HHAI 2025
D. Pedreschi et al. (Eds.)

doi:10.3233/FAIA250649

Social AI for a Healthier Lifestyle: Four Competencies to Manage and Prevent Chronic Diseases

Mark NEERINCX [a,b,1], Jasper VAN DER WAA [a], Myrthe L. TIELMAN [b], Chenxu HAO [b], Liv ZIEGFELD [a], Davide DELL'ANNA [c] and Shihan WANG [c]

[a] *TNO, The Netherlands*
[b] *Delft University of Technology, The Netherlands*
[c] *Utrecht University, The Netherlands*

Abstract. Lifestyle-related diseases like type 2 diabetes mellitus (T2DM) and chronic obstructive pulmonary disease (COPD), have a major impact on society, asking for comprehensive disease management support. While AI technology has advanced for diagnosis and disease detection, its implementation into eHealth and mHealth applications remains limited, with low adoption rates and limited evidence of effectiveness. To achieve the necessary levels of client engagement and self-efficacy in chronic disease lifestyle management (CDLM), Artificial Intelligence (AI) support must demonstrate social competencies throughout its entire lifecycle—an under-researched topic. This paper introduces a novel Social AI Competence framework designed to provide durable personalized CDLM-support. The framework defines four complementary core competencies: (1) supporting meaningful activities, (2) providing responsible actionable explanations, (3) engaging persons in reflective interactions, and (4) strengthening and leveraging support networks. Underlying these competencies are eleven key social skills, detailed in terms of their foundation, functionality, state-of-the-art advancements, and research and development challenges. The CDLM system under development employs interactive modeling techniques to incorporate the experience and expertise of both experts and clients into these skills, supported by a modular architecture that ensures adaptability and scalability. Integrating social AI functions into the competency framework enables systematic assessment and optimization of their proportional effectiveness in real-world use cases.

Keywords. hybrid intelligence, social intelligence, socially interactive agent, collaboration patterns, personalization, lifestyle, chronic disease, diabetes, COPD

1. Introduction

Lifestyle-related diseases, such as type 2 diabetes mellitus (T2DM) and chronic obstructive pulmonary disease (COPD), have significant societal and economic impacts. Poor lifestyle choices, such as unhealthy diets and physical inactivity, are key contributors [1]. As populations age and urbanize, these diseases are rising, exacerbating health inequal-

[1] Corresponding Author: Mark Neerincx, mark.neerincx@tno.nl.

ities and threatening sustainable development, making effective prevention and management crucial.

Extensive research focuses on advancing AI technology for (semi-)automated detection and diagnosis of specific diseases like T2DM [2] and COPD [3,4]. However, this technology has only been partially integrated into various eHealth and mHealth applications for the management of these chronic diseases. The adoption of these applications remains low, and there is limited evidence supporting their effectiveness [5]. To address this, several researchers [6,7] propose Hybrid Intelligence (HI) as a socio-technical approach to enhance Chronic Disease Lifestyle Management (CDLM), where human and artificial intelligence work together, complementing and augmenting each other's knowledge and capabilities. So far, this research has shown a focus on advancing the AI-modeling of the concerning disease (e.g. for classification, prediction or treatment planning) and the corresponding human-AI information exchange (e.g., interactive knowledge graphs, explanations and personalization of advice) to enable human-AI knowledge sharing. For a specific disease like T2DM and based on a corresponding data-set, core HI-functions of the CDLM-system have been explored concerning patient profiling, management & care activity prioritizing, and shared decision making [7]. Important HI-research challenges concern for a major part social functions like stakeholder involvement, engaging human-AI interactions, keeping up with rapidly evolving (distributed) domain knowledge, and accounting for the interdependence of the required collaborative activities between the different actors in disease management [6].

To develop integrated long-term AI-support for individuals in their personal contexts, we need a coherent trans-disciplinary research approach in which the various subtopics of the challenges are addressed in conjunction. This paper presents such an approach in which different HI-research groups worked-out a concrete overarching CDLM research & development objective that covers these subtopics, driving the required scientific collaboration. This objective centers on the specific *competencies* needed to support a healthy lifestyle [8] and to actually change the behaviors accordingly [9]. A competency consists of multiple related skills, along with knowledge and behaviors. As discussed above, important gaps in the concerning HI-research concern the *social* aspects of lifestyle-related disease management. Therefore, our main shared objective is to identify and develop the *social competencies* and underlying *skills* that AI should have in a hybrid CDLM-system. These social AI competencies are *not* copies but complements or derivatives of the human competencies involved. The AI-competencies involve models of the health-promoting goals, activity planning (interventions), information exchanges, and assessments *in relation to* the individual's state, personal values, group norms and circumstances (including the social environment). The corresponding model-driven social CDLM-support should bring about the social conditions for progress and maintenance of a healthy lifestyle, such as the required levels of inclusiveness, trust, engagement and self-efficacy [10,11,12]. We aim at the development of competencies that generalize over specific use cases and their focused, data-set driven, analyses of (variants of) diseases, enabling theory building and transfer over specific applications in the CDLM-domain.

2. Building AI's Social Competencies

Recent HI-developments for supporting chronic disease prevention and management provide a modular architecture that allows for adding social functionality. By building on

such an architecture (with the available models), this functionality can be worked-out, prototyped and tested in an effective and efficient way. For example, a good starting point for such research is a decision support system with a supervised learning AI module that predicts whether a client will develop T2DM in the future, based on physiological- and lifestyle-data. This system contains a basic dialogue module, providing up-to-date information in the form of knowledge graphs that enable the reasoning for personalized support [13]. Such a knowledge graph needs to be updated during its life-cycle, but manually updating is labour intensive and prone to errors. Syntactic and semantic metrics for change evaluation were developed and tested, providing empirical evidence of metrics' efficacy in assessing modifications to knowledge graphs from different domains [14]. How far this knowledge graph improves the prediction of diabetes type 2, is currently being studied with the Kaggle Diabetes Prediction dataset [2].

