Value-Sensitive Design for AI Technologies: Proposition of Basic Research Principles Based on Social Robotics Research

Theresa Schmiedel, Vivienne Jia Zhong and Janine Jäger

FHNW University of Applied Sciences and Arts Northwestern Switzerland
theresa.schmiedel@fhnw.ch
viviennejia.zhong@fhnw.ch
janine.jaeger@fhnw.ch

Abstract. Artificial intelligence (AI) technologies, such as social robots, increasingly influence economic and social life, with both opportunities and risks. The added value of such technologies depends enormously on how they are designed. In this conceptual paper, we report on our research experience and call for developing AI technologies including human values relevant to the particular stakeholder groups of such technologies as suggested by value-sensitive design (VSD) – an established design approach. To facilitate the application of VSD for designing AI technologies, we propose three basic research principles for VSD and call for more research particularly covering the design of AI technologies, such as social robots and beyond.

Keywords: Value-sensitive Design, Artificial Intelligence, Human Values, AI Technologies, Social Robots.

1 The importance to design AI technologies considering human values

Artificial Intelligence (AI) technologies are increasingly part of all areas of social and economic life (e.g., smart home applications, self-driving cars). AI systems can mimic or even surpass human intelligence and decision-making. They are able to learn independently, interact with their environment, and make decisions. Thus, with increasing autonomy, AI becomes an independent actor in social and economic life. While AI bears innovation potential with high societal benefits in various areas, AI also comes along with unforeseeable risks and challenges for economy and society.

Social robots are a good example to demonstrate the benefits and challenges of AI. Social robots are physically embodied robots designed to engage in social interactions with humans in a socially acceptable way [1]. They engage with humans using various interaction modalities, for example, conversing with users in natural language, recognizing user's face and emotion and reacting accordingly. To enable such humanlike, natural interactions, a wide range of AI technologies (e.g., spoken dialog system, face recognition, motion planning, etc.) are employed and aligned with each other. Social robots are used for various contexts (e.g., in education [2], healthcare [3–5], tourism [6–9]). Across the applications, numerous benefits are reported. Research reports on several positive outcomes (e.g., improvement of social behaviors in children) involving social robots in autism therapy [10] or diverse benefits from using social robots for mental health (e.g., improved mood and loneliness reduction) . Despite these benefits, researchers also raise numerous concerns with regard to the deployment of social robots. For example, researchers have identified several ethical and societal issues related to the use of social robots for therapeutic applications in mental health services (e.g., loss of patient autonomy, lack of guidance on development, emotional dependency on robots) and developed corresponding recommendations [12]. In the same vein, findings from a series of transdisciplinary workshops involving social robot experts indicate further concerns from ethical, social and legal aspects (e.g., replacement of human interaction) [13].

As social robots and other AI technologies have the potential to influence human relationships [12, 14] and shape social practices [15], we need to mitigate potential negative impacts connected to the use of AI. Thus, we need to design AI with human values in mind that reflect what is important to the relevant user groups of each AI technology in its particular use context [16]. Only by employing a value-sensitive approach to AI, the use of AI is sustained in the society.

2 Value-sensitive design

Value-Sensitive Design (VSD) is an established approach that incorporates various user-centered design methods and aims to design technologies in such a way that human values are reflected in the technology design

and its applications [17, 18]. Values are subconscious needs, that members of a group share (e.g., autonomy, security, or collaboration) [19], and which guide their behavior (e.g., to use or avoid certain technologies) [20]. The advantage of VSD is that technology adapts to human needs and not the other way around. In other words, humans shape the technology in a sustainable and proactive way and are not passively influenced or controlled by the technology [21–23]. VSD is thus a proactive approach that translates human values into technological requirements and thus influences the design of technologies early in the development process to incorporate stakeholder-relevant requirements in a fundamental and timely manner [21]. Particularly, VSD consists of a tripartite methodology of empirical, conceptual, and technical investigations. In the conceptual investigation, stakeholders are identified. Moreover, human values from literature and those important to relevant stakeholder groups are identified and conceptualized through analytic, theoretical and philosophical investigation. In the empirical investigation, methods from social science research are used to examine how stakeholders experience values in the particular sociotechnical context. Specially, competing values are prioritized and value conflicts are solved. Design requirements are then formulated. In the technical investigation, one can employ retrospective analyses to evaluate the compliance of an existing technology to the identified values in the conceptual investigation, or design the technology according to the identified values in a proactive way [21, 24]. It is important to stress that the application of the tripartite methodology does not follow a specific order. Researchers can start with any of the three types of investigation. Indeed, it is recommended to apply the three types of investigation in an iterative and integrative manner [21].

