad. 2021. When the social becomes non-human: Young people's perception of social support in chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 257, 13 pages. doi:10.1145/3411764.3445318

[19] Eleanor R. Burgess, Kathryn E. Ringland, Jennifer Nicholas, Ashley A. Knapp, Jordan Eschler, David C. Mohr, and Madhu C. Reddy. 2019. "I think people are powerful": The sociality of individuals managing depression. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 41 (Nov. 2019), 29 pages. doi:10.1145/3359143

[20] Victoria Cabales. 2019. Muse: Scaffolding metacognitive reflection in design-based research. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI EA 2019)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3308450

[21] Nadia Campo Woytuk, Marie Louise Juul Søndergaard, Marianela Ciolfi Felice, and Madeline Balaam. 2020. Touching and being in touch with the menstruating body. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2020)*. Association for Computing Machinery, New York, NY, USA, Article 344, 14 pages. doi:10.1145/3313831.3376471

[22] Charles S. Carver and Michael F. Scheier. 2000. Autonomy and self-regulation. *Psychological Inquiry* 11, 4 (2000), 284–291.

[23] Eun Kyoung Choe, Bongshin Lee, Sean Munson, Wanda Pratt, and Julie A Kientz. 2013. Persuasive performance feedback: The effect of framing on self-efficacy. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, Bethesda, MD, USA, 825–833.

[24] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, ON, Canada) *(CHI 2014)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. doi:10.1145/2556288.2557372

[25] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A. Kientz, and Sean A. Munson. 2016. Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, CA, USA) *(CSCW 2016)*. Association for Computing Machinery, New York, NY, USA, 770–786. doi:10.1145/2818048.2819926

[26] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT 2021)*. Association for Computing Machinery, New York, NY, USA, 598–609. doi:10.1145/3442188.3445921

[27] Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. 2008. Flowers or a robot army? Encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th International Conference on Ubiquitous Computing* (Seoul, Korea) *(UbiComp 2008)*. Association for Computing Machinery, New York, NY, USA, 54–63. doi:10.1145/1409635.1409644

[28] Eric Corbett and Emily Denton. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT 2023)*. Association for Computing Machinery, New York, NY, USA, 1624–1634. doi:10.1145/3593013.3594104

[29] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI 2015)*. Association for Computing Machinery, New York, NY, USA, 3207–3216. doi:10.1145/2702123.2702154

[30] John W. Creswell and Vicki L. Plano Clark. 2017. *Designing and conducting mixed methods research (3rd ed.).* SAGE Publications, London, UK.

[31] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2020)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376638

[32] Xianghua (Sharon) Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. 2021. Data engagement reconsidered: A study of automatic stress tracking technology in use. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 535, 13 pages. doi:10.1145/3411764.3445763

[33] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3338286.3340116

[34] Elizabeth V. Eikey and Madhu C. Reddy. 2017. "It's definitely been a journey": A qualitative study on how women with eating disorders use weight loss apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA) *(CHI 2017)*. Association for Computing Machinery, New York, NY, USA, 642–654. doi:10.1145/3025453.3025591

[35] Tina Ekhtiar, Armağan Karahanoğlu, Rúben Gouveia, and Geke Ludden. 2023. Goals for goal setting: A scoping review on personal informatics. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS 2023)*. Association for Computing Machinery, New York, NY, USA, 2625–2641. doi:10.1145/3563657.3596087

[36] Daniel Adam Epstein. 2018. *Everyday Personal Informatics.* Ph. D. Dissertation. University of Washington. https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/43016/Epstein_washington_0250E_19042.pdf

[37] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the device in the drawer: Lapses as a design opportunity in personal informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp 2016)*. Association for Computing Machinery, New York, NY, USA, 829–840. doi:10.1145/2971648.2971656

[38] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp 2015)*. Association for Computing Machinery, New York, NY, USA, 731–742. doi:10.1145/2750858.2804250