The social AI research builds on this combined machine-learning & knowledge engineering approach, distinguishing 4 social competencies to develop: (1) supporting meaningful activities, (2) providing responsible actionable explanations, (3) engaging persons in reflective interactions, and (4) strengthening and leveraging support networks (Fig 1).

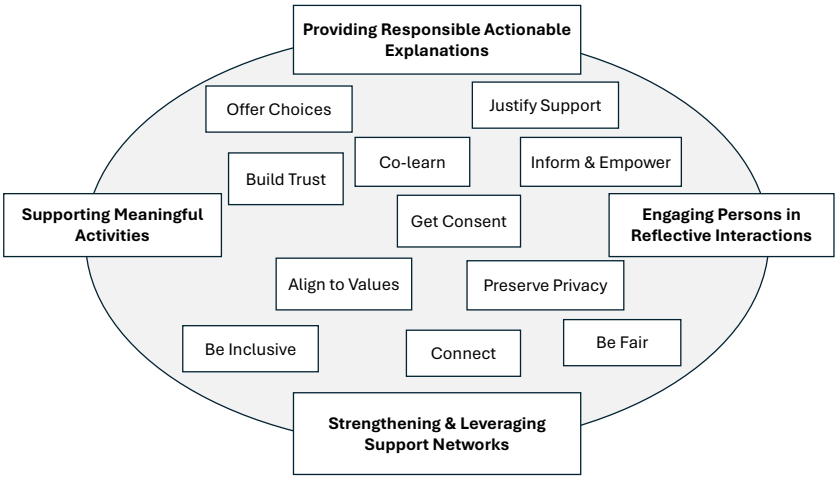

Figure 1. The Social AI Competency Framework, consisting of four competencies (presented in bold) and eleven underlying skills (in the running text, these skills are formatted with SMALL CAPITALS).

2.1. Supporting Meaningful Activities

An activity is meaningful when it is aligned with the personal VALUES of the stakeholders, respecting the values of other stakeholders, and conforming with the applicable norms (e.g., on INCLUSIVENESS). In a similar way, the information and advises of the CDLM-system should be meaningful for the clients. Two characteristics of values are relevant for value alignment: They change or evolve over time as people interact with changing environments [15,16] and can be shown in emotional responses (e.g. motivation; [17]).

[2] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Values are typically defined as beliefs that transcend specific situations, referring to desirable goals which motivate action, and which are used as evaluation criteria. Multiple theories around values exist, which generally organize values in some way, for instance identifying values which are more similar (e.g., Loyalty & Responsibility) or often conflict (e.g., Security & Privacy) [18]. Value-sensitive design (VSD) is an approach which aims to effectively incorporate the values and norms of stakeholders into the design of systems and products [19]. This means involving stakeholders and end-users early, and translating their value preferences to more concrete norms and design decisions in a methodological way. This may also involve breaking down higher-order values into sub-factors. Hereby it is important to investigate the links between different values or sub-factors, especially when designing a broader set of social AI capabilities. For instance, the health literacy of a patient (sub-factor of autonomy) may influence the lifestyle a patient adopts (sub-factor of patient well-being), which could thus also affect the degree to which a patient follows the lifestyle advice provided by the CDLM-system. Hence, the relationships between different values are important to consider to maximize the meaningfulness of the advice.

Current value sensitive design methods are mainly focused on values of groups of stakeholders, and do not necessarily allow a system to align with an individual user's unique values. In recent years, people have increasingly been working on potential methods to represent individual's values and norms in a way which allows systems, including embodied systems, to adapt and learn at run-time based on interactions with users [20,21,22]. This also requires eliciting and updating these values in a personal and flexible way, for instance through conversations [23,13]. This personalized approach to elicitation is crucial as different individuals may have different understandings of a given value. As mentioned above, values refer to goals as a motivation to perform the related actions. Supporting value-aligned goal-setting enhances the meaningfulness of the specific activities for the person concerned, further increasing the adoption and effectiveness of the CDLM-support [24,25]. Reinforcement learning can be used to further personalize the support over time [25]. Only if the CDLM-system aligns with the personal values and goals can optimal meaningfulness and effectiveness be achieved.

Meaningfulness is an important condition to BUILD TRUST in the CDLM-system, which is crucial for adoption [26,27,28]. Levy et al. (2023) developed a conceptual model of trust in Health-Behavior-Change AI apps [29]. Their model highlights that aside from the characteristics of the system (e.g. safety, transparency, & explainability), the user (e.g. AI literacy and self-efficacy) and environmental characteristics (e.g. services & support) also play a role. Moreover, trust in the technical system isn't the only relevant trust, but trust in the mediator (i.e. company) is also relevant. Given the changeable nature of people it is unavoidable that the system might occasionally be wrong in it's advice, leading to trust violations [30]. In these cases, trust repair is possible [31], but should be accompanied by strategies which aim to constantly, interactively update the system's knowledge to better align with users in the future [23].

When a CDLM-system interacts with a user, the system may need to handle sensitive information such as medical records, recommend actions, share information with others (software or humans), or consult others when a decision involves them. These interactions necessitate a careful regulation of the system's autonomy to ensure alignment with user VALUES and to ensure that adequate TRUST can be put in the CDLM-system. In this sense, we argue that the notion of CONSENT should play a critical role in moderating the

CDLM-system's autonomy. Similar to how humans seek consent before acting on someone else's behalf or when they need resources from others, when a software collaborates with a human, we should expect it to determine when consent is needed, from whom, and for what actions. The type and degree of consent required depend on the context, the individuals involved, and established norms. Not all actions demand consent; however, the ability to compute when consent is required, and to manage and express it effectively, would empower a trustworthy CDLM-system to act responsibly on behalf of, and in support to, the human. Consent must regulate not only data sharing but also the reactive, proactive, and social behaviors of a software system. Existing software systems currently lack the capability to manage these complex consent interactions and dynamics. Towards this objective, formal representations of consent have been developed, which enable software agents to track its evolution over interactions [32]. Algorithms have been built that enable consent mechanisms to be computationally expressed and that allow autonomous software systems to dynamically regulate the use of their autonomy in alignment with user norms and preferences. This approach aligns with the principles of personalized and ethical support, essential for a CDLM-system. Future work is needed to effectively integrate such rich consent management life-cycle as part of CDLM-system hybrid social intelligence.

