Team Design Patterns for Moral Decisions in Hybrid Intelligent Systems: A Case Study of Bias Mitigation

Jip J. van Stijn[a], Mark A. Neerincx[b,c], Annette ten Teije[a] and Steven Vethman[c]

[a] *Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*
[b] *Technische Universiteit Delft, Mekelweg 5, 2628 CD Delft, The Netherlands*
[c] *Netherlands Organisation for Applied Scientific Research (TNO), Kampweg 55, 3769 DE Soesterberg, The Netherlands*

Abstract

Increasing automation in the healthcare sector calls for a Hybrid Intelligence (HI) approach to closely study and design the collaboration of humans and autonomous machines. Ensuring that medical HI systems' decision-making is ethical is key. The use of Team Design Patterns (TDPs) can advance this goal by describing successful and reusable configurations of design problems in which decisions have a moral component and facilitating communication in multidisciplinary teams designing HI systems. For this research, TDPs were developed describing a set of solutions for a design problem in a medical HI system: mitigating harmful biases in machine learning algorithms. The Socio-Cognitive Engineering (SCE) methodology was employed, integrating operational demands, human factors knowledge, and a technological analysis into a set of TDPs. A survey was created to assess the usability of the patterns with regards to their understandability, effectiveness, and generalizability. Results showed that TDPs are a useful method to unambiguously describe solutions for diverse HI design problems with a moral component on varying abstraction levels, usable by a heterogeneous group of multidisciplinary researchers. Additionally, results indicated that the SCE approach and the developed questionnaire are suitable methods for creating and assessing TDPs.

Keywords

Hybrid intelligence, socio-cognitive engineering, value-sensitive design, bias mitigation, team design patterns, moral decision-making

1. Introduction

Over the past decades, the healthcare domain has witnessed a steep increase in automation. eHealth applications allow for higher quality and more cost-effective care [1], robot-assisted surgery has proven to be effective and safe in several medical domains [2], and machine learning algorithms are capable of classifying radiology images with malicious cancers better than many radiologists [3]. While automation has unprecedented potential, concerns have been voiced in the academic realm and society, pointing at undesirable effects of recent autonomous systems (e.g. [4]). It is of utmost importance to design such systems carefully to make sure

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

jipvanstijn@gmail.com (J.J. van Stijn); mark.neerincx@tno.nl (M.A. Neerincx); annette.ten.teije@vu.nl (A. ten Teije); steven.vethman@tno.nl (S. Vethman)
0000-0002-1822-8797 (J.J. van Stijn); 0000-0002-8161-5722 (M.A. Neerincx); 0000-0002-9771-8822 (A. ten Teije)

that their actions align with human goals and values. The moral component of autonomous machines has been the subject of recent research, especially in the field of machine ethics. This discipline attempts to contribute to the creation of Artificial Moral Agents (AMAs) that follow certain ethical rules [5]. An important current challenge is posed by the discipline of Hybrid Intelligence (HI), aimed at utilizing the complementary strengths of human and artificial intelligence, so that they can perform better than either of the two separately [6, 7].

Using a HI approach, it is key to not only include machine requirements in the study of AMAs, but also to look into the cognitive capabilities of the humans they perform teamwork with. As this is a highly multidisciplinary endeavor, it is first and foremost imperative to establish a common language to talk about moral situations in HI systems. A promising format for this is the use of Team Design Patterns (TDPs): combinations of text and pictorial language to describe possible solutions to recurring design problems [8]. These patterns represent reusable and generic HI design solutions in a coherent way and are aimed at facilitating the multidisciplinary HI design process. Van Diggelen and Johnson [9] developed a simple and intuitive graphical TDP language, expressing different types of work, different degrees of engagement, and different environmental constraints. Van der Waa et al. [10] made a first attempt to apply this language to the moral domain, describing the task allocation of human-computer teams in morally sensitive situations. However, the TDP language has minimal sufficiency for expressing human cognitive components and requirements, which is a vital element for a truly HI approach toward moral decision making in autonomous systems. Additionally, the TDP method has so far solely been used as a taxonomy and has not been utilized in the design process of HI, making it difficult to evaluate its effects.