A wide range of technologies have benefited from VSD (e.g., [25–27]). In the field of social robotics, several works have advanced the design of value-sensitive human-robot interaction. In her works, van Wynsberghe [28, 29] focuses the application of VSD on care robotics — research on robots that are used related to the care of persons — and develops the "care centered value sensitive design methodology" as well as the "care centered framework" that allow one to design and evaluate care robots from an ethical perspective. Building upon on her works, Umbrello and his colleagues [24, 30] proposed a multi-tiered approach VSD-AI for Social Good (AI4SG) for care robotics that extend the scope of values to be considered (e.g., including UN Sustainable Development Goals as value sources). Beyond the healthcare context, VSD has been applied to social robots in education to uncover parents' moral values with respect to the introduction of social robots in primary schools [31]. Furthermore, various works have been conducted in the field of child-robot interaction to gain an understanding of children's perception of and attribution of moral and social standing to social robots . Finally, using VSD as a research background, Zhong et al. [32] investigated factors that affect robot likeability.

The existing literature shows that VSD research traditionally has a strong background in ethics. However, recent discussion calls on broadening the view of values by giving more attention to the context, which affects stakeholders' view on values [30, 33]. In particular, the context plays a vital role in how stakeholders perceive, interpret and prioritize values relevant to the design of technology [33]. Complementary to the ethical perspective, we take a managerial and contextual viewpoint in our application of VSD.

3 Proposition of three basic research principles for VSD

Based on our research experience on VSD in social robotics, we propose three basic research principles, which we outline next.

3.1 Employ a multi-iterative VSD approach

The first principle concerns the iterative characteristic of VSD, which essentially ensures a technology design that respects stakeholders' values. With regard to VSD as an iterative approach of three types of investigation (conceptual, empirical, technical), a recent review reveals a lack of iteration in most VSD projects and calls for considering this aspect to enhance technology design [34]. While existing research (e.g., [24]) suggests that each design process once contains the three types of investigation of VSD, we argue that the iterative nature of VSD calls for a more frequent iteration of the three investigation types within a design process. Centering values in the core of the technology design, we propose a multi-iterative VSD approach. This multi-iterative approach should follow the identification of stakeholders and their benefits and harms in the particular sociotechnical context. Based on our research on human-robot interactions, we distinguish three main phases of VSD for specifying the multi-iterative approach: 1) value identification, 2) value embedding, and 3) value evaluation. Fig 1 displays the multi-iterative VSD approach we propose as the first basic research principle of applying VSD for designing AI technologies, and particularly social robots.

Value identification. This phase focuses on the identification, conceptualization, and prioritization of values relevant to the stakeholders not only in general but especially in the particular use case context of the particular technology. All three types of investigations can be part of this phase. For instance, the identification of values in HRI can be done *conceptually* by drawing on ethical frameworks identified in the literature [28], but also by conducting *empirical* investigations with stakeholders for the stake of contextual values. Furthermore, *technical* artifacts such as mockups can be used for the elicitation of values as well. Particularly, we learned that mockups are useful for identifying values in HRI as most people do not have any experience in interacting with social robots yet [32]. The outcomes of the empirical and technical investigation inform the conceptual investigation, for instance, with regards to the conceptualization of values. Another important aspect of this phases is addressing potential tensions between values, also through conceptual, empirical, and potentially technical investigations.

