ICIS 2022 Proceedings

Digital Learning and IS Curricula

Dec 12th, 12:00 AM

Educating about Responsible AI in IS: Designing a course based on Experiential Learning

Charlotte Husom Grøder
Norwegian University of Science and Technology, charlotte.h.groder@ntnu.no

Stefan Schmager
University of Agder, stefan.schmager@uia.no

Elena Parmiggiani
Norwegian University of Science and Technology, parmiggi@ntnu.no

Polyxeni Vasilakopoulou
University of Agder, polyxeni.vasilakopoulou@uia.no

Ilias Pappas
NTNU, ilpappas@ntnu.no

See next page for additional authors

Presenter Information
Charlotte Husom Grøder, Stefan Schmager, Elena Parmiggiani, Polyxeni Vasilakopoulou, Ilias Pappas, and Sofia Papavlasopoulou

Educating about Responsible AI in IS: Designing a Course Based on Experiential Learning

Completed Research Paper

Charlotte Husom Grøder
Norwegian University of Science and Technology
Department of Computer Science
7491 Trondheim, Norway
charlotte.h.groder@ntnu.no

Stefan Schmager
University of Agder
Department of Information Systems
4604 Kristiansand, Norway
stefan.schmager@uia.no

Elena Parmiggiani
Norwegian University of Science and Technology
Department of Computer Science
7491 Trondheim, Norway
parmiggi@ntnu.no

Polyxeni Vassilakopoulou
University of Agder
Department of Information Systems
4604 Kristiansand, Norway
polyxeni.vasilakopoulou@uia.no

Ilias Pappas
University of Agder, Kristiansand & Norwegian University of Science and Technology
Department of Computer Science
7491 Trondheim, Norway
ilpappas@ntnu.no

Sofia Papavlasopoulou
Norwegian University of Science and Technology
Department of Computer Science
7491 Trondheim, Norway
spapav@ntnu.no

Abstract

Responsible AI (RAI) is an emerging topic in the Information Systems (IS) literature. RAI entails ensuring ethical, transparent, and accountable use of AI technologies in line with societal values, expectations, and norms. The challenge for research on IS education at university level is to accompany the growing research on RAI with approaches to educate students about this emerging theme. Research on IS education on responsible AI remains scarce to date, however. We ask: How can we design a course to educate students about responsible AI? We build on earlier research and an experiential learning-based approach to propose a course design promoting students' multidisciplinary, problem-based learning about RAI applied to the case of public welfare services. Our study is based on participatory observations of student groups and group interviews after a project, acting as an arena where the students could reflect on the learning process and evolving awareness of RAI.

Keywords: Experiential learning, exploratory study, digital learning, responsible AI, attitudes toward AI

Introduction

Artificial Intelligence (AI) has become part of everyday life as it is implemented in several services in both public and private sectors, such as healthcare, education, and employment (Dignum, 2019; Orr & Davis, 2020). Based on incoming data, AI-infused systems can solve complex problems with increased efficiency and quality while learning from existing datasets (van den Broek et al., 2021). Without human intervention and being explicitly instructed, AI systems can easily identify patterns and make predictions and classifications. However, one of the challenges is that they tend to lack transparency, i.e., it is difficult for humans to understand AI models' inner processes to generate outputs after they have been trained. This challenge is addressed as the black-box problem (Norwegian Data Protection Authority, 2018; Asatiani et al., 2021).

The non-transparent nature of black-box systems has been found to negatively affect the trust of end users (Burrell, 2016; Gilkson and Woolley, 2020), particularly citizens in the public sector, where a growing number of AI systems are tested and introduced to increase the efficiency of public services with unanswered questions related to reporting the models' inner workings to citizens (Vassilakopoulou, 2020). To address this problem, AI systems need to be handled as tools for "enhancing human agency, without removing human responsibility" (Floridi et al., 2018, p. 692). Accordingly, a growing body of research calls for a responsible design and development of AI - coined as Responsible AI (RAI). RAI entails ensuring ethical, transparent, and accountable use of AI technologies in line with societal values, expectations, and norms (Mikalef et al., 2022). Emerging research in IS and neighboring fields explores ways to improve trust among users by explaining how AI models work and thus promoting a responsible approach to AI (Asatiani et al., 2021; Kane et al., 2021).