1.1. Research aims

The aim of the current research is twofold. Firstly, it attempts to advance the conceptualization of moral decision-making in HI systems in the medical domain by creating Team Design Patterns for the process of bias mitigation in a HI digital assistant for diabetes type II care. These patterns should serve as reusable entities for solving similar design challenges. This paper contributes to the existing TDP literature by addressing two issues: (1) the expression of both human cognitive components and AI requirements for moral decision-making and (2) the development of TDPs not only as a library of successful and reusable design solutions, but also in their application as a method for the design process HI by a multidisciplinary team. In this paper, we study the potential of TDPs as an empirical approach for involving expert knowledge from a wide range of disciplines early in the design process of HI systems.

Secondly, this research aims to contribute to a scientific standard in the methodology of conceptualizing moral decision-making in HI systems, as there is currently no such standard yet. A documentation of the used methods for arranging and evaluating the TDPs may serve as a benchmark in the development of a universal methodological framework for the conceptualization of moral decision-making in human-computer teams.

The two aims mentioned above resulted in two main research questions:

1. How can Team Design Patterns describe moral decision-making for bias mitigation in a medical HI system, so that they are usable by researchers from the various disciplines involved in the design of such systems?

2. Which methodological tools are suitable for the creation of Team Design Patterns?

1.2. Methodology and Research Design

Though important for the universal adoption of the TDP language across disciplines, a standardized set of methods for the creation of Team Design Patterns is still missing. This methodology has several requirements. It should (1) be geared to HI by combining technical AI knowledge with human factors knowledge, (2) be able to incrementally improve the patterns based on its application to new use cases, (3) be able to incorporate moral values of the stakeholders involved in the patterns' application domain.

Socio Cognitive Engineering (SCE) was developed for the design of hybrid intelligent systems, combining elements from cognitive engineering, user-centered design, and requirement analysis [11]. It has been implemented in a wide range of systems in various domains, including a digital support assistant for children with diabetes [12]. An overview of SCE is illustrated in Figure 1.

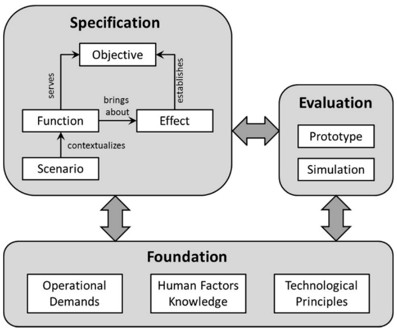


Figure 1: SCE overview.

The SCE methodology is always applied to satisfy a specific demand in a particular context. The *foundation* layer consists of three components. Firstly, it includes an analysis of operational demands, which revolves around inspecting the work domain of the hybrid intelligent system and the support that is needed. Additionally, it includes an analysis of human factors relevant to the system, as well as an analysis of the technological principles that may be appropriate for the envisioned support. In the *specification* component of SCE, a number of objectives of the envisioned system is defined. This leads to the recognition of functions of the system, which are contextualized by scenario-like descriptions of the supposed human-machine interactions. The functions are supposed to bring about certain effects, which are called claims. Lastly, the *evaluation* component uses a prototype or simulation to test whether the specified functions really have the claimed effect. The results of the evaluation can then be utilized to revise and enhance the foundation, specification and evaluation components, incrementally advancing the product. The current research was designed to follow the SCE methodology as presented in [13]. In this adaptation of the methodology, elements of value-sensitive design were included in the foundation and specification layers in order to take the stakeholders' moral values into account throughout the design process.