2.2. Providing Responsible Actionable Explanations

So far, the extensive research on explainable AI did not yet provide concrete models, methods or tools for the development of human-centered explanations [33]. The identification and support of meaningful activities, as described above, provides a sound, human-centered, foundation for the creation of the desired explanations (such as the VALUE and TRUST models; cf. [33]). This subsection discusses the consequential explanation competencies that have been developed.

The CDLM-system communicates explanations that provide insight into how the AI's output came to be. This is done through a combined contrastive and counterfactual explanation. A contrastive explanation compares two outputs in terms of the decision rule that separates them [34]. A counterfactual explanation communicates a hypothetical input that would result in a different output [35]. By combining contrastive rules and counterfactuals, an explanation is obtained that provides insight into the AI's functioning. For example, a client might receive the advice to walk more often. The accompanying contrastive explanation could be: "Walking is suggested as an activity instead of cycling because you value family activities". The counterfactual explanation could be: "If you would value friendly competition, the cycling activity would have been suggested". The combination leads to an explanation such as: "Walking is suggested because you value family activities, if you also valued friendly competition cycling would have been suggested instead". Notice that since the CDLM-system includes the client's VALUES (i.e., family, competition), the counterfactual addresses values and preferences the client might wish to add to the CDLM-system, realizing a hybrid CO-LEARNING process.

Both contrastive and counterfactual explanations aim to offer insight into the functioning of an AI. However, in the case of the CDLM-system the purpose of the explanation is not for the client to understand all intricate details of the AI's functioning. Instead, the purpose is to support and motivate the client in their choice of activities that are beneficial to them (OFFER CHOICES). As such, the CDLM-system communicates

the contrastive and counterfactual explanations in a supportive and *actionable* manner. The purpose of these actionable explanations is for the client to select and take appropriate action. In this case, this involves performing an activity that provides benefit for the client. For an explanation to be actionable in this setting, we designed it to INFORM AND EMPOWER the client to be able to select an activity and be motivated to perform it. The principles of informing and empowering leads to a set of properties the contrastive and counterfactual explanation should adhere to.

For the explanation to inform the client, we argue that it should JUSTIFY the proposed activity, be as *explicit* as possible, and propose only those activities that are *feasible* for the client to perform. The justification is meant to convey the underlying argument why a certain activity is beneficial (e.g. "walking is an entry-level activity with immediate health benefits"). When the explanation is explicit, it becomes clear what the activity actually entails (e.g., "walking should be done for half an hour each day"). Finally, only what the client can do should be included, which puts clear constraints on what can be proposed and mostly affects how the explanation is generated as non-feasible activities will not be communicated. The feasibility of the proposal is widely recognized as a property required for an explanation to become actionable [36,37,38]. However, offering a JUSTIFICATION and being explicit are both novel properties. The combination of these properties might lead to an explanation such as; "Walking for half an hour a day is suggested because you value family activities. It is also an entry-level activity with immediate health benefits. Cycling for 15 to 30 minutes each day is another activity with similar effects. Cycling would have been proposed instead of walking if you also valued friendly competition". This explanation serves to provide the information the client needs to determine if they want to follow the suggestion or not.

To empower the client to also perform the suggested activity, the explanation should OFFER CHOICES between activities, remind the client of *past joys and successes*, communicate *effects* that can be expected, and make the activity *manageable*. In the CDLM setting it is especially important to empower clients, as adopting a proposed activity in their daily lives implies a behavioral change. To make behavioral changes is notoriously difficult for people and requires prolonged commitment of often months [39]. To ensure and support this commitment over time, it is important that the explanations OFFER CHOICES to support the client's sense of autonomy (for example, "both walking and cycling are equally beneficial for you"). If the CDLM-system would only offer one choice, it is likely that the client would feel limited autonomy and agency [40,41]. This is detrimental to commitment, as the sense of autonomy and agency in people's action is viewed as a strong incentive to follow through [42,43]. Second, to boost commitment the explanation should remind the client of any past joys or successes pertaining to the proposed activity (for example, "In the past you enjoyed a brisk walk with your grandchildren"). This allows the client to reflect on what the activity brought them in the past which can motivate or distract from negative thoughts when considering the activity [44,45]. By mentioning future effects that can be expected serves a similar purpose, as the client is reminded of why the activity should be performed (for example, "when walking regularly your body can learn to make better use of insulin"). Finally, the activity should be communicated in a manageable or operable manner (for example, "you can start with walking half an hour every week, and slowly increase the frequency until you walk half an hour every day"). This recognizes that a behavioral change only comes for most through gradual steps of small changes [39]. With this property, the explanation should split a

proposed activity into concrete steps over time and ideally adhere to the client's progress. When we integrate all of these empowering properties, an example explanation might be: "By being active for 30 minutes a day, your body can learn to make more better use of insulin. This can be done walking or cycling every day. In the past, you enjoyed a brisk walk on Sunday with your grandchildren, consider making this a regular activity. As you get used to this, you will get this advice more frequently until you walk every day". None of these properties that aim to empower the client offer new insights into the functioning of the AI. This is not surprising as the act of empowerment has little to do with how the AI functions in the CDLM-system. Instead, empowering the client has much more to do with the explanation treating the client as a person and integrating our knowledge of behavioral change and its difficulties into the design of the explanation.