[39] Deborah Estrin. 2014. Small data, where n = me. *Communications of the ACM* 57, 4 (April 2014), 32–34. doi:10.1145/2580944

[40] Jixiang Fan, Derek Haqq, Morva Saaty, Wei-Lu Wang, and Scott McCrickard. 2025. Diary study as an educational tool: An experience report from an HCI course. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (Pittsburgh, PA, USA) *(SIGCSETS 2025)*. Association for Computing Machinery, New York, NY, USA, 311–317. doi:10.1145/3641554.3701845

[41] Jixiang Fan, Morva Saaty, and D. Scott Mccrickard. 2024. Education in HCI outdoors: A diary study approach. In *Proceedings of the 6th Annual Symposium on HCI Education* (New York, NY, USA) *(EduCHI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 3, 10 pages. doi:10.1145/3658619.3658621

[42] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI 2018)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174047

[43] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: Framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OzCHI 2010)*. Association for Computing Machinery, New York, NY, USA, 216–223. doi:10.1145/1952222.1952269

[44] Joseph L. Fleiss. 1981. *The measurement of interrater agreement.* Vol. 2. New York: John Wiley, New York, NY, USA. 212–236 pages.

[45] Margaret Flemings, Shanzay Kazmi, Rachel Pak, and Orit Shaer. 2018. Crimson Wave: Shedding light on menstrual health. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (Stockholm, Sweden) *(TEI 2018)*. Association for Computing Machinery, New York, NY, USA, 343–348. doi:10.1145/3173225.3173292

[46] Raluca Alexandra Fulgu and Valerio Capraro. 2024. Surprising gender biases in GPT. *Computers in Human Behavior Reports* 16 (2024), 100533.

[47] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, FL, USA) *(CHI 2003)*. Association for Computing Machinery, New York, NY, USA, 233–240. doi:10.1145/642611.642653

[48] ATLAS.ti Scientific Software Development GmbH. 2025. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[49] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2018. Activity tracking in vivo. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI 2018)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173936

[50] Nitesh Goyal and Susan R. Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, CA, USA) *(CSCW 2016)*. Association for Computing Machinery, New York, NY, USA, 288–302. doi:10.1145/2818048.2820071

[51] Anthony M. Grant, John Franklin, and Peter Langford. 2002. The Self-Reflection and Insight Scale: A new measure of private self-consciousness. *Social Behavior and Personality: An International Journal* 30, 8 (2002), 821–835. doi:10.2224/sbp.2002.30.8.821 Place: New Zealand Publisher: Society for Personality Research.

[52] Elize Herrewijnen, Meagan B. Loerakker, Marloes Vredenborg, and Paweł W. Woźniak. 2024. Requirements and attitudes towards explainable AI in law enforcement. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS 2024)*. Association for Computing Machinery, New York, NY, USA, 995–1009. doi:10.1145/3643834.3661629

[53] Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology* 26, 2 (June 2024), 38. doi:10.1007/s10676-024-09775-5

[54] Michael Hoefer and Stephen Voida. 2023. Being, having, doing, and interacting: A personal informatics approach to understanding human need satisfaction in everyday life. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS 2023)*. Association for Computing Machinery, New York, NY, USA, 2593–2610. doi:10.1145/3563657.3596120

[55] Sarah Homewood, Kari Okholm Just, and Olivia Bramm Johansson. 2024. The unanticipated use of fitness tracking technologies during post-COVID syndrome. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS 2024)*. Association for Computing Machinery, New York, NY, USA, 556–570. doi:10.1145/3643834.3661617

[56] Sarah Homewood and Anna Vallgårda. 2020. Putting phenomenological theories to work in the design of self-tracking technologies. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS 2020)*. Association for Computing Machinery, New York, NY, USA, 1833–1846. doi:10.1145/3357236.3395550