2. Foundation

2.1. Operational Demands

The current research employs SCE for the development of a system that aids with Diabetes Type II (DT2) care [14]. The operational demands analysis of the SCE methodology revolves around the question: What kind of support is needed in the application domain? [11] The envisioned DT2 system is to provide support to health care professionals and patients in the prevention, diagnosis, treatment, and management of the disease through models based on existing patient data and domain knowledge. These models may, for example, be geared towards predicting patient's risk of developing diabetes or suggesting a diagnosis. Alternatively, they may predict the best type and dose of medicine or type of lifestyle change to make the disease as unintrusive as possible. Through several modules, these predictions or suggestions are presented to the relevant patient or healthcare professional. Finally, the models are improved and updated by patients' medical and behavioral data. Our focus is on the support of moral decision-making, therefore the next step is to localize processes in this system in which actors face choices that have a moral component. This analysis was based on medical guidelines, first drafts of the envisioned system, and interviews with four experts in the domain of lifestyle-related diseases and their care. We identified three design challenges in which moral decision-making plays a large role: (1) the mitigation of harmful biases in learning algorithms, (2) the sharing of patients' medical and behavioral data and the recording of consents, and (3) suggestions and interventions into the patients' lifestyle. In this paper we will only discuss the former.

Bias mitigation is necessary when learning models develop biases that may result in unfair treatment by the system. For example, the underrepresentation of certain ethnic groups in the input data can result in a racial bias in the system. People of those ethnic backgrounds may then receive worse care than others. The system's developers can employ techniques to mitigate the harmful bias (see section 2.2), but this usually results in a lower average accuracy of the predictions [15]. Hence, the value-tension in this moral issue is between the system's (average) effectivity and fair treatment of each patient.

2.2. Analysis of Technological Principles

To incorporate bias mitigation in HI team design patterns, an understanding of bias mitigation in AI systems is required. The science of fair machine learning is a recent and upcoming science that faces the challenge of formalizing the definition of fairness and the related bias [16]. Two types of formulations for fairness are widely used: the statistically defined group fairness and the more locally defined individual fairness [16, 17]. Group fairness measures assess whether subjects in a demographic group are classified similarly according to statistical measures relative to other groups in the sample or population. Measures often relate to statistical parity, which requires the chance of receiving a false positive or false negative to be independent from certain sensitive features such as ethnicity, gender or sexuality [18]. Individual fairness, in contrast, entails that people with similar traits with respect to a certain task should be treated similarly. This is usually measured by a context-specific distance metric [18]. These two metrics are often at par with each other, illustrating that there is no firm consensus for a universal approach to quantifying fairness [19].

Although there is no consensus on the formalization of fairness, a substantial amount of research has focused on several measure- and context-specific solutions to tackle bias and unfairness. In a comprehensive review of the currently available methods for reducing unfairness in machine learning algorithms, [20] identify three types of methods based on where the bias mitigation is performed: pre-processing, in-processing, or post-processing. Pre-processing methods focus on bias in the training data, e.g. [15], while in-processing methods aim to modify the algorithm to reduce unfair prediction in the learning phase, e.g. [21] Post-processing methods make adjustments after learning to satisfy fairness constraints, e.g.[22].

Regardless of the exact formalization and method chosen, fairness is usually an objective that exists alongside a conventional utility measure, such as accuracy. These competing objectives then result in a moral question to which extent the optimization of average effectiveness is allowed to diverge from the ignorant solution to comply to the additional fairness constraint. Independent of a specific measure or method, we therefore focus on incorporating this important aspect in fair machine learning and bias mitigation: the trade-off between fairness measures and overall utility [23, 24]. This provides the TDP with the flexibility and versatility necessary for the context-dependent nature of bias and new developments in the field of fair machine learning.

2.3. Human Factors Analysis

McLaughlin et al. [25] identify five main cognitive processes: attention, memory, perception, decision-making, and knowledge aids. Each of these classes consists of several more specific processes. For example, attention aids can support humans in selective, orienting, sustained, or divided attention. Moral decision-making is a complex mental and social process for which philosophers and scientists do not have a univocal explanation. It is apparent, however, that the process requires all five categories specified by [25]. We aim to use these five main cognitive processes to conceptualize technological requirements in the Team Design Pattern language.