Finally, it is important the actionable explanations are provided *responsibly*. The generation of counterfactual explanations has some severe limitations due to them dealing directly with a data set whose points reflect individual clients. First, the counterfactuals found can be of different quality depending on the client [46]. This can overlap with certain minority groups if the data set is unbalanced or biased, leading to unfair treatment as some clients will receive lower quality explanations. Second, counterfactuals are sensitive to PRIVACY attacks [47]. A receiver of a counterfactual explanation is able to derive personal information from other clients if sufficient explanations are collected. We adopted the concept of k-anonymity to mitigate leaking private data [48]. The underlying principle here is that the identified counterfactuals are sufficient generic that a potential attacker cannot distinguish between k amount of people. As k increases, the explanation will become more generic and thus affecting the quality and relevance of the explanation. However, it has been shown that even with small values for k, privacy can be preserved with limited impact on the quality. For more information we refer to recent work on this method [48].

To mitigate the effects of potentially unfair explanations, we derived a metric that uses simulation to estimate the quality of a particular explanation. This can be used as either a way to assess if a selected technology that identifies the counterfactual used in the explanation is FAIR. Similarly, it can be adopted as an optimization criteria to prevent the selection of low quality counterfactuals. The proposed metric defines the quality of counterfactual as its degree of similarity to the current situation while still being sufficiently different to result in a different AI output. In other words, it should limit the number of necessary changes or additions in a client's values while still resulting in a different advice for an activity. This is a common property attributed to quality counterfactuals in the literature [36,37,38]. In addition, a quality counterfactual is also one that is located in a high data density area where the AI performs well. In other words, the counterfactual should be a data point that leads to an activity advice that is correct and common. This is another property found in the literature used to identify quality counterfactuals [36,49]. These two known principles identify a low quality counterfactual as one that is highly dissimilar to the client, is rare and potentially incorrect. We propose that the combination of these two known aspects into a quality metric can be used to assess the fairness of explanations. By simulating the generation of explanations for different clients and applying the metric, one can obtain insights if the used technology to find counterfactuals treats some client groups unfairly.

Lastly, a responsible actionable explanation is an explanation that is INCLUSIVE. Inclusiveness is an important norm, meaning that technology should be accessible for

all who might benefit. This accessibility goes beyond giving access, but also means that interactions should be understandable to all, and explanations should be aligned. As proposed by [50], a system is only explainable insofar users can properly understand and engage with the explanations given. In diabetes care for example, especially people with lower health literacy should not be forgotten. We recognize this as important and thus do not propose that the examples of our explanation provided earlier to be used as such. Instead, future work is needed to create various designs to communicate the explanation and evaluate them with a highly diverse group of clients.

2.3. Engaging Persons in Reflective Interactions

Reflection is a meta-cognitive activity, commonly defined as a process of gaining insights, new perspectives, and making changes through rethinking about the past. It can be beneficial for motivating behavior change and supporting lifestyle changes [51].

Intelligent systems can provide support for reflection by either promoting self-reflection, i.e., an individual reflective process, or engaging users in collaborative reflection. Existing systems that support self-reflection are often integrated with personal informatics systems, where information about personal behaviors is summarized and visualized for users to reflect on. However, there still are many challenges. One is adapting to individual user's rhythm of life and promoting reflection at the appropriate moment [52]. Another is providing explanations along with the personal information, as past work suggests that merely showing information does not trigger reflection necessarily [52]. In addition, past work has emphasized that to achieve the goal of guiding future behaviors, reflective learning requires guided and structured processes [53,54].

One possibility to address the issues of adaptability, helping users interpret information and providing a guided process, is to integrate a *conversational agent* into the dialogue module and to tailor the dialogue to the user's VALUES while engaging users in collaborative reflection. User's values can provide the system with knowledge about user's life rhythm and preferences, and can help the system guide the user to interpret personal information. In addition, these values can be updated through the continued human-AI interactions to allow the system support users better.

For the human-AI interaction, including the explanations of the previous subsection, a conversational agent is being integrated into the dialogue module that allows for more open and continuing information exchanges, addressing the cognitive and affective aspects of CDLM in the conversation. Two versions of a lifestyle-advice dialogue module (chatbot) were developed and tested for different realistic health scenarios: an empathetic and neutral one. The usability of the empathic chatbot proved to be higher (Chatbot Usability Questionnaire; [55]), showing the importance of empathetic communication with CDLM-systems (i.e., crucial for engagement and satisfaction) [56].

Reinforcement Learning (RL) methods offer significant potential for personalizing the interaction and supporting engagement over a long period of time [57]. As a fast-developing machine learning technique, RL targets sequential decision-making tasks and achieves a long-term learning goal via interactions with the environment [58], which is very much aligned with the long-term engagement and support needs of CDLM-systems. Since RL is particularly powerful by taking the feedback from users into account and updating the interactive strategy adaptively according to actual needs, in recent years, RL methods have been widely developed for optimizing human-AI interactions in dia-

logue systems [59], and shown as a well-suited solution in personalized (health) behavior change support [60]. For example, deep RL has become a mainstream method for training dialogue policies to optimize human-AI interactive tasks, achieving significant success in pipeline-based dialogue systems [61,62]. In contrast, reinforcement learning from human feedback (RLHF) has become essential for large language model (LLM)-based dialogue systems, as it integrates RL with human feedback to better align with human VALUES and preferences [63], facilitating CO-LEARNING. We therefore argue that RL methods can play a pivotal role in enhancing personalized interactions, ensuring they are adaptive and reflective towards the evolving needs and preferences of users, particularly in health-related and long-term engagement applications like CDLM-systems.

A subsequent challenge is to support long-term dialogues which need underpinning by some kind of a conversational memory [64,65,66]. Like activities, it should focus on meaningful experiences and relate to the values at stake. Such a meaningful memory allows for value-centered reflections, increasing the understanding of own and other drivers for healthy behaviors (i.e., increasing the health literacy and self-efficacy).