This situation presents an unprecedented opportunity for the IS field to contribute to educating future generations: "The most important role of the information systems community is to educate new generations of professionals whose work focuses on the use of information systems to transform the ways in which organizations and societies are structured and operate to achieve their goals." (Topi, 2019, p. 1) In light of the current state and future of IS education due to the growth of AI and automation "that often exceeds the human capability to adapt" (ibid.), there is a need to offer effective and complete educational experiences for students to help them understand the implications and consequences of their work. The challenge for research on IS education, however, is to accompany the growing research on RAI with approaches to educate students about this emerging theme. Research on IS higher education on responsible AI remains scarce to date, however. We, therefore, ask the following research question: *How can we design a course to educate students about responsible AI?*

We build on previous work on education in IS (e.g., Matthee and Turpin, 2019; Goh, Gangi, and Gunnells, 2020; Connolly, Rush, and Mutchler, 2020) and present an approach to educating students in RAI based on experiential learning (Kolb, 2014). We implemented a collaborative, multidisciplinary course on RAI drawing on an existing experiential course container at master level at a large Scandinavian University. The novelty presented in this work consists of integrating innovative methods and tools into an existing course infrastructure to enhance experiential learning. An important element is the multidisciplinary nature of the course, as students from very different study programs and backgrounds work together in groups to solve societal problems. In our study, students were asked to propose approaches on how to present and communicate the inner workings of AI models in welfare services to end-users (citizens) and reflect on emerging concerns and understanding of AI. With this study, we contribute to IS research on education by proposing a protocol based on the design of a course to educate about RAI. The protocol is evaluated based on students' reflections after the course.

The remainder of this paper is organized as follows. First, we present relevant literature related to educating students through experiential learning on emerging topics such as RAI. We then describe the background of the course we designed before we elaborate on our contribution to the course. Next, we describe our research methods and present our findings in detail. Our findings are discussed based on the background literature. We conclude by summarizing limitations and suggesting future research directions.

Theoretical background: Toward educating students in responsible AI in IS

Previous research in IS investigates different methods to teach and educate students on emergent topics (Matthee and Turpin, 2019; Connolly, Rush and Mutchler, 2020; Shahrasbi, Jin and Zheng, 2021; Shankaranarayanan et al., 2021). Given the societal relevance of RAI, we lean on experiential learning, in which "[l]earning is the process whereby knowledge is created through the transformation of experience" (Kolb, 2014, p. 38). Different conceptualizations of the learning model in experiential learning exist, but its tenet is a problem-centric approach based on a continuous feedback loop between concrete experiences, reflection, generalization, and testing (ibid).

Experiential learning in practice thus relies on collaboration and active learning (Shahrasbi, Jin and Zheng, 2021; Beard, 2010; Hall 2018) by engaging students in different activities such as discussions, problem-solving, role-play (e.g., Connolly, Rush, and Mutchler, 2020) or group projects (e.g., Goh, Gangi, and Gunnells, 2020; Hod et al., 2022). For example, Shahrasbi, Jin, and Zheng (2021) challenge the traditional way of teaching software programming and app development courses that involve intensive coding exercises and thus little interaction among students. They propose a teaching protocol in a mobile app development course that focuses on a design thinking model, taking advantage of the models' user involvement. Matthee and Turpin (2019) identified a lack of work in IS on how to teach foundational skills such as critical thinking and problem-solving to students and thus focused on developing such skills in a course taught to first-year IS students. Along the same lines, Connolly, Rush, and Mutchler (2020) implemented role-play in a project management class to teach students soft skills such as empathy, emotional intelligence, and communication due to the lack- and slow adaption of education on such skills. Goh, Gangi, and Gunnells (2020) employed team-based learning in an online introductory IS course to consolidate individual learning with collaboration. Hall (2018) applies experiential learning in a teaching case on service-learning and finds that students feel they have more ownership of their projects, which will be an asset to future employers.

Neighboring fields outside the IS literature provide examples of research on (responsible) AI and education. For example, to understand and reevaluate current AI teaching and practices, Wollowski et al. (2016) present the results from two studies designed and conducted on AI practitioners and AI educators, aiming to investigate the different topics and techniques used in practice, and current teaching practices and possible changes, respectively. They conclude with it seeming to be a "healthy match between instructed topics and the practitioners' needs" (p.6). Moreover, their studies showed for example that an essential skill among practitioners that seemed to be lacking in education was the ability to take a perspective on AI tools and techniques. Dignum (2021) investigates how we can bring AI and education together by discussing and relating AI's vision for responsibility and trustworthiness to education. She explains that traditional education systems need to be reformed due to AI's social transformation and presents different aspects to include in education curricula to contribute to the reform, for example, distributed collaboration to be able to work efficiently in person or online across distance and time, and to include humanity subjects into technology curricula to approach problems in a creative way using, for example, soft skills such as empathy. Moreover, Hod et al. (2022) developed a course design structure on responsible AI specifically for students and professionals coming from the law and data science fields. They deliberately focused on this audience as these professions are likely to meet and work together in "real world" projects: "It is our belief that to increase the likelihood that AI systems are "responsible", an effective multidisciplinary dialogue between data scientists and lawyers is needed." (p. 35) They designed the curricula according to the signature pedagogies of each of the disciplines, namely the common practices and styles of instruction of the specific profession. Thus, the tasks were designed in a way that they could only be addressed by a joint effort, fostering collaboration and multidisciplinary dialogues. Their approach describes a standalone course design, yet the authors also suggest integrating the course into core computing courses.