Value embedding. In this phase, values are embedded into technology design through translating the values into design requirements and implementing them into the design. Deriving design requirements can be *conceptually* guided by existing frameworks (e.g., [35]), complementing *empirical* investigations involving stakeholders. The operationalization of the design requirements can then draw from theoretical *concepts* (e.g., design patterns [40] for specifying values-sensitive HRI), followed by the *technical* realization in form of prototypes. Before moving on to the next phase of value evaluation, the developed prototype should be tested not only with regard to functional aspects but also with regard to usability aspects. The latter is important to ensure a potential misperception of values in the evaluation phase is not due to a lack of usability. We suggest that the technical prototype shall ensure good usability before conducting the value evaluation. This requires an iterative cycle of *technical* prototyping and *empirical* usability testing. Furthermore, from past projects, we learned that during prototyping, there might be unforeseen technical limitations that require adaptions of the operationalization of design requirements, which requires further iterations within this phase.

Value evaluation. This phase focuses on the *empirical* assessment whether the values embedded into the developed prototype are actually perceived by stakeholders as intended. The results of the evaluation can either confirm the design or indicate two types of required improvement. The first type concerns the insufficient technical specification of the developed prototype (e.g., misguiding robot behaviors), which consequently needs a revision of design requirements. In this case, researchers shall revisit the second phase, value embedding. The second type concerns unanticipated values, which stakeholders become aware of to be relevant, and value tensions, which usually become apparent after the deployment of the prototype [36]. The evaluation ideally spans a period of several weeks to be able to uncover this type of required improvement. If such improvements are necessary, researchers are advised to revisit the first phase, value identification. To avoid serious deviations from stakeholders' expectations over time that lead to a major redesign of the technology, we propose to revisit the three phases frequently.

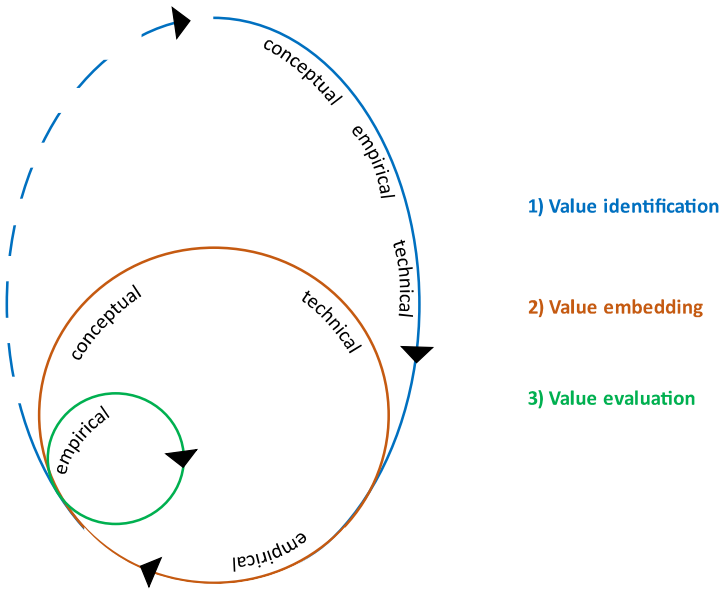

Fig. 1. Specification of the multi-iterative VSD approach

3.2 Start with the use case

The second principle addresses the beginning of a VSD project. Traditionally, VSD does not prescribe how to start a VSD project. Instead, Friedman et al. [36] suggest beginning with the most relevant aspect to the researchers. Therefore, over the past years, we started our VSD projects with the context of use and the technology — social robots. For example, we investigated the use of social robots in public spaces such as a campus library [38], a reception area of a company [37], and the application of social robots in the healthcare sector. We learned that it is highly beneficial to focus on specific social robot use cases to identify values that are relevant to the various stakeholders of a social robot such as in the healthcare context. Even though it is possible to investigate which generic values are important to the application of social robots in the healthcare field, we found that such values are relevant but insufficient for specifying design requirements that foster acceptance and usage of a social robot. For example, in a survey (N = 127) investigating values important in the healthcare context, we found that some universal values such as power and universalism as identified by Schwatz [39] have no importance as opposed to context-specific values such as compassion and respect for person as specified in prior research [40]. Furthermore, participants prioritized the values in a use case-dependent way. Using a robot as a companion in the case of a longer hospital stay, the top three values are compassion, respect for person, hedonism, while the values of security, benevolence and self-direction share the fourth rank. Using a robot as a distraction during needle injection, the top five values are compassion, respect for person, security, hedonism, benevolence and conformity. Our explanation is that the technology can be used for a very diverse set of applications. For example, social robots may serve as assistants in the registration process of patients or they may support staff through providing medical information to patients. In each particular use case, the set of values relevant to the specific stakeholders of a social robot can differ largely. In the example, efficiency may be one of the dominant values relevant in the registration process of patients, while empathy may be especially important when giving medical information to the patients. Thus, we suggest establishing the research principle of a use case-guided specification of stakeholder-relevant values for a particular technology.