2.4. Strengthening and Leveraging Support Networks

Social support is essential for successful lifestyle management [9,67]. Particularly, the perceived support and belief that help is available of social support can have a positive effects on someone's health; strengthening individual's coping abilities is an important factor [67]. Support from informal (e.g., family, friends) and formal caregivers (e.g., health professionals) can enhance healthy lifestyle habits and mental health substantially [68,69]. The CDLM-system should strengthen and leverage the available and potential networks, making CONNECTIONS aligned with the VALUES of the stakeholders involved, and harmonized to the existing or envisioned socio-technical structures in which it will operate. This is crucial for its adoption and continued usage. Note that its adoption may in turn influence the current healthcare landscape, particularly with respect to social phenomena. For instance, if CDLM-systems are used on a larger scale this might have implications for the required tech-savviness and AI-literacy of healthcare providers', requiring up-skilling [70]. Doctors might be asked more frequently to help their patients with their CDLM-system or the healthcare professionals themselves might use decision-support systems. In the research and development of CDLM-systems, it should be identified and accommodated what healthcare practitioners need to understand about AI's workings in order to use the systems responsibly. Further, user studies should be conducted to supplement the substantial lack of empirical knowledge about the effects of these systems on work practices, relationships and end-users [71,72].

Aside from investigating effects of the CDLM-system on the client and practitioner in isolation, it should also be assessed how the usage of the CDLM-system influences the patient-doctor *relationship* and how they are CONNECTED. AI tools are frequently proposed as potential means to alleviate the high workloads faced by healthcare practitioners, which may support more person-centered care and might allow doctors to build more empathetic and meaningful doctor-patient relationships [73]. Critics however warn that AI tools may in fact further remove the doctor from the patient, have dehumanizing effects and might only result in even higher patient throughput and workloads for doctors [74,73]. Further, with the increased adoption of self-management tools, the frequency at which patients seek help of human doctors might significantly change. Concerns have

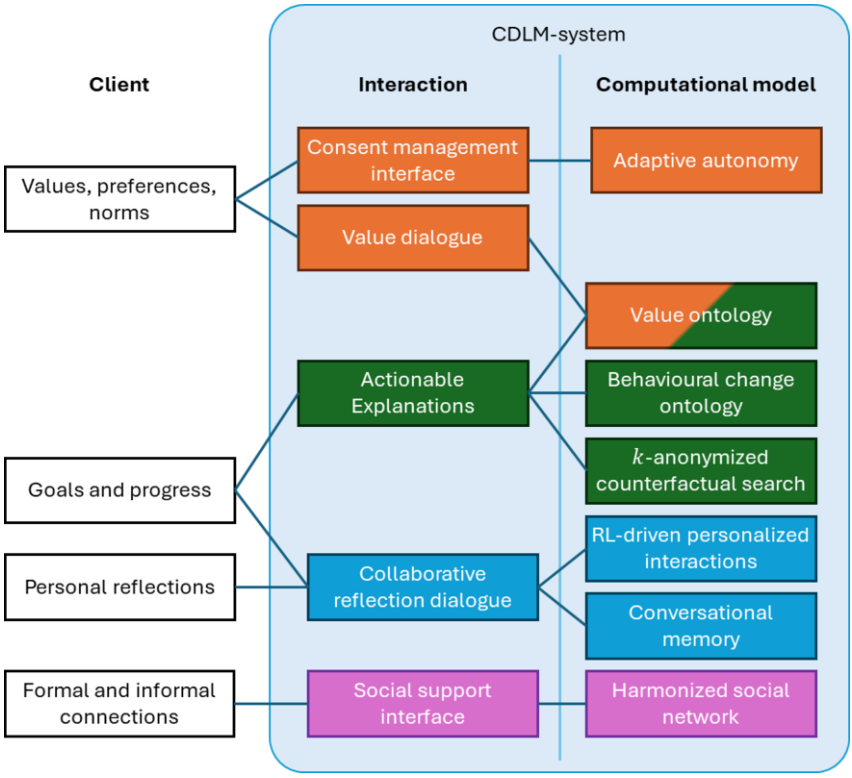

Figure 2. An overview of the Chronic Disease Lifestyle Management (CDLM) system with its interaction and computational modeling components necessary to obtain the core competencies and underlying skills.

been raised about the emergence of a two-tier access to healthcare as a result of the introduction of AI systems, whereby the contact with human doctors is reserved for wealthier patients due to the high costs, and where less wealthy individuals must rely more on AI tools [70]. The opposite is also a danger promoting inequality, by which less wealthy nations might have less means to set up computational infrastructures and other necessities for enabling use of self-management AI tools [70]. Note that the societal dynamics are rather complex, requiring insights in the changes over time and dependencies, such as: 1) how do value preferences change as a result of an altered patient-doctor relationship (e.g. patient autonomy, empathy of doctors, accessibility of care), 2) how does the social support that a CDLM-system may provide compare to the support from a human healthcare practitioner and which synergies may arise here? By addressing such questions in CDLM-research and development, we will be better able to harmonize the support to the broader societal context and the evolving social dynamics.

2.5. CDLM-System Components

Figure 2 illustrates the different technical components a CDLM-system requires to obtain the four described competencies. Consent was stated as vital to dictate the system's allowed autonomy and build a trust relation with the client. To do so, a consent management interface is required where given consent can be reviewed and where the system

can actively request for consent for particular behaviors. To do so, it utilizes the values, preferences and norms of the client obtained through an open dialogue and modeled in an ontology of values. These values should dictate when consent is required. In addition, these values play an important role in providing actionable explanations. For the CDLM-system to provide such explanations as described previously, the client's goals and progress are required and to be combined with its value ontology and a domain ontology that describes the effect of certain behavioral changes. With this knowledge, the system can employ a safe and secure way of identifying the necessary contrastive rules and counterfactuals to provide the actionable explanations. To do so, the privacy preservation technology of *k*-anonymization is used. Throughout the use of the system, the client can engage in a reflective dialogue where the system supports self-reflection using a memory of previous discussed topics and reinforcement learning to adapt the conversation. By involving the value model into this dialogue, clients get better insight in their health-related behavior choices. Finally, the CDLM-system requires a representation of the client's social network, harmonized to their formal and informal care context. This social network can be employed to provide a means for the client to interact with others to receive their support as well as to engage in joint activities.