[57] Kristina Höök. 2006. Designing familiar open surfaces. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI 2006)*. Association for Computing Machinery, New York, NY, USA, 242–251. doi:10.1145/1182475.1182501

[58] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury. 2014. Personal visualization and personal visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (2014), 420–433. doi:10.1109/TVCG.2014.2359887

[59] Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in Automated Speaker Recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT 2022)*. Association for Computing Machinery, New York, NY, USA, 230–247. doi:10.1145/3531146.3533089

[60] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes from the past: How technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI 2013)*. Association for Computing Machinery, New York, NY, USA, 1071–1080. doi:10.1145/2470654.2466137

[61] Karin A.M. Janssens, Elisabeth H. Bos, Judith G.M. Rosmalen, Marieke C. Wichers, and Harriëtte Riese. 2018. A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology* 18, 140 (2018), 12 pages. doi:10.1186/s12874-018-0579-6

[62] Gyuwon Jung, Sangjun Park, and Uichin Lee. 2024. DeepStress: Supporting stressful context sensemaking in personal informatics systems using a quasi-experimental approach. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 1000, 18 pages. doi:10.1145/3613904.3642766

[63] Manasa Kalanadhabhatta, Adrelys Mateo Santana, Zhongyang Zhang, Deepak Ganesan, Adam S. Grabell, and Tauhidur Rahman. 2022. EarlyScreen: Multi-scale instance fusion for predicting neural activation and psychopathology in preschool children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2, Article 60 (July 2022), 39 pages. doi:10.1145/3534583

[64] Armağan Karahanoğlu, Aykut Coskun, Dees Postma, Bouke Leonard Scheltinga, Rúben Gouveia, Dennis Reidsma, and Jasper Reenalda. 2024. Is it just a score? Understanding training load management practices beyond sports tracking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 313, 18 pages. doi:10.1145/3613904.3642051

[65] Rohit Ashok Khot, Larissa Hjorth, and Florian 'Floyd' Mueller. 2014. Understanding physical activity through 3D printed material artifacts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, ON, Canada) *(CHI 2014)*. Association for Computing Machinery, New York, NY, USA, 3835–3844. doi:10.1145/2556288.2557144

[66] Hankyung Kim, Dong Yoon Koh, Gaeun Lee, Jung-Mi Park, and Youn-kyung Lim. 2019. Designing personalities of conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI EA 2019)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312887

[67] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. 2018. Can a machine tend to teenagers' emotional needs? A study with conversational agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI EA 2018)*. Association for Computing Machinery, New York, NY, USA, 6 pages. doi:10.1145/3170427.3188548

[68] Jens Kleesiek, Yonghui Wu, Gregor Stiglic, Jan Egger, and Jiang Bian. 2023. An opinion on ChatGPT in health care—written by humans only. *Journal of Nuclear Medicine* 64, 5 (2023), 701–703. doi:10.2967/jnumed.123.265687

[69] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior* 160 (2024), 108352. doi:10.1016/j.chb.2024.108352

[70] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2, Article 70 (July 2018), 26 pages. doi:10.1145/3214273

[71] Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. The personality dimensions GPT-3 expresses during human-chatbot interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2, Article 61 (May 2024), 36 pages. doi:10.1145/3659626

[72] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875* (2022), 23 pages.

[73] Sylvie D Lambert and Carmen G Loiselle. 2008. Combining individual interviews and focus groups to enhance data richness. *Journal of Advanced Nursing* 62, 2 (2008), 228–237. doi:10.1111/j.1365-2648.2007.04559.x

[74] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (March 1977), 159–174. doi:10.2307/2529310

[75] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 632, 15 pages. doi:10.1145/3411764.3445312

[76] Effie Lai-Chong Law, AsbjØRn FØLstad, and Nena Van As. 2022. Effects of humanlikeness and conversational breakdown on trust in chatbots for customer service. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3546155.3546665