3.3 Sustain the knowledge of value-sensitive design

The third principle focuses on perpetuating the findings of VSD projects, so both researchers and engineers can build upon the already generated design knowledge. Some researchers have already generated an overview of VSD projects, such as Friedman and Hendry [21] who categorized VSD projects by application domain, values, and technology. To foster the exchange of the VSD community, in particular, with regard to design practices and knowledge, we suggest to go beyond such categorizations and extend the overview into a technology-specific catalogue of use cases, stakeholder-relevant values, and particularly also the related design requirements, and technical specifications. Such a catalogue can then develop into a reference frame for researchers and engineers for 1) leveraging current best practices and 2) further developing value-sensitive AI technology. Thus, we propose developing a reference frame that shall serve as a basis for technology design, especially shaping the interaction of humans and technology, such as through specifying typical design patterns.

4 Conclusion

Various AI technologies are currently in their infancies and, thus, naturally a predominant focus lies on their technical performance and development. In this context, the end users and affected social stakeholder groups are not necessarily sufficiently considered. Yet, to develop AI technologies in a responsible and sustainable manner requires integrating what these stakeholder groups deem important. While user-centered design approaches focus on integrating the perspectives of the end user, VSD goes beyond in that it involves all stakeholders of the technology and focuses on a value perspective rather than a user requirements perspective.

To ease the design of AI technologies, we propose three basic research principles of VSD. The first research principle employs the multi-iterative VSD approach that consists of 1) value identification, 2) value embedding, and 3) value evaluation. The second research principle starts with the use case to understand and identify contextual values relevant to the stakeholders. The third research principle sustains knowledge of value-sensitive design. The proposed research principles are our first attempt to generalize our learnings from the past VSD projects. Currently, we are refining and further developing the proposed principles, in particular, using social robots as an exemplary technology.

We intend to create awareness for the relevance of values in AI design. Our overall goal is to trigger interdisciplinary exchange and collaboration on the important topic of designing AI with a focus on stakeholder relevancy and societal wellbeing.