Figure 2 shows how the interactive modeling approach is structured within the modular system design, enabling adaptability and scalability. Future work will address the key challenge of further developing these components and integrating the specified functions and technologies into a comprehensive, collaborative CDLM system. The next step will be to evaluate the extent to which the system embodies the described social AI competencies and underlying skills.

3. Conclusions

Hybrid Intelligence (HI) offers great potential to reduce the magnitude of a significant health problem, the increasing number of people who must cope with lifestyle-related chronic diseases, such as type 2 diabetes mellitus (T2DM) and chronic obstructive pulmonary disease (COPD). The social aspects of developing a desired healthy lifestyle are very important for adoption and long-term adherence to such a lifestyle, whereas the required coherent trans-disciplinary research on these aspects has been scarce. This paper presents a collaborative research line with a shared objective for the research an development of Chronic Disease Lifestyle Management (CDLM) support: the identification and development of the *social competencies* that AI should have, as complements or derivatives of the human competencies involved.

The main outcome of the presented collaborative research consist of a *social AI competency framework*, distinguishing four related competencies (Fig 1).

Supporting Meaningful Activities. This reflects the ability to understand a client's values, interests, and motivations and to guide them in integrating activities that align with these aspects into their daily lives. It demonstrates the ability to collaborate effectively, in which clients BUILD appropriate TRUST in the CDLM-support and their own lifestyle self-management capabilities (self-efficacy). This competence involves, among other things, the acquisition and formalization of stakeholders' VALUES in an adjustable model, which is adapted and refined by user interactions during its complete life-cycle. For example, life-style related advices and intervention advices are attuned to these val-

ues and daily routines, ensuring that the activities suggested by the CDLM-system resonate with the user's personal values and contexts.

Providing Responsible Actionable Explanations. This involves communicating information in a way that is clear, practical, and ethically sound, enabling clients to make informed decisions. It is tailored to the support of meaningful activities and aligned to the corresponding VALUES of the stakeholders. For example, contrastive and counterfactual explanations are presented in a practical, actionable manner to INFORM AND EMPOWER clients while ensuring PRIVACY, FAIRNESS, and INCLUSIVITY. These explanations help clients understand the rationale behind the system's recommendations, fostering TRUST and informed decision making.

Engaging Persons in Reflective Interactions. Facilitating reflective discussions helps clients examine their own thoughts, emotions, and behaviors, fostering self-awareness and motivation for change. It encourages users to evaluate their lifestyle choices and behaviors over time, helping them adapt and make more informed health decisions through reflective dialogue with AI. This competency encompasses the ability to create a safe, nonjudgmental space for dialogue. For example, the dialogue module provides a conversational agent that engages users in collaborative reflections with more open and continuing information exchanges, addressing the cognitive and affective aspects of CDLM, and enhancing the understanding of own and other's values in the behavioral choices (CO-LEARNING).

Strengthening and Leveraging Support Networks. Developing and leveraging a supportive network (e.g., healthcare providers, community resources, or peer groups) ensures that clients have access to comprehensive support options. This requires abilities in networking, and boundary-setting to connect clients with appropriate resources while maintaining professionalism. For example, the CDLM-system maintains an overview of active and potential relationships, and of resources that have been or can be consulted. In mixed-initiative dialogues, appropriate assessments and suggestions are made for maintaining, adjusting and consulting this network (i.e., to CONNECT).

Understanding stakeholders' values is an important aspect of all four competencies and, consequently the value models provide coherence in the expressions of these competencies, such as respectively, activity advice, explanation, interaction and network facilitation. These competencies also involve personalization and providing choice options to establish the required levels of inclusiveness, trust attribution, engagement and self-efficacy for the desired sustainable behavior change. The CDLM system is currently under development, utilizing interactive modeling techniques to integrate the experience and expertise of both experts and clients during its complete life-cycle.

By researching and developing social AI-functions within the CDLM framework, their proportional effectiveness in real-world use cases can be accurately assessed and optimized. As the framework distinguishes competencies and skills that generalize over specific use cases and chronic disease variants, it facilitates theory building and transfer over specific applications in the CDLM-domain.

References

[1] Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes research and clinical practice. 2022;183:109119.

