Preparing Psychologists and Social Workers for the Daily use of AI

Fredrik Åhs[1], Peter Mozelius[2] and Majen Espvall[1]
[1]Department of Psychology and Social Work, Mid Sweden University, Sweden
[2]Department of Computer and System Science, Mid Sweden University, Sweden
Peter.Mozelius@miun.se

Abstract: A daily use of Artificial Intelligence (AI) is becoming a fact in many fields today, and two of them are psychology and social work. At the same time as AI systems are used for predicting psychological treatments and for decisions in social welfare, higher education has few AI courses for these professions. Moreover, there are several examples in these fields where AI can make unethical decisions that need to be corrected by humans. To better understand the possibilities and challenges of AI in psychology and social work, professional users of AI services need a tailored education on how the underlying technology works. The aim of this paper is to present a project concept for the design and evaluation of a novel course in AI for professional development in psychology and social work. For the design and development of the course the guiding research question should be: What are the strengths and challenges with contemporary AI techniques regarding prediction, adaptivity and decision systems? The suggested AI course should be given as a technology enhanced online training to enable the idea of anytime and anywhere for full-time working participants. Course content and activities are divided into the four separate sections of: 1) The history of AI structured around the 'Three waves of AI', with o focus on the current third wave. 2) A section with a focus on AI techniques for prediction and adaptivity. Underlying techniques such as machine learning, neural networks, and deep learning will be conceptually described and discussed, but not on a detailed level. 3) An elaborated discussion on the relevance, usefulness and trust, and the at the difference between AI-based decision systems and AI-based decision support systems. 4) Finally, the fourth section should comprise the ethical aspects of AI, and discuss transparency and Explainable AI. An innovative approach of the project is to use a neuroscientific assessment of the education to understand how the education changes brain function relevant to evaluate AI based decision. This should be complemented with a qualitative evaluation based on semi-structured interviews.

Keywords: artificial intelligence, AI, human compatible AI, professional development, explainable artificial intelligence, XAI

1. Introduction and aims

In the rapid development of Artificial Intelligence (AI), the implementation of new AI-based services is an ongoing process in many fields today (Zhang & Lu, 2021). Two of them are psychology (Fiske, Henningsen & Buyx, 2019; Taylor & Taylor, 2021), and social work (Goldkind, 2021; Hodgson et al., 2021), where AI-services such as expert systems and neural networks have been part of the discussion since the 1990s (Patterson & Cloud, 1999; Goldkind, 2019). AI systems are used in these professions for predicting psychological treatments, and for decision making in social welfare, but there are still few AI courses for these professions in higher education.

In the Swedish context, society faces a challenge when it comes to the demographical development, and that the number of inhabitants that will need welfare services increases. The national e-health strategy has stated that Sweden should be best in the world to use the possibilities of digitalisation to further develop e-health. Digitalisation is described as an important part of the puzzle to make our welfare services more efficient and useful. A potential positive effect of digitalisation in this area is the possibility to offer welfare services to a larger number of citizens (Edbacken & Vernmark, 2021). Softbots or AI assistants, are used by social services and around 48% of the municipalities in Sweden have implemented softbots for decision support (Socialstyrelsen, 2019). It has been argued that softbots can make the decisions more neutral (Wihlborg, Larsson and Hedström, 2016; Ranerup & Henriksen, 2019), but there is also research reporting about the risks. There are studies that have reported on the risk of digital decision-making systems that can concentrate human bias (Goldkind, 2021), and also on the risk of bias in the algorithms themselves (Meilvang & Dahler, 2022). Moreover, Nordesjö, Scaramuzzino and Ulmestig (2021) describe how structural aspects of poverty are neglected, and that the algorithms that are used include hidden values, such as racism (Svensson, 2019).

The European Union Agency for Fundamental rights has recently written a report about AI and fundamental rights. They argue that AI will create new opportunities, but that it also challenges and threats humans and their fundamental rights. Furthermore, they emphasize the importance of increased attention to issues related to AI and fundamental rights. Kahneman et al. (2021) describe that we always are influenced by an invisible problem when we make decisions, namely noise. This flaw in human judgement has to be something that we should learn to handle when we collaborate with devices such as robots and softbots. This development already has had an effect of the professionals that work in the welfare sector such as social workers and psychologist.

Artificial intelligence (AI) is increasingly used in the welfare sector to judge if individuals fulfil requirements to receive financial aid or other societal assistance. Also in healthcare, usage of AI is predicted to increase in the future to assist in prognosis of treatment outcomes and diagnosis. These changes will have profound impact on how work is performed in the caring professions. While AI assistance can save time that can be used for other work activities, such as meeting clients, there are also known pitfalls that need to be adequately dealt with. It is for example known that racial biases in AI judgments can occur. Therefore, there is a need to teach health care professionals and social welfare workers how AI works, and when AI decisions may need to be corrected by human personnel.