References

1. Schmiedel, T., Jäger, J., Zhong, V.J.: Social Robots in Organizational Contexts: The Role of Culture and Future Research Needs. In: Dornberger, R. (ed.) New Trends in Business Information Systems and Technology: Digital Innovation and Digital Business Transformation. pp. 163–177. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-48332-6_11.
2. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: A review. Science Robotics. 3, eaat5954 (2018). https://doi.org/10.1126/scirobotics.aat5954.
3. Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., Horne, N., Gregorio, M., Mann, J., Chaudhury, H.: The benefits of and barriers to using a social robot PARO in care settings: a scoping review. BMC Geriatr. 19, 232 (2019). https://doi.org/10.1186/s12877-019-1244-6.
4. Arent, K., Kruk-Lasocka, J., Niemiec, T., Szczepanowski, R.: Social robot in diagnosis of autism among preschool children. In: 2019 24th International Conference on Methods and Models in Automation and Robotics, MMAR 2019. pp. 652–656 (2019). https://doi.org/10.1109/MMAR.2019.8864666.
5. Riek, L.D.: Healthcare Robotics. Communications of the ACM. 60, 68–78 (2017).
6. Murphy, J., Hofacker, C., Gretzel, U.: Dawning of the age of robots in hospitality and tourism: Challenges for teaching and research. European Journal of Tourism Research. 15, 104–111 (2017).
7. Ivanov, S.H., Webster, C., Berezina, K.: Adoption of Robots and Service Automation by Tourism and Hospitality Companies. Social Science Research Network, Rochester, NY (2017).
8. Chung, M.J.-Y., Cakmak, M.: "How was Your Stay?": Exploring the Use of Robots for Gathering Customer Feedback in the Hospitality Industry. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 947–954. IEEE, Nanjing (2018). https://doi.org/10.1109/ROMAN.2018.8525604.
9. Germak, C., Lupetti, M.L., Giuliano, L., Ng, M.E.K.: Robots and Cultural Heritage: New Museum Experiences. Journal of Science and Technology of the Arts. 7, 47–57 (2015).
10. Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., Pioggia, G.: Autism and social robotics: A systematic review. Autism Res. 9, 165–183 (2016). https://doi.org/10.1002/aur.1527.
11. Rabbitt, S.M., Kazdin, A.E., Scassellati, B.: Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. Clinical Psychology Review. 35, 35–46 (2015). https://doi.org/10.1016/j.cpr.2014.07.001.
12. Fiske, A., Henningsen, P., Buyx, A.: Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. Journal of Medical Internet Research. 21, e13216 (2019). https://doi.org/10.2196/13216.
13. Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A.: Gathering Expert Opinions for Social Robots' Ethical, Legal, and Societal Concerns: Findings from Four International Workshops. Int J of Soc Robotics. 12, 441–458 (2020). https://doi.org/10.1007/s12369-019-00605-z.
14. Volpe, G., Schulte-Althoff, M., Dillmann, D., Maurer, E., Niedenzu, Y., Schließer, P., Fürstenau, D.: Humanoid Social Robots and the Reconfiguration of Customer Service. In: Bandi, R.K., C. R., R., Klein, S., Madon, S., and Monteiro, E. (eds.) The Future of Digital Work: The Challenge of Inequality. pp. 310–325. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-64697-4_23.
15. Samani, H., Saadatian, E., Pang, N., Polydorou, D., Fernando, O.N.N., Nakatsu, R., Koh, J.T.K.V.: Cultural Robotics: The Culture of Robotics and Robotics in Culture Regular Paper. International Journal of Advanced Robotic Systems. 10, 400 (2013). https://doi.org/10.5772/57260.
16. Umbrello, S., De Bellis, A.F.: A Value-Sensitive Design Approach to Intelligent Agents. In: Yampolskiy, R. (ed.) Artificial Intelligence Safety and Security (2018). CRC Press, Rochester, NY (2018).
17. Friedman, B.: Value-sensitive design. interactions. 3, 16–23 (1996). https://doi.org/10.1145/242485.242493.
18. Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A.: Value Sensitive Design and Information Systems. In: Doorn, N., Schuurbiers, D., van de Poel, I., and Gorman, M.E. (eds.) Early Engagement and New Technologies: Opening up the Laboratory. pp. 55–95. Springer Netherlands, Dordrecht (2013). https://doi.org/10.1007/978-94-007-7844-3_4.
19. Ros, M., Schwartz, S.H., Shoshana, S.: Basic Individual Values, Work Values, and the Meaning of Work. Applied psychology. 48, 49–71 (1999).
20. Maio, G.R.: The Psychology of Human Values. Routledge, London (2016).
21. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge, MA (2019).