[2] Chaki J, Ganesh ST, Cidham S, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences. 2022;34(6):3204-25.
[3] Bećirović LS, Deumić A, Pokvić LG, Badnjevic A. Artificial Inteligence Challenges in COPD management: a review. In: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE). IEEE; 2021. p. 1-7.
[4] Xu Y, Long ZA, Setyohadi DB. A Comprehensive Review on the Application of Artificial Intelligence in Chronic Obstructive Pulmonary Disease (COPD) Management. In: 2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM). IEEE; 2024. p. 1-8.
[5] Wang Y, Min J, Khuri J, Xue H, Xie B, Kaminsky LA, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR mHealth and uHealth. 2020;8(4):e15400.
[6] Dudzik BJ, van der Waa JS, Chen PY, Dobbe R, de Troya ÍM, Bakker RM, et al. Hybrid Intelligence Supports Application Development for Diabetes Lifestyle Management. Journal of Artificial Intelligence Research. 2024;80:919-29.
[7] de Boer MH, van der Waa J, van Gent S, Smit QT, Korteling W, van Stokkum RM, et al. A contextual hybrid intelligent system design for diabetes lifestyle management. In: International Workshop Modelling and Representing Context, ECAI. vol. 23; 2023. .
[8] Singh HK, Kennedy GA, Stupans I. Competencies and training of health professionals engaged in health coaching: A systematic review. Chronic illness. 2022;18(1):58-85.
[9] Deci EL, Ryan RM. Self-determination theory. Handbook of theories of social psychology. 2012;1(20):416-36.
[10] Igwama GT, Olaboye JA, Cosmos C, Maha MDA, Abdul S. AI-powered predictive analytics in chronic disease management: Regulatory and ethical considerations. International Journal Of Engineering Research And Development. 2024;20(4):405-10.
[11] Karekla M, Kasinopoulos O, Neto DD, Ebert DD, Van Daele T, Nordgreen T, et al. Best practices and recommendations for digital interventions to improve engagement and adherence in chronic illness sufferers. European Psychologist. 2019.
[12] Farley H. Promoting self-efficacy in patients with chronic disease beyond traditional education: A literature review. Nursing open. 2020;7(1):30-41.
[13] Chen PY, Baez Santamaria S, de Boer MHT, den Hengst F, Kamphorst BA, Smit Q, et al. Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning. In: HHAI 2024: Hybrid Human AI Systems for the Social Good. IOS Press; 2024. p. 457-9. Available from: https://ebooks.iospress.nl/doi/10.3233/FAIA240223.
[14] Bakker R, de Boer M. Dynamic knowledge graph evaluation. to appear.
[15] Van de Poel I, Kudina O. Understanding technology-induced value change: A pragmatist proposal. Philosophy & Technology. 2022;35(2):40.
[16] de Wildt T, van de Poel I. Modelling value change: An exploratory approach. Journal of Artificial Societies and Social Simulation. 2024;27(1).
[17] Nelissen RM, Dijker AJ, de Vries NK. Emotions and goals: Assessing relations between values and emotions. Cognition and Emotion. 2007;21(4):902-11.
[18] Schwartz SH. An Overview of the Schwartz Theory of Basic Values. Online readings in Psychology and Culture. 2012;2(1):2307-0919. Available from: https://scholarworks.gvsu.edu/orpc/vol2/iss1/11/.
[19] Friedman B, Kahn PH, Borning A. Value Sensitive Design and Information Systems. The handbook of information and computer ethics. 2008:69-101. Publisher: Wiley Online Library. Available from: https://link.springer.com/chapter/10.1007/978-94-007-7844-3_4.
[20] Tielman ML, Jonker CM, van Riemsdijk MB. What Should I Do? Deriving Norms from Actions, Values and Context. In: MRC@ IJCAI; 2018. p. 35-40.
[21] Cranefield S, Winikoff M, Dignum V, Dignum F. No Pizza for You: Value-based Plan Selection in BDI Agents. In: IJCAI; 2017. p. 178-84.
[22] Dell'Anna D, Jamshidnejad A. SONAR: An Adaptive Control Architecture for Social Norm Aware Robots. Int J Soc Robotics. 2024;16(9):1969-2000.
[23] Chen PY, Tielman ML, Heylen DK, Jonker CM, Van Riemsdijk MB. Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues. In: HHAI 2023: Augmenting Human Intellect. IOS Press; 2023. p. 93-107.

[24] Neerincx MA, Van Vught W, Blanson Henkemans O, Oleari E, Broekens J, Peters R, et al. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. Frontiers in Robotics and AI. 2019;6:118.
[25] Albers N, Hizli B, Scheltinga BL, Meijer E, Brinkman WP. Setting physical activity goals with a virtual coach: vicarious experiences, personalization and acceptance. Journal of medical systems. 2023;47(1):15.
[26] Orji R, Vassileva J, Mandryk RL. Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. User Modeling and User-Adapted Interaction. 2014;24:453-98. Publisher: Springer.
[27] Choung H, David P, Ross A. Trust in AI and Its Role in the Acceptance of AI Technologies. International Journal of Humanâ€“Computer Interaction. 2023 May;39(9):1727-39. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2050543. Available from: https://doi.org/10.1080/10447318.2022.2050543.
[28] Bhat S, Lyons JB, Shi C, Yang XJ. Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. HRI '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 32-41. Available from: https://dl.acm.org/doi/10.1145/3610977.3634921.
[29] Levy M, Pauzner M, Rosenblum S, Peleg M. Achieving trust in health-behavior-change artificial intelligence apps (HBC-AIApp) development: A multi-perspective guide. Journal of Biomedical Informatics. 2023 Jul;143:104414. Available from: https://www.sciencedirect.com/science/article/pii/S1532046423001351.
[30] Tolmeijer S, Weiss A, Hanheide M, Lindner F, Powers TM, Dixon C, et al. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction; 2020. p. 3-12.
[31] Kox ES, Kerstholt JH, Hueting TF, de Vries PW. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Autonomous Agents and Multi-Agent Systems. 2021;35(2):30. Publisher: Springer.
[32] Apeiron AS, Dell'Anna D, Murukannaiah PK, Yolum P. Model and Mechanisms of Consent for Responsible Autonomy. In: Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2025. International Foundation for Autonomous Agents and Multiagent Systems / ACM; 2025. .
[33] van Leersum CM, Maathuis C. Human Centred Explainable AI Decision-Making in Healthcare. Journal of Responsible Technology. 2025:100108.
[34] van der Waa J, Nieuwburg E, Cremers A, Neerincx M. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial intelligence. 2021;291:103404.
[35] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv JL & Tech. 2017;31:841.
[36] Poyiadzi R, Sokol K, Santos-Rodriguez R, De Bie T, Flach P. FACE: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020. p. 344-50.
[37] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency; 2020. p. 607-17.
[38] Schleich M, Geng Z, Zhang Y, Suciu D. GeCo: quality counterfactual explanations in real time. Proceedings of the VLDB Endowment. 2021;14(9):1681-93.
[39] Brandt CJ, Clemensen J, Nielsen JB, Søndergaard J. Drivers for successful long-term lifestyle change, the role of e-health: a qualitative interview study. BMJ open. 2018;8(3):e017466.
[40] Kulakova E, Khalighinejad N, Haggard P. I could have done otherwise: Availability of counterfactual comparisons informs the sense of agency. Consciousness and cognition. 2017;49:237-44.
[41] Barlas Z, Obhi SS. Freedom, choice, and the sense of agency. Frontiers in human neuroscience. 2013;7:514.
[42] Bandura A. The role of self-efficacy in goal-based motivation. New developments in goal setting and task performance. 2013:147-57.
[43] Alkire S. Subjective quantitative studies of human agency. Social indicators research. 2005;74:217-60.
[44] Markman KD, McMullen MN. A reflection and evaluation model of comparative thinking. Personality