[77] Kwangyoung Lee and Hwajung Hong. 2018. MindNavigator: Exploring the stress and self-interventions for mental wellness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI 2018)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174146

[78] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI 2019)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300932

[79] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2020)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376175

[80] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI 2010)*. Association for Computing Machinery, New York, NY, USA, 557–566. doi:10.1145/1753326.1753409

[81] Nira Liberman, Yaacov Trope, and Cheryl Wakslak. 2007. Construal level theory and consumer behavior. *Journal of Consumer Psychology* 17, 2 (April 2007), 113–117. doi:10.1016/S1057-7408(07)70017-7

[82] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *Proceedings of the 8th International Conference on Ubiquitous*

Computing (Orange County, CA) *(UbiComp 2006)*. Springer-Verlag, Berlin, Heidelberg, 261–278. doi:10.1007/11853565_16

[83] Edwin A Locke and Gary P Latham. 2019. The development of goal setting theory: A half century retrospective. *Motivation Science* 5, 2 (2019), 93. doi:10.1037/mot0000127

[84] Meagan B. Loerakker, Jasmin Niess, Marit Bentvelzen, and Paweł W. Woźniak. 2023. Designing data visualisations for self-compassion in personal informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4, Article 169 (Dec. 2023), 22 pages. doi:10.1145/3631448

[85] Deborah Lupton. 2014. Self-tracking cultures: towards a sociology of personal informatics. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (Sydney, New South Wales, Australia) *(OzCHI 2014)*. Association for Computing Machinery, New York, NY, USA, 77–86. doi:10.1145/2686612.2686623

[86] Sonja Lyubomirsky, Kennon M. Sheldon, and David Schkade. 2005. Pursuing happiness: The architecture of sustainable change. *Review of General Psychology* 9, 2 (2005), 111–131. doi:10.1037/1089-2680.9.2.111

[87] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can we talk? Design implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. doi:10.1145/3469595.3469600

[88] Lena Mamykina, Daniel A. Epstein, Predrag Klasnja, Donna Sprujt-Metz, Jochen Meyer, Mary Czerwinski, Tim Althoff, Eun Kyoung Choe, Munmun De Choudhury, and Brian Lim. 2022. Grand challenges for personal informatics and AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA 2022)*. Association for Computing Machinery, New York, NY, USA, Article 76, 6 pages. doi:10.1145/3491101.3503718

[89] Lena Mamykina, Arlene M Smaldone, and Suzanne R Bakken. 2015. Adopting the sensemaking perspective for chronic disease self-management. *Journal of Biomedical Informatics* 56 (Aug. 2015), 406–417. doi:10.1016/j.jbi.2015.06.006

[90] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047* (2019), 8 pages.

[91] Catherine Marshall and Gretchen B. Rossman. 2014. *Designing qualitative research (6th ed.)*. SAGE Publications, USA.

[92] Thomas Mildner, Orla Cooney, Anna-Maria Meck, Marion Bartl, Gian-Luca Savino, Philip R. Doyle, Diego Garaialde, Leigh Clark, John Sloan, Nina Wenig, Rainer Malaka, and Jasmin Niess. 2024. Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 307, 18 pages. doi:10.1145/3613904.3642542

[93] Robert R. Morris, Stephen M. Schueller, and Rosalind W. Picard. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: Randomized controlled trial. *Journal of Medical Internet Research* 17, 3 (2015), e72. doi:10.2196/jmir.4167

[94] Christine Murad and Cosmin Munteanu. 2020. Designing Voice Interfaces: Back to the (Curriculum) Basics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376522

[95] Jasmin Niess, Kristina Knaving, Alina Kolb, and Paweł W. Woźniak. 2020. Exploring fitness tracker visualisations to avoid rumination. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI 2020)*. Association for Computing Machinery, New York, NY, USA, Article 6, 11 pages. doi:10.1145/3379503.3405662

[96] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting meaningful personal fitness: The Tracker Goal Evolution Model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI 2018)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173745

[97] Donald A. Norman. 1988. *The psychology of everyday things.* Basic Books, New York, NY, USA.