3. Specification

In the specification phase of SCE, the foundational knowledge was brought together into TDPs describing possible solutions for the bias mitigation design problem. The operational demands and technology analyses were used to create the patterns' team process configurations. The (dis)advantages of each pattern address the value tensions identified in the analysis of operational demands. The taxonomy of cognitive aids [25] was used for the patterns' human requirements, and the hybrid AI boxology [26] was used for to identify AI requirements. Table 1 shows a simple and common TDP to solve the bias mitigation design challenge, while Table 2 depicts a solution which requires more advanced technology. While addressing this challenge, it became clear that other design challenges could be nested within these patterns. For example, in Table 2, the human and machine agent have a joint responsibility to change the model, which presents a design challenge of its own. We designed several *sub-patterns* to address this sub-challenge, one of which is shown in Table 3.

Table 1
Pattern 1: The Human Moral Decision-Maker

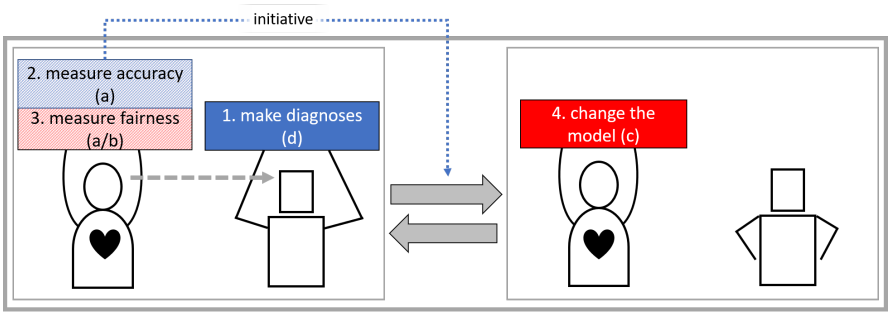
Name:	**Human Moral Decision Maker**
Description:	In this pattern, the machine agent solely performs a machine learning task: predicting the diagnosis (e.g. diabetes) of patients. The human AI developer supervises this process, measuring both the overall accuracy and the fairness of the predictions. If the human agent thinks that the balance between these two measures is off (e.g. because people with a specific ethnic background receive significantly less accurate diabetes diagnoses), the human initiates a takeover. In this takeover, the machine stops its task, while the human changes the model.
Structure:	 initiative 2. measure accuracy (a) 3. measure fairness (a/b) 1. make diagnoses (d) 4. change the model (c) 1. Machine uses learning models to make predictions about diagnosis. 2. Human performs task supervision: Are the predictions accurate? 3. Human performs moral supervision: are the predictions accurate for everybody? Is there social discrimination based on subgroups? 4. If the balance between overall accuracy and fairness of the model is off, the human initiates a takeover, deciding whether and how to change the model.
Human req.:	(a) Sufficient working memory for task supervision and moral supervision (b) Sufficient moral attention to recognize morally sensitive situation (c) Sufficient moral knowledge and domain knowledge to make moral decision
AI req.:	(d) Machine Learning
Advantages:	+ Human is accountable for recognizing and making moral decision + Machine does not require moral competencies
Disadvantages:	— Cognitive under- or overload of the human may result in missing moral choice situations or optional solutions — Human may be turned into moral scapegoat