22. Orlikowski, W.J.: The Duality of Technology: Rethinking the Concept of Technology in Organizations. Organization Science. 3, 398–427 (1992).
23. Šabanović, S.: Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. International Journal of Social Robotics. 2, 439–450 (2010). https://doi.org/10.1007/s12369-010-0066-7.
24. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics. 1, 283–296 (2021). https://doi.org/10.1007/s43681-021-00038-3.
25. Maathuis, I., Niezen, M., Buitenweg, D., Bongers, I.L., van Nieuwenhuizen, C.: Exploring Human Values in the Design of a Web-Based QoL-Instrument for People with Mental Health Problems: A Value Sensitive Design Approach. Sci Eng Ethics. 26, 871–898 (2020). https://doi.org/10.1007/s11948-019-00142-y.
26. Thornton, S.M., Lewis, F.E., Zhang, V., Kochenderfer, M.J., Christian Gerdes, J.: Value Sensitive Design for Autonomous Vehicle Motion Planning. In: 2018 IEEE Intelligent Vehicles Symposium (IV). pp. 1157–1162 (2018). https://doi.org/10.1109/IVS.2018.8500441.
27. Gazzaneo, L., Padovano, A., Umbrello, S.: Designing Smart Operator 4.0 for Human Values: A Value Sensitive Design Approach. Procedia Manufacturing. 42, 219–226 (2020). https://doi.org/10.1016/j.promfg.2020.02.073.
28. van Wynsberghe, A.: Designing Robots for Care: Care Centered Value-Sensitive Design. Science and Engineering Ethics. 19, 407–433 (2013). https://doi.org/10.1007/s11948-011-9343-6.
29. van Wynsberghe, A.: Service Robots, Care Ethics, and Design. Ethics and Information Technology. 18, 311–321 (2016). https://doi.org/10.1007/s10676-016-9409-x.
30. Umbrello, S., Capasso, M., Balistreri, M., Pirni, A., Merenda, F.: Value Sensitive Design to Achieve the UN SDGs with AI: A Case of Elderly Care Robots. Minds & Machines. (2021). https://doi.org/10.1007/s11023-021-09561-y.
31. Smakman, M., Jansen, B., Leunen, J., Konijn, E.: Acceptable social robots in education: A value sensitive parent perspective. In: INTED2020 Proceedings. pp. 7946–7953. International Academy of Technology, Education and Development (IATED), Valencia, Spain (2020). https://doi.org/10.21125/inted.2020.2161.
32. Zhong, V.J., Mürset, N., Jäger, J., Schmiedel, T.: Exploring Variables That Affect Robot Likeability. In: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction. pp. 1140–1145. IEEE Press, Sapporo, Hokkaido, Japan (2022).
33. Saket, M.: Putting Values in Context: an augmentation of Value Sensitive Design (VSD). J. Eth. Emerg. Tech. 31, 1–9 (2021). https://doi.org/10.55613/jeet.v31i2.86.
34. Winkler, T., Spiekermann, S.: Twenty Years of Value Sensitive Design: A Review of Methodological Practices in VSD Projects. Ethics and Information Technology. (2018). https://doi.org/10.1007/s10676-018-9476-2.
35. van de Poel, I.: Translating Values into Design Requirements. In: Michelfelder, D.P., McCarthy, N., and Goldberg, D.E. (eds.) Philosophy and Engineering: Reflections on Practice, Principles and Process. pp. 253–266. Springer Netherlands, Dordrecht (2013). https://doi.org/10.1007/978-94-007-7762-0_20.
36. Friedman, B., Hendry, D.G., Borning, A.: A Survey of Value Sensitive Design Methods. HCI. 11, 63–125 (2017). https://doi.org/10.1561/1100000015.
37. Zhong, V.J., Schmiedel, T.: A User-Centered Agile Approach to the Development of a Real-World Social Robot Application for Reception Areas. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. pp. 76–80. Association for Computing Machinery, Boulder, CO, USA (2021). https://doi.org/10.1145/3434074.3447132.
38. Sabbioni, G., Zhong, V.J., Jäger, J., Schmiedel, T.: May I Show You the Route? Developing a Service Robot Application in a Library Using Design Science Research. In: Ahram, T. and Taiar, R. (eds.) Human Interaction, Emerging Technologies and Future Systems V. pp. 306–313. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-85540-6_39.
39. Schwartz, S.H.: Are There Universal Aspects in the Structure and Contents of Human Values? Journal of Social Issues. 50, 19–45 (1994). https://doi.org/10.1111/j.1540-4560.1994.tb01196.x.
40. Rider, E.A., Kurtz, S., Slade, D., Longmaid, H.E., Ho, M.-J., Pun, J.K., Eggins, S., Branch, W.T.: The International Charter for Human Values in Healthcare: An interprofessional global collaboration to enhance values and communication in healthcare. Patient Education and Counseling. 96, 273–280 (2014). https://doi.org/10.1016/j.pec.2014.06.017.