[98] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. doi:10.1126/science.aax2342

[99] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. 2018. "Suddenly, we got to become therapists for each other": Designing peer support chats for mental health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI 2018)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173905

[100] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research* 34, 5 Part II (Dec. 1999), 1189–1208.

[101] James W. Pennebaker and Cindy K. Chung. 2011. Expressive writing: Connections to physical and mental health. In *The Oxford Handbook of Health Psychology*. Oxford University Press, Oxford, UK, 417–437. doi:10.1093/oxfordhb/9780195342819.013.0018

[102] Laura Pina, Sang-Wha Sien, Clarissa Song, Teresa M. Ward, James Fogarty, Sean A. Munson, and Julie A. Kientz. 2020. DreamCatcher: Exploring how parents and school-age children can track and review sleep information together. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1, Article 70 (May 2020), 25 pages. doi:10.1145/3392882

[103] Laura R. Pina, Sang-Wha Sien, Teresa Ward, Jason C. Yip, Sean A. Munson, James Fogarty, and Julie A. Kientz. 2017. From personal informatics to family informatics: Understanding family practices around health monitoring. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, OR, USA) *(CSCW 2017)*. Association for Computing Machinery, New York, NY, USA, 2300–2315. doi:10.1145/2998181.2998362

[104] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. McLean, VA, USA, 2–4.

[105] Judith J. Prochaska, Erin A. Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. *Journal of Medical Internet Research* 23, 3 (2021), e24850. doi:10.2196/24850

[106] Aare Puussaar, Adrian K. Clear, and Peter Wright. 2017. Enhancing personal informatics through social sensemaking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA) *(CHI 2017)*. Association for Computing Machinery, New York, NY, USA, 6936–6942. doi:10.1145/3025453.3025804

[107] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (2016), 1–17. doi:10.1016/j.ijhcs.2016.05.006

[108] Amon Rapp and Lia Tirabeni. 2020. Self-tracking while doing sport: Comfort, motivation, attention and lifestyle of athletes using personal informatics tools. *International Journal of Human-Computer Studies* 140 (2020), 102434.

[109] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending chatbots to probe users: Enhancing complex decision-making through probing conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, UK) *(CUI 2022)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. doi:10.1145/3543829.3543832

[110] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's good to talk: A comparison of using voice versus screen-based interactions for agent-assisted tasks. *ACM Transactions on Computer-Human Interaction* 29, 3 (2022), 41. doi:10.1145/3484221

[111] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I interrupt? Diverging opinions on proactive smart speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI 2021)*. Association for Computing Machinery, New York, NY, USA, Article 34, 10 pages. doi:10.1145/3469595.3469629

[112] Yvonne Rogers. 2012. *HCI theory: Classical, modern, and contemporary*. Vol. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool Publishers, San Rafael, CA, USA. doi:10.2200/S00418ED1V01Y201205HCI014

[113] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4, Article 159 (Jan. 2018), 23 pages. doi:10.1145/3161187

[114] Yvon Ruitenburg, Minha Lee, Wijnand IJsselsteijn, and Panos Markopoulos. 2024. Seeking truth, comfort, and connection: How conversational user interfaces can help couples with dementia manage reality disjunction. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI 2024)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3640794.3665547

[115] Herman Saksono, Carmen Castaneda-Sceppa, Jessica Hoffman, Vivien Morris, Magy Seif El-Nasr, and Andrea G. Parker. 2020. Storywell: Designing for family fitness app motivation by using social rewards and reflection. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2020)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376686

[116] Herman Saksono, Ashwini Ranade, Geeta Kamarthi, Carmen Castaneda-Sceppa, Jessica A. Hoffman, Cathy Wirth, and Andrea G. Parker. 2015. Spaceship Launch: Designing a collaborative exergame for families. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW 2015)*. Association for Computing Machinery, New York, NY, USA, 1776–1787. doi:10.1145/2675133.2675159

[117] Katharine Sanderson. 2023. GPT-4 is here: What scientists think. *Nature* 615, 7954 (2023), 773.