The overall aim of this paper is to present a project concept for the design and evaluation of a novel course in AI for professional development in psychology and social work. The specific aims for the project idea are:

Aim 1: Determine if AI education can change psychologists' and social workers' perception and usage of AI assistance. Course participants will be interviewed following course completion to assess their perception of AI and their confidence in using AI assistance. A group that did not participate in the course will also be interviewed for comparison.

Aim 2: Dissociate trustworthiness judgements and brain responses to social agents and computers following a course in AI. Trustworthiness of information given by a human or an AI assistant will be compared. Brain responses to information from a human and an AI assistant will be compared using functional magnetic resonance imaging (fMRI). A socioemotional neural network, including the temporoparietal junction, is expected to differentiate between information sources. One group of participants that have passed a course in AI will be compared to a group that have not taken the course. It is predicted that following the course in AI, differences in neural responses to information from human and AI assistants will diminish.

Aim 3: Determine differences in brain responses to transparent (white-box) and non-transparent (black-box) AI assistance. We predict that course participants will be better at discriminating between white-box (transparent) and black-box (hidden) AI methods, than psychologists and social workers that did not take the course. They will also be better to judge when AI decisions need to be checked by humans. We predict that increased discrimination between white-box and black-box methods will be accompanied by increased neural responses in networks important for working memory and saliency detection.

2. Course design

Human Compatible AI for Psychology and Social Work, should be developed as a 2,5 hp (roughly 2 week fulltime) course, that introduces the opportunities and challenges with contemporary AI techniques. A course with a design suitable for full-time working professionals in the fields of psychology and social work. The course should be given at 25% study pace with around 30 participants. Technology enhanced and asynchronous self-studies are facilitated by four online meetings with synchronous activities in the video conferencing system Zoom. Three full-day meetings online, should involve a mix of lectures, workshops, discussions and facilitating, complemented by a final online seminar. The suggested virtual learning environments are the Moodle platform for asynchronous activities, and the Zoom conferencing system for synchronous learning and teaching activities.

The suggested pedagogical approach is heutagogy, based on technology enhanced and self-directed collaborative learning (Blaschke, 2021). With the heutagogical idea that theoretical introductions are complemented with hand-on workshops where students can choose subjects that they find relevant and useful for their daily work-life. In the workshop activities, students should explore various existing real-world systems and tools in the AI field. Besides the practical assignments, there will also be essay assignments where students should reflect on the opportunities and obstacles with AI techniques in the fields of psychology and social work. The course is divided into for sub-sections that all are described more in detail, one by one here below.

The first initial course section should be built around the fact that AI on several occasions has been optimistically presented as something that will change our daily living (Russel, 2019; Ågren, 2019). Which of the promising AI techniques that have had actual contributions could be related to the idea of humans tending to overestimate the impact of technology in the short perspective, but to underestimate the long-term effects (Wärnestål, 2021). An introduction of the AI history will be structured around the so called 'Three waves of AI', with o focus on the current third wave (Ågren, 2019). Which concepts in the two first waves have been realized into useful daily

techniques, and which concepts in the ongoing third wave can we expect as mature technology in the near future? Moreover, the first section will introduce important AI concepts and terminology that also should be assessed in an auto-correcting online assignment.

In the second course section, content and activities will focus on AI techniques for prediction and adaptivity. Underlying techniques such as machine learning, neural networks, and deep learning will be conceptually described and discussed, but not on a detailed level. The relatively short technical presentation will be followed up by a more elaborated discussion on relevance, usefulness and trust. A central theme in the second course second should be the idea of a 'Human compatible AI', and the need for human control of AI systems (Russell, 2019). This section will also include a discussion on the-state-of-the-art of AI systems in psychology and social work.

Main concepts to be introduced in the following third course section are agency and collaboration. Important to look at the difference between AI-based decision systems and AI-based decision support systems, combined with the discussion on if AI and humans together is stronger and more reliable than just one of them (Humble & Mozelius, 2019). Furthermore, this section will also involve a presentation of contemporary research ideas on how humans successfully can coexist with the increasing number of AI system in our daily lives. Without getting carried away into science fiction the third course section will bring up how AI might proceed in the future, and which conceptual breakthroughs that are expected to come.