In sum, these studies underline the importance of developing novel approaches to educate about AI based on interdisciplinary collaboration and grounding in real-world problems. From an educational perspective, however, RAI poses the additional challenge of helping students learn about complex themes such as ethics, transparency, and accountability. The students need to be able to both understand these underlying concepts and become and remain aware of the human-centered nature of AI i.e., ensure that AI tools do not disregard users' humanity (Kane et al., 2021). Experiential learning is a very promising approach in this direction because it promotes the empirical investigation and inclusion of different people's perspectives in concrete contexts. The importance of including different perspectives in the learning process about RAI relates to the particular nature of AI. AI model's inner workings can be difficult to understand from a technical point of view (Berente et al., 2021) and are never fully knowable (Orr and Davis, 2020). AI models are often black-boxed (Norwegian Data Protection Authority, 2018) because they are non-transparent and put in place opaque processes for reaching results, limiting humans in their ability to shape, operate, and monitor them. This affects and goes beyond human control over AI model systems, making it difficult to trust them, calling for "increasing levels of accountability, that is, for organizations to be responsible for the consequences of AI in many circumstances" (Berente et al., 2021, p. 1434). RAI corresponds to principles to design and adopt AI-based systems in a way that is transparent, fair, and equitable (Bovens, 2007), ensuring that the social and technical constructs of the systems are in place to provide responsibility and trust in AI-based systems: "Besides choosing the proper algorithms, you also need to consider the ingredients (e.g., the data) to use and the composition of the team using it" (Dignum, 2019, p.4).

In sum, RAI is widely recognized as an important guidance for the development of future AI-based systems (European Union, 2021), and the sociotechnical soul of the IS field has the pedigree to elicit the required design knowledge on human-AI interaction and responsible design and use of AI-based solutions. However, research on operationalizing the general principles of RAI into actionable insights and models is lacking so far (cf. Mikalef et al., 2022), also in IS education. We believe experiential learning can be an important means to address RAI's human-centered, inclusive, and complex nature. We hypothesize that an experiential learning approach grounded in student collaboration addressing concrete problems can be fruitful in allowing students to learn to reflect critically on themes with complex societal ramifications such as RAI (cf. Vassilakopoulou, 2020). In what follows, we expand on existing trends in IS education about emerging topics to propose an experiential learning-based approach to develop a course on RAI.

Context

This study was conducted in a master's degree course container at a Scandinavian University called Experts in Teamwork (Sortland, 2001). Experts in Teamwork is mandatory for all students in master programs and programs of professional study with the purpose of educating students in multidisciplinary teamwork skills, as described by Sortland (2001, p.8): "By working in multidisciplinary teams where each member initially may have unlike perspectives on the problem at hand as well as be used to different problem-solving methods, the students will develop attitudes and skills related to teamwork". The main purpose of Experts in Teamwork is thus to guide the students in teamwork and help them acquire the necessary skills to work in diverse teams on real-world projects relevant to society (Wallin et al., 2017). The focus is to practice collaboration and communication skills by following experience-based learning and reflection (Sortland, 2001). Therefore, we chose this course in line with our theoretical interest in experiential learning.

In the spring semester of 2022, the students enrolled in the course could choose from approximately 105 classes that are referred to as 'villages', each village focused on a unique topic of societal relevance such as automation and society, responsible AI in welfare, space technology, and environmental sustainability. The villages were either run intensively (three weeks of project work) or longitudinal (one day a week of project work during the semester). A village always consists of a village leader, two learning assistants, and around 25 to 30 enrolled students, divided into smaller multidisciplinary teams of about four to six students. The purpose of the team's work is that students work together across disciplines (Sortland, 2001), bringing their competencies and challenging themselves to gain new ones. The unique topic for each village is often given by an external partner (either within or outside the university) who follows the team's work together with the village leader who is the main responsible. For this paper, we refer to the external partner as the challenge owner. To initiate the students' work, the challenge owner presents the topic of the village that involves challenges and tasks for the student's project work. In addition, throughout the course, the challenge owner is in contact with the teams during their projects' development. At the end of the project, the teams present their project work to the challenge owner. The students are evaluated based on the delivery of two final reports: One process report (that describes the reflections on the teamwork situations) and one project report (that describes the project development and solution given by the students) which together aim to show the interdisciplinary teamwork skills that the students developed through the course.