[118] Kim Sauvé, Saskia Bakker, Nicolai Marquardt, and Steven Houben. 2020. LOOP: Exploring physicalization of activity tracking data. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI 2020)*. Association for Computing Machinery, New York, NY, USA, Article 52, 12 pages. doi:10.1145/3419249.3420109

[119] Christopher L. Schaefbauer, Danish U. Khan, Amy Le, Garrett Sczechowski, and Katie A. Siek. 2015. Snack Buddy: Supporting healthy snacking in low socioeconomic status families. In *Proceedings of the 18th ACM Conference on*

Computer Supported Cooperative Work & Social Computing (Vancouver, BC, Canada) *(CSCW 2015)*. Association for Computing Machinery, New York, NY, USA, 1045–1057. doi:10.1145/2675133.2675180

[120] Jessica Schroeder, Chia-Fang Chung, Daniel A. Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A. Munson. 2018. Examining self-tracking by people with migraine: Goals, needs, and opportunities in a chronic health condition. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS 2018)*. Association for Computing Machinery, New York, NY, USA, 135–148. doi:10.1145/3196709.3196738

[121] Jessica Schroeder, Ravi Karkar, Natalia Murinova, James Fogarty, and Sean A. Munson. 2020. Examining opportunities for goal-directed self-tracking to support chronic condition management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4, Article 151 (Sept. 2020), 26 pages. doi:10.1145/3369809

[122] Anjeli Singh and Sareeka Malhotra. 2013. A researcher's guide to running diary studies. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction* (Bangalore, India) *(APCHI 2013)*. Association for Computing Machinery, New York, NY, USA, 296–300. doi:10.1145/2525194.2525261

[123] Michael W. Sjoding, Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. 2020. Racial bias in pulse oximetry measurement. *New England Journal of Medicine* 383, 25 (2020), 2477–2478. doi:10.1056/NEJMc2029240

[124] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI 2017)*. Association for Computing Machinery, New York, NY, USA, 2696–2707. doi:10.1145/3025453.3025516

[125] Petr Slovák, Joris Janssen, and Geraldine Fitzpatrick. 2012. Understanding heart rate sharing: Towards unpacking physiosocial space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI 2012)*. Association for Computing Machinery, New York, NY, USA, 859–868. doi:10.1145/2207676.2208526

[126] Katarzyna Stawarz, Chris Preist, David Coyle, et al. 2019. Use of smartphone apps, social media, and web-based resources to support mental health and well-being: Online survey. *JMIR Mental Health* 6, 7 (2019), e12546. doi:10.2196/12546

[127] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating fitness: Leveraging large language models for reflective fitness tracker data interpretation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 646, 16 pages. doi:10.1145/3613904.3642032

[128] Sruti Subramanian, Katrien De Moor, Markus Fiedler, Kamil Koniuch, and Lucjan Janowski. 2023. Towards enhancing ecological validity in user studies: A systematic review of guidelines and implications for QoE research. *Quality and User Experience* 8, 1 (2023), 6. doi:10.1007/s41233-023-00059-2

[129] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (2024), pgae346. doi:10.1093/pnasnexus/pgae346

[130] Tammy Toscos, Anne Faber, Shunying An, and Mona Praful Gandhi. 2006. Chick Clique: Persuasive technology to motivate teenage girls to exercise. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (Montréal, QC, Canada) *(CHI EA 2006)*. Association for Computing Machinery, New York, NY, USA, 1873–1878. doi:10.1145/1125451.1125805

[131] Paul D. Trapnell and Jennifer D. Campbell. 1999. Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology* 76, 2 (1999), 284. doi:10.1037/0022-3514.76.2.284