Table 2
Pattern 2: The Coactive Moral Decision-Maker

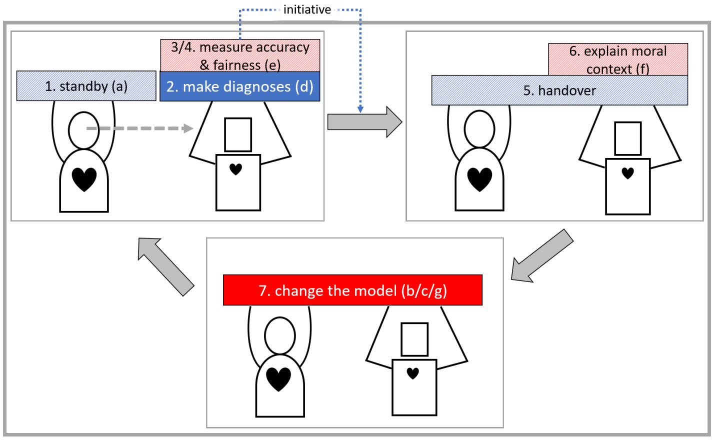
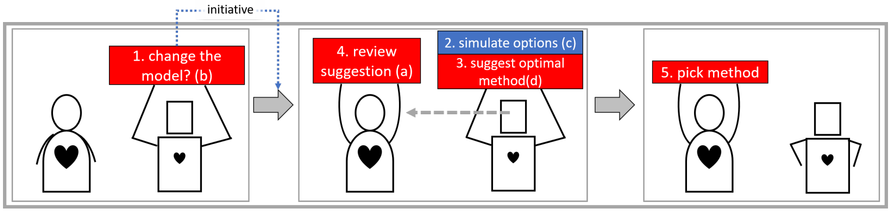
Name:	**The Coactive Moral Decision-Maker**
Description:	In this pattern, the machine has more moral responsibilities. The machine performs moral supervision on itself: it measures whether its own predictions are equally accurate for each subgroup (e.g. if diabetes diagnoses are equally accurate for patients from different etnical backgrounds). The human is on stand-by. If the machine measures bias in its own predictions, it initiates a handover. The machine explains which exceeded thresholds necessitate a moral decision (e.g. because the model is racially biased towards people with a specific ethnic background). The human and the computer then make a joint decision in changing the model.
Structure:	 initiative 3/4. measure accuracy & fairness (e) 1. standby (a) 2. make diagnoses (d) 6. explain moral context (f) 5. handover 7. change the model (b/c/g) 1. Human is on stand-by 2. Machine uses learning models to make predictions about diagnosis 3. Machine performs task supervision: Are the predictions accurate? 4. Machine performs moral supervision: Are the predictions accurate for everybody? Is there social discrimination based on subgroups? 5. If the balance between overall accuracy and fairness of the model is outside preset human-made boundaries, the machine initiates a handover. 6. Machine explains the moral context: which preset thresholds are exceeded? 7. Human and machine jointly decide whether and how to change the model.
Human req.:	(a) Sufficient trust in machine to recognize morally sensitive situations (b) Sufficient understanding of moral implications (c) Sufficient moral knowledge and domain knowledge to make moral decision
AI req.:	(d) Machine Learning (e) Ability to recognize morally sensitive situations (f) Ability to sufficiently explain the moral context (g) Moral decision-support
Advantages:	+ Human is on stand-by, allowing them to do different tasks + Human is accountable for moral consequences, but can receive support
Disadvantages:	— Morally sensitive situations may not be recognized by the machine — Human may be biased by machine's explanations and suggestions