and Social Psychology Review. 2003;7(3):244-67.
[45] Markman KD, McMullen MN, Elizaga RA, Mizoguchi N. Counterfactual thinking and regulatory fit. Judgment and Decision Making. 2006;1(2):98-107.
[46] Dai J, Upadhyay S, Aivodji U, Bach SH, Lakkaraju H. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society; 2022. p. 203-14.
[47] Goethals S, Sörensen K, Martens D. The privacy issue of counterfactual explanations: explanation linkage attacks. ACM Transactions on Intelligent Systems and Technology. 2023;14(5):1-24.
[48] Berning S, Dunning V, Spagnuelo D, Veugen T, van der Waa J. The Trade-off Between Privacy & Quality for Counterfactual Explanations. In: Proceedings of the 19th International Conference on Availability, Reliability and Security; 2024. p. 1-9.
[49] Zhang S, Chen X, Wen S, Li Z. Density-based reliable and robust explainer for counterfactual explanation. Expert Systems with Applications. 2023;226:120214.
[50] Tielman ML, Suárez-Figueroa MC, Jönsson A, Neerincx MA, Siebert LC. Explainable AI for All-a Roadmap for Inclusive XAI for people with Cognitive Disabilities. Technology in Society. 2024:102685.
[51] Bentvelzen M, Woźniak PW, Herbes PS, Stefanidi E, Niess J. Revisiting reflection in hci: Four design resources for technologies that support reflection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2022;6(1):1-27.
[52] Bentvelzen M, Niess J, Woźniak PW. 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: ACM; 2021. p. 1-12. Available from: https://dl.acm.org/doi/10.1145/3411764.3445505.
[53] Krogstie BR, Prilla M, Pammer V. Understanding and supporting reflective learning processes in the workplace: The csrl model. In: Scaling up Learning for Sustained Impact: 8th European Conference, on Technology Enhanced Learning, EC-TEL 2013, Paphos, Cyprus, September 17-21, 2013. Proceedings 8. Springer; 2013. p. 151-64.
[54] Slovák P, Frauenberger C, Fitzpatrick G. 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; 2017. p. 2696-707.
[55] Holmes S, Moorhead A, Bond R, Zheng H, Coates V, McTear M. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In: Proceedings of the 31st European Conference on Cognitive Ergonomics; 2019. p. 207-14.
[56] Reguera Gomez C. Building a Natural Conversational Agent for Healthcare by Examining Empathetic Language. Utrecht University; 2024.
[57] Zou L, Xia L, Ding Z, Song J, Liu W, Yin D. Reinforcement learning to optimize long-term user engagement in recommender systems. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 2810-8.
[58] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
[59] Uc-Cetina V, Navarro-Guerrero N, Martin-Gonzalez A, Weber C, Wermter S. Survey on reinforcement learning for language processing. Artificial Intelligence Review. 2023;56(2):1543-75.
[60] Brons A, Wang S, Visser B, Kröse B, Bakkes S, Veltkamp R. Machine Learning Methods to Personalize Persuasive Strategies in mHealth Interventions That Promote Physical Activity: Scoping Review and Categorization Overview. J Med Internet Res. 2024 Nov;26:e47774. Available from: https://www.jmir.org/2024/1/e47774.
[61] Zhao Y, Wang Z, Zhu C, Wang S. Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy. In: Moens MF, Huang X, Specia L, Yih SWt, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 4311-23. Available from: https://aclanthology.org/2021.emnlp-main.354/.
[62] Zhao Y, Dastani M, Long J, Wang Z, Wang S. Rescue Conversations from Dead-ends: Efficient Exploration for Task-oriented Dialogue Policy Optimization. Transactions of the Association for Computational Linguistics. 2024 11;12:1578-96. Available from: https://doi.org/10.1162/tacl_a_00717.
[63] Wang H, Wang L, Du Y, Chen L, Zhou J, Wang Y, et al. A survey of the evolution of language model-based dialogue systems. arXiv preprint arXiv:231116789. 2023.
[64] Saravanan A, Tsfasman M, Neerincx MA, Oertel C. Giving social robots a conversational memory

for motivational experience sharing. In: 2022 31st IEEE international conference on robot and human interactive communication (RO-MAN). IEEE; 2022. p. 985-92.
[65] Ligthart ME, Neerincx MA, Hindriks KV. Memory-based personalization for fostering a long-term child-robot relationship. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE; 2022. p. 80-9.
[66] Campos J, Kennedy J, Lehman JF. Challenges in exploiting conversational memory in human-agent interaction. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems; 2018. p. 1649-57.
[67] Drageset J. Social support. Health promotion in health care–Vital theories and research. 2021:137-44.
[68] Mohebi S, Sharifirad G, Feizi A, Botlani S, Hozori M, Azadbakht L. Can health promotion model constructs predict nutritional behavior among diabetic patients? Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences. 2013;18(4):346.
[69] Bøen H, Dalgard OS, Bjertness E. The importance of social support in the associations between psychological distress and somatic health problems and socio-economic factors among older adults living at home: a cross sectional study. BMC geriatrics. 2012;12:1-12.
[70] Steering. Committee for Human Rights in the Fields of Biomedicine and Health (CDBIO). Report on the application of artificial intelligence in healthcare and its impact on the 'patient-doctor' relationship. Council of Europe; 2024.
[71] Delaney E, Pakrashi A, Greene D, Keane MT. Counterfactual explanations for misclassified images: How human and machine explanations differ. Artificial Intelligence. 2023;324:103995.
[72] Keane MT, Kenny EM, Delaney E, Smyth B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:210301035. 2021.
[73] Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Medical Informatics and Decision Making. 2023;23(1):73.
[74] Sparrow R, Hatherley J. High hopes for "Deep Medicine"? AI, economics, and the future of care. Hastings Center Report. 2020;50(1):14-7.