[132] Laia Turmo Vidal and Jared Duval. 2024. Ambiguity as a resource to design for a plurality of bodies. In *Proceedings of the Halfway to the Future Symposium* (Santa Cruz, CA, USA) *(HttF 2024)*. Association for Computing Machinery, New York, NY, USA, Article 21, 9 pages. doi:10.1145/3686169.3686176

[133] Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* 16, 3 (2018), 252–260. doi:10.1108/JICES-06-2018-0056

[134] Niels van Berkel, Matthew J Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R Davidson, and Ann Blandford. 2020. Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics* 110 (2020), 103553.

[135] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI 2020)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376210

[136] Ron Wakkary and Leah Maestri. 2007. The resourcefulness of everyday design. In *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition* (Washington, DC, USA) *(C&C 2007)*. Association for Computing Machinery, New York, NY, USA, 163–172. doi:10.1145/1254960.1254984

[137] Ron Wakkary and Leah Maestri. 2008. Aspects of everyday design: Resourcefulness, adaptation, and emergence. *International Journal of Human–Computer Interaction* 24, 5 (2008), 478–491. doi:10.1080/10447310802142276

[138] WHOOP. 2023, Sept. 26. Introducing WHOOP Coach, Powered By OpenAI. Retrieved April 26, 2025 from https://www.whoop.com/us/en/thelocker/introducing-whoop-coach-powered-by-openai/?srsltid=AfmBOop4HX7nE2DBOhsza4qhmKsNhGbIQd4jyiLU37UjgXYb6Fcr_lGx

[139] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, can you help us solve this problem? How conversations with smart personal assistant tutors increase task group outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI EA 2019)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3313090

[140] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (May 2024), 1–35. doi:10.1145/3659625

[141] Sofia Yfantidou, Marios Constantinides, Dimitris Spathis, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2023. Beyond accuracy: A critical review of fairness in machine learning for mobile and wearable computing. *arXiv preprint arXiv:2303.15585* (2023), 30 pages.

[142] Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. 2023. Uncovering bias in personal informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3, Article 139 (Sept. 2023), 30 pages. doi:10.1145/3610914

[143] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health* 6, 1 (2024), e12–e22.

[144] Nima Zargham, Vino Avanesi, Thomas Mildner, Kamyar Javanmardi, Robert Porzel, and Rainer Malaka. 2024. HASI: A model for Human-Agent Speech Interaction. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI 2024)*. Association for Computing Machinery, New York, NY, USA, Article 44, 8 pages. doi:10.1145/3640794.3665885

[145] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, UK) *(CUI 2022)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. doi:10.1145/3543829.3543834

[146] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments* 11, 1 (2024), 28. doi:10.1186/s40561-024-00316-7

[147] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023), 33 pages.

[148] Zhirun Zhang, Xinzhi Zhang, and Li Chen. 2021. Informing the design of a news chatbot. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Japan) *(IVA 2021)*. Association for Computing Machinery, New York, NY, USA, 224–231. doi:10.1145/3472306.3478358

A Instruction Video Content

The instruction video used in the study contained the following explanations and contents:

- **Aesthetic:** The video approaches the participant in a more 'personal' manner, with the first author being present in the video. Initially, they explain the goal of the study. Besides videos of themselves, the first author also voiced over videos of the tracker and screenshots of the WHOOP app.
- **Charging:** How to charge the WHOOP tracker with the battery pack, as well as what the different-coloured lights communicate.
- **App:** How to connect the tracker to the WHOOP app with its built-in Bluetooth connection, including what to do when the Bluetooth does not work automatically.
- **Personalisation:** How to personalise the app according to some basic bodily information, including weight, height, and gender.
- **Metrics:** A mention of some basic metrics that are included in the app, like sleep, heart rate variability (HRV), recovery, strain and menstrual cycle tracking.