Table 3
Pattern 2.1: The Suggesting Machine

Name:	**The Suggesting Machine**
Description:	In this pattern, the machine has more moral responsibilities, while the human only has a reviewing role. In the first frame, not the human, but the machine decides whether a model change is desirable, based on preset conditions (e.g. if the learning model's diabetes diagnoses are over 10% less accurate for people with a specific ethnic background). After this, the machine simulates all possible methods to change the model, and their effects on the accuracy-fairness trade-off. The machine then suggests the optimal method, which the human reviews. The human agent takes this into consideration, and finally picks the preferred method.
Structure:	 initiative 1. change the model? (b) 4. review suggestion (a) 2. simulate options (c) 3. suggest optimal method(d) 5. pick method 1. Machine considers whether a model change is necessary, based on preset rules. 2. If so, it initiates a transition. The machine simulates all possible methods to mitigate bias. 3. Machine provides decision support: it gives a suggestion of the optimal method. 4. Human reviews this suggestion. 5. Human picks the method to change the model.
Human req.:	(a) Sufficient trust in the machine agent's suggestions
AI req.:	(b) Ability to recognize and take responsibility for moral choice (c) Capability to simulate the effects of all possible options (d) Sufficient moral understanding for picking the optimal choice
Advantages:	+ Low cognitive demands for human + Clearly defined boundaries to which situations demand a moral response
Disadvantages:	— High demands for machine's computing power — Risk of machine missing morally sensitive situations if it is not included in preset boundaries — Human overtrust in machine may result in little moral deliberation — Machine only considers premade set of moral responses, and cannot think 'outside the box'

4. Evaluation

4.1. Data collection and analysis

A usability evaluation was performed with four metrics: (1) understandability, (2) coherency, (3) effectiveness, and (4) generalizability. A questionnaire was filled in by the pattern's prospective direct users: researchers and designers of HI systems. Thirty researchers and developers from various research areas at TNO were approached through a network sample. Twenty of them responded, taking around 45 minutes to fill in the questionnaire. The first section of questions consisted of five-point Likert scale questions inquiring the participant's background knowledge in several disciplines relevant to moral decision-making in HI systems. After that, a video was presented explaining the basic elements of TDPs for moral decision-making. The next section of the questionnaire presented the bias mitigation design challenge as described in section 2.1, including its relevance, the actors involved, and the moral tension between beneficence and fairness it entails. It then presented the proposed patterns, each with four Likert scale statements addressing the metrics mentioned above. Each of the statements was followed by the prompt 'Please explain your answer' and a long answer text field, resulting in a combination of quantitative and qualitative data for each pattern. Additionally, participants were asked which tasks or concepts were missing in or should be added to the patterns.

For the quantitative part of the analysis, the Likert scale questions regarding background knowledge and the ratings regarding the metrics (understandability, coherency, effectiveness, and generalizability) of each of the patterns were analyzed. Due to the nonparametric nature of Likert scale data and the small sample size (N=20), the mode and median were used as an indication for the distribution of the responses. For the same reason, Spearman's rank test was performed to test for correlations between variables, while Wilcoxon's ranked sum test was used to test for statistically significant differences between ratings for the patterns.

The qualitative analysis aimed to get insight into the metrics described above. Additionally, it aimed to reveal concepts and requirements that are still missing from the patterns from the perspective of their anticipated users. Hence, the qualitative data was analyzed thematically and largely data driven. The four metrics were used as predetermined themes, in which subthemes were inferred by categorizing the responses on an increasingly abstract level.

4.2. Results

Distributions of the participants' self-reported background knowledge in the prespecified disciplines showed that (self-reported) experts from all specified research disciplines were present in the participant sample. This suggests that the sample was a good representation of the user target group: researchers and designers of hybrid intelligent systems with varying knowledge in the related disciplines. Overall, the majority of respondents was positive regarding the understandability of all patterns. Pattern 2, with a mode of 4, was rated less understandable than pattern 1 (p=0.055) and pattern 2.1 (p=0.011), possibly due to its higher complexity. The fact that there were no correlations between self-reported background knowledge and understandability suggests that the patterns were understandable for researchers and designers regardless of their area of expertise, which is a key requirement for their purpose of facilitating communication between disciplines. However, the qualitative data indicated that more specific examples would benefit the understanding of some of the participants with little background in AI.

Participants were largely positive regarding the coherency and generalizability of the patterns (with modes of 4 and 5). Ratings for both measures were stable regardless of the specific pattern, with high internal correlations (0.50<r<0.75, p<0.02). This underwrites that structure and reusability are among the main strengths of the TDP methodology, as indicated in [8].