Finally, the fourth course section should bring up the ethical aspects of AI that often are neglected. One important issue is the one on biased data and biased AI-systems (Wärnestål, 2021), another is the quest for more transparent AI systems to replace the frequent blackbox design. A widely accepted alternative here is Explainable artificial intelligence, where the ingoing parameters are visible for the end users (Åhs, Mozelius & Dobslaw, 2020). Furthermore, there are also the more philosophical aspects of the relationship between humans and AI systems, and the widely discussed question regarding if, and when 'Strong AI' or Artificial general intelligence (AGI) will surpass human intelligence. What will be at risk of 'Overly intelligent AI', if that happens? (Bostrom, 2017; Tegmark, 2017). Which types of AI systems might meet the specified requirement of a 'Provably beneficial AI' (Russell, 2019)?

As for many other short courses on professional development the suggested grading scale is Pass/Fail. However, for the further evaluation course participants' learning outcomes will be measured by a 10 graded scale.

Learning objectives: After a completed course the participant should be able to:

- Compare and discuss the similarities and the differences between the three waves of AI
- Understand and use the fundamental terminology in the field of artificial intelligence
- Tell the difference between predictive AI systems, decision systems and decision support system
- Analyse and discuss ethical aspects of transparent and non-transparent AI systems, and the idea of a provably beneficial AI
- Describe and present an overview of existing AI services in the fields of psychology or social work

Course literature

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Ågren, P. O. 2019. *Den tredje AI-vågen: Essäer om AI, samhället och individen (The third wave of AI: Essays on AI, society and the individual)*. BoD-Books on Demand.

3. Evaluation design

A total of 60 participants will be recruited for the project. Thirty of these will be randomized to the AI education group, while the others will serve as a control group. Half of the participants will be working as psychologists and the other half as social workers in the Swedish counties of Jämtland, Västernorrland, and Västerbotten. To ensure that the participants are well acquainted with professional practice and have a larger part of their professional life ahead of them (for any follow-up studies), participants with 3 to 10 years of professional experience will be included. We strive for a heterogeneous group of professionals, but the participants will not

be included with regard to gender. Because the study includes a Magnetic Resonance (MR) examination, exclusion criteria are: contraindications for MR, age below 18 or over 65 years, presence of severe somatic disease or serious psychiatric disorder such as psychosis or severe major depression, treatment for any psychiatric disorder (ongoing or terminated within three months), pregnancy, menopause, and drug or alcohol abuse/dependency.

Moreover, qualitative research interviews will be conducted with the course participants 1-2 months after the end of the course. The purpose of the interviews is to obtain in-depth knowledge of the course participants' learning processes and the applicability of the course content in professional life. Particular attention is paid to critical and ethical aspects and the risks of various forms of discrimination in the use of AI technology in professional practice. More detailed interview questions will also be prepared on the basis of essay assignment where the course participants reflect on opportunities and obstacles by using AI technology in the fields of psychology and social work. The qualitative research interviews, which are estimated to take approximately one hour, will be conducted over Zoom. Transcribed interviews will then be analyzed using qualitative content analysis.

Functional magnetic resonance imaging: Two experiments will be performed. The first fMRI experiment, will consist of a single-card poker game where participants either play a computer or a human agent, as in the study of Carter et al (2012). Briefly, each trial begins with a picture of the opponent (i.e., a photograph of the human opponent or of a computer). The participant's card is then revealed, and they decide whether to bet or fold. If the participant chooses to fold, the trial ends. If the participant chooses to bet on the card, control passes to the opponent who has 3 s to decide whether or not to match the participant's bet or fold. Results from this experiment has shown that the temporo-parietal junction is a critical part of the brain for integrating information about the agent that you play, human or computer, and chances of success in the game. We predict that the group of individuals who received. AI education will process the information of agency (human or computer) with similar activity level in the temporo-parietal junction as compared to the control group. As a control task, participants will view pictures of faces displaying emotional expressions. In this task, we expect the same pattern of responses in both groups.

In a second fMRI experiment, participants will read scenarios of situations where an AI agent has made decisions regarding social welfare or treatment given to a fictional client. The participants will get information on the socioeconomic and ethnic background of the client. In some cases, information on what parameters were important for the AI decision will be given (White-box), but in others they will just be given the accuracy with which the AI algorithm is expected to perform. We will compare brain responses to scenarios where the ethnic background of the client is non-Swedish born to scenarios where the ethnic background is Swedish born. We predict that participants who have taken the AI course will engage working memory networks to a larger degree than the control group who did not receive AI education. We also predict that the saliency network (amygdala, insula, anterior cingulate cortex) will be less activated in participants who received AI education than in the control group. This would be consistent with the group that received AI education making more informed judgements.

4. Concluding discussion

There is an explosive increase in the development and usage of AI methods in the health and welfare sector. Results from the proposed research program can therefore be predicted to capture the attention of a large audience. The ground-breaking usage of neuroimaging to understand how brain functions discriminate between human and AI assistance is also of a general interest. Findings from this neuroscientific investigation will aid in understanding if emotion related brain networks are more involved in processing information from human agents than from AI assistants. Such information can be of importance for future designs of AI assistants, that are treated more similarly to how humans think and the human brain works. As highlighted by (Mitchell, 2019), AI has a new spring with an impressing progress in the field of machine learning, but also with a number of caveats to consider.

The increased usage of AI in the workplace needs to be accompanied by an increased education of users in the technical and ethical aspects of AI methods. The current research program will develop and implement a new course in AI for psychologists and social workers. A novel aspect of the proposed research project is the use of functional neuroimaging to determine how the course can change neural responses to AI based decisions. This

would inform on how the brain responds to AI assistants and whether brain responses change after taking the presented course, with a validation of the results in comparison with findings from the interviews. To develop and give the described course would be a contribution itself, but the unique contribution in this paper is the course design and the evaluation design together. Authors hope to present an evaluation of the first course version at ECIAIR 2023. The plan for future research is a thorough and iterative evaluation of a series of future course versions, if the presented project idea will get funded.

References

Ågren, P.O., 2019. Den tredje AI-vågen: Essäer om AI, samhället och individen (The third wave of AI: Essays on AI, society and the individual). BoD-Books on Demand.

Åhs, F., Mozelius, P. and Dobslaw, F., 2020. Artificial Intelligence Supported Cognitive Behavioral Therapy for Treatment of Speech Anxiety in Virtual Reality Environments. In European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2020)

Blaschke, L.M., 2021. The dynamic mix of heutagogy and technology: Preparing learners for lifelong learning. British Journal of Educational Technology, 52(4), pp.1629-1645.

Bostrom, N. 2017. Superintelligence: Paths, Dangers, Strategies, Oxford University Press

Carter, R.M., Bowling, D.L., Reeck, C. and Huettel, S.A., 2012. A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science, 337(6090), pp.109-111.

Edbacken, J. and Vernmark, K. 2021. Digital psykologi (Digital psychology), Studentlitteratur ISBN:9789144141565

Fiske, A., Henningsen, P. and Buyx, A., 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of medical Internet research, 21(5), p.e13216.

Goldkind, L., 2021. Social Work and Artificial Intelligence: Into the Matrix. Social Work, 66(4), pp.372-374.

Hodgson, D., Goldingay, S., Boddy, J., Nipperess, S. and Watts, L., 2021. Problematising Artificial Intelligence in Social Work Education: Challenges, Issues and Possibilities. The British Journal of Social Work.

Humble, N. and Mozelius, P., 2019. Teacher-supported AI or AI-supported teachers. In European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2019) (pp. 157-164).

Kahneman, D., Sibony, O. and Sunstein, C.R., 2021. Noise: A flaw in human judgment. Little, Brown.

Meilvang, M.L. and Dahler, A.M., 2022. Decision support and algorithmic support: the construction of algorithms and professional discretion in social work. European Journal of Social Work, pp.1-13.

Mitchell, M. (2019). Artificial intelligence: A guide for thinking humans. Penguin UK.

Nordesjö, K., Scaramuzzino, G. and Ulmestig, R., 2022. The social worker-client relationship in the digital era: a configurative literature review. European Journal of Social Work, 25(2), pp.303-315.

Patterson, D. A., & Cloud, R. N. 1999. The application of artificial neural networks for outcome prediction in a cohort of severely mentally ill outpatients. Journal of Technology for Human Services, 16(2–3), 47–61.

Ranerup, A. and Henriksen, H.Z., 2019. Value positions viewed through the lens of automated decision-making: The case of social services. Government Information Quarterly, 36(4), p.101377.

Russell, S., 2019. Human compatible: Artificial intelligence and the problem of control. Penguin.

Socialstyrelsen / The Swedish Board of Health and Welfare 2019. E-hälsa och välfärdsteknik i kommunerna 2019 (E-health and welfare technique in municipalities). Stockholm: Socialstyrelsen

Svensson, L. 2019. Automatisering–till nytta eller fördärv? (Automation – of use or of destruction?). Socialvetenskaplig tidskrift, 26(3-4), 341-362.

Taylor, J.E.T. and Taylor, G.W., 2021. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. Psychonomic Bulletin & Review, 28(2), pp.454-475.

Tegmark, M., 2017. Life 3.0: Being human in the age of artificial intelligence. Vintage.

Wihlborg, E., Larsson, H., & Hedström, K. 2016. "The Computer Says No!"--A Case Study on Automated Decision-Making in Public Authorities. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 2903-2912). IEEE.

Wärnestål, P. 2021. Design av AI-drivna tjänster (Design of AI-driven services). Studentlitteratur. ISBN: 9789144139746 52(4)

Zhang, C. and Lu, Y., 2021. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, p.100224.