The village on Responsible AI in welfare

Based on previous research presenting different perspectives on education in IS (Matthee and Turpin, 2019; Goh, Gangi, and Gunnells, 2020; Shahrasbi, Jin and Zheng, 2021; Connolly, Rush, and Mutchler, 2020), we designed a village in the Experts in Teamwork course container with the name "Responsible AI in Welfare" to teach students about AI and responsibility in the public sector as an emerging IS topic in Spring 2022. The village was run online due to restrictions because of COVID-19, enabling students to participate both in and outside of Scandinavia. The platform used for the village was Zoom which enabled teamwork by organizing the students into so-called breakout rooms where each team was assigned an online room to discuss the project and collaborate. The focus of the village was on exploring ethics, transparency, and accountability aspects in the adoption of AI in welfare. Specifically, the students had to explore the type, content, and representation of information needed to communicate the explainability of the model used by AI applications for end-users.

Our contribution of a novel process design to educate students on responsible AI needed to be aligned with two existing conceptual configurations - the structure of Experts in Teamwork and the Experiential Learning Cycle (Kolb, 2014). The Experts in Teamwork course provided overarching scaffolding, describing roles, group compositions, and challenge context. On the other hand, we used Kolb's Experiential Learning Cycle as the educational foundation of the process design, to ensure each learning phase was reasonably integrated into the Experts in Teamwork structure. We provided the students with appropriate frameworks, methods, and tools to facilitate each learning phase, including the design thinking methodology, personas, and prototyping techniques. Further, we also offered dedicated forums for reflection at the end of the course to enable critical contemplation about the topic, the learnings, and conclusions.

The main stakeholders for the village were the challenge owner contributing to the project, the village leader, two learning assistants, and the students. The structure of Experts in Teamwork and how the authors contributed to the village are visualized in Figure 1.

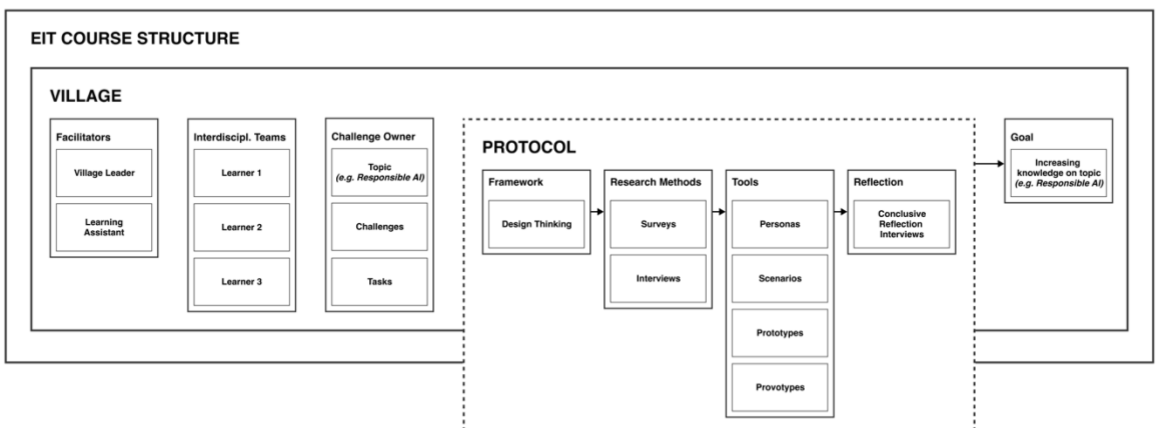

Figure 1. Authors' contribution integrated into the village and larger Experts in Teamwork course structure

The village consisted of 28 students – 10 female and 18 male – that came from 17 different study programs such as financial economics, digitalization, administration and work, globalization and sustainable development, neuroscience, psychology, informatics, and medicine. Most students did not have heavy technical backgrounds. Based on their study programs and gender, they were divided into five teams with four to six students per team to ensure that the teams were multi- and interdisciplinary. In addition, the teams had members from Europe, North America, Africa, and Asia, ensuring diversity. Being a multidisciplinary field including for example both a technical and an experiential learning perspective, the IS field is well-positioned for educating about RAI, and Experts in Teamwork is a suitable course for this purpose as the course engages students from different study programs and nationalities to participate.

At the beginning of the village, the teams were given two guidelines to facilitate the process: First, they had to choose one out of three following fictional cases on AI models used in a public service context: (1)

suggesting what kind of support is better for an unemployed person, (2) suggesting high-risk case for student loan holders and (3) suggesting the expected total duration of sick leave. In all three cases, users overlapped with citizens accessing public services. Second, the teams had to develop an understanding of the relevant user groups in the AI case they selected by choosing a research method for collecting data. Based on their initial research, the students developed personas and scenarios and then