Effectiveness scores of all patterns were low compared to the other measures, with a mode of 3 (=neutral). This may partly be explained by ambiguous phrasing of the question, as it was not clearly defined what 'appropriate moral decision-making' meant. Participants expressed through qualitative responses that they found it difficult to judge the pattern's effectiveness this early in the design process, without any knowledge of the specific implementation. Several participants noted, however, that the patterns resulted in the discussion and explication of important moral requirements of the HI system, which may be a more appropriate effectiveness measure in the early design phase. Some requested to make the patterns more specific, which resonates with an important tension that runs through the core of TDPs, between their requirement to be generalizable, abstract, and intuitive on the one hand, and the desire to include ample detailed information to make them applicable to specific situations on the other.

5. Conclusions and future work

The SCE methodology was used to combine state-of-the-art knowledge of human factors, available technologies, and operational demands to create a set of patterns for moral design problems. This research has provided reusable conceptualizations for solving the bias mitigation (and the data sharing design problems in the full report of this research [27]) from a Hybrid Intelligence perspective. It demonstrated that the TDP language [8] is suitable for the unambiguous description and communication of design problems and their (hypothetical) solutions. In this paper we show that this was possible for the bias mitigation problem on varying degrees of abstraction. The research illustrated that TDPs can be created and applied in the design process of actual Hybrid Intelligence systems and provided an indication that the created patterns were usable for researchers and designers from a range of relevant disciplines. A main strength of these solutions is that they can be improved iteratively in the design process of a HI system, as well as through their application to other, similar design challenges. One such challenge is the identification of more complex biases towards unaccounted (combinations of) properties.

The TDP language would benefit considerably by the inclusion of several concepts. Particularly, references to the interaction design would likely make the patterns more understandable and concrete. Connections to legal requirements would draw the users' attention to laws that may be important for the systems' design. The inclusion of evaluation loops and the possibility to go "outside" the pattern and communicate with external actors (development team, organization, auditing committee, legislator, etc.) would likely lead to improved moral decision-making. Additionally, future work may focus on the development of a tool specifically geared towards creating TDPs. A specialized tool for the creation of the patterns is crucial for a universal adoption of the TDP language and the building of a library of reusable design patterns.

A final line of research that would greatly contribute to the development of ethical HI systems is the study of AI requirements in the TDP language. The current research addressed high-level functions the machine agent should have to take various support roles, including 'the ability to explain moral consequences' and 'the ability to recognize morally sensitive situations.' A lot of progress is yet to be made in formalizing these functions.

Acknowledgments

The study was part of the TNO FATE-project; we thank the TNO researchers for their contribution. It was further supported by the Hybrid Intelligence Center, funded by the Dutch Ministry of Education, Culture and Science (through NWO).

References

[1] N. J. Elbert, H. van Os-Medendorp, W. van Renselaar, A. G. Ekeland, L. Hakkaart-van Roijen, H. Raat, T. E. Nijsten, S. G. Pasmans, Effectiveness and cost-effectiveness of ehealth interventions in somatic diseases: a systematic review of systematic reviews and meta-analyses, Journal of medical Internet research 16 (2014) e110.

[2] T. L. Ghezzi, O. C. Corleta, 30 years of robotic surgery, World journal of surgery 40 (2016) 2550–2557. doi:10.1007/s00268-016-3543-9.

[3] P. Lakhani, A. B. Prater, R. K. Hutson, K. P. Andriole, K. J. Dreyer, J. Morey, L. M. Prevedello, T. J. Clark, J. R. Geis, J. N. Itri, et al., Machine learning in radiology: applications beyond image interpretation, Journal of the American College of Radiology 15 (2018) 350–359. doi:10.1016/j.jacr.2017.09.044.

[4] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453.

[5] M. Anderson, S. L. Anderson, Machine ethics: Creating an ethical intelligent agent, Ai Magazine 28 (2007) 15–15.

[6] M. M. Peeters, J. van Diggelen, K. Van Den Bosch, A. Bronkhorst, M. A. Neerincx, J. M. Schraagen, S. Raaijmakers, Hybrid collective intelligence in a human–ai society, AI & SOCIETY (2020) 1–22.

[7] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, et al., A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, Computer 53 (2020) 18–28.

[8] J. Van Diggelen, M. Neerincx, M. Peeters, J. M. Schraagen, Developing effective and resilient human-agent teamwork using team design patterns, IEEE int. syst. 34 (2018) 15–24.

[9] J. van Diggelen, M. Johnson, Team design patterns, in: Proc. of the 7th Int. Conf. on Human-Agent Interaction, 2019, pp. 118–126.

[10] J. van der Waa, J. van Diggelen, L. C. Siebert, M. Neerincx, C. Jonker, Allocation of moral decision-making in human-agent teams: A pattern approach, in: Int. Conf. on Human-Computer Interaction, Springer, 2020, pp. 203–220.

[11] M. Neerincx, W. Vught, O. Blanson Henkemans, E. Oleari, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer, D. Fumagalli, Socio-cognitive engineering of a robotic partner for child's diabetes self-management, Frontiers in Robotics and AI 6 (2019) 118.

[12] R. Looije, M. A. Neerincx, K. V. Hindriks, Specifying and testing the design rationale of social robots for behavior change in children, Cogn. Syst. Research 43 (2017) 250–265.

[13] M. Harbers, M. A. Neerincx, Value sensitive design of a virtual assistant for workload harmonization in teams, Cognition, Technology & Work 19 (2017) 329–343.

[14] J. de Greeff, M. de Boer, F. Hillerström, F. Bomhof, W. Jorritsma, M. Neerincx, The FATE system: Fair, transparent and explainable decision making, in: AAAI-MAKE, 2021.

[15] F. du Pin Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, 2017, pp. 3992–4001.

[16] R. Binns, Fairness in machine learning: Lessons from political philosophy, in: Conf. on Fairness, Accountability and Transparency, PMLR, 2018, pp. 149–159.

[17] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, volume 28 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2013, pp. 325–333. URL: http://proceedings.mlr.press/v28/zemel13.html.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proc. of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[19] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices, in: Y. Guo, F. Farooq (Eds.), Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, ACM, 2018, pp. 2239–2248.

[20] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proc. of the Conf. on Fairness, Accountability, and Transparency, ACM, 2019, pp. 329–338. doi:10.1145/3287560.3287589.

[21] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: P. A. Flach, T. D. Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, volume 7524 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 35–50. doi:10.1007/978-3-642-33486-3_3.

[22] M. P. Kim, A. Ghorbani, J. Zou, Multiaccuracy: Black-box post-processing for fairness in classification, in: Proc.AAAI/ACM Conf. on AI, Ethics, and Society, 2019, pp. 247–254.

[23] J. S. Kim, J. Chen, A. Talwalkar, Fact: A diagnostic for group fairness trade-offs, in: Int. Conf. on Machine Learning, PMLR, 2020, pp. 5264–5274.

[24] A. K. Menon, R. C. Williamson, The cost of fairness in binary classification, in: Conf. on Fairness, Accountability and Transparency, 2018, pp. 107–118.

[25] A. C. McLaughlin, V. E. Byrne, A fundamental cognitive taxonomy for cognition aids, Human Factors 62 (2020) 865–873. doi:10.1177/0018720820920099.

[26] F. van Harmelen, A. ten Teije, A boxology of design patterns forhybrid learningand reasoning systems, J. Web Eng. 18 (2019) 97–124. doi:10.13052/jwe1540-9589.18133.

[27] J. van Stijn, Moral decision-making in medical hybrid intelligent systems: A team design patterns approach to the bias mitigation and data sharing design problems, 2021. arXiv:2102.11211, MSc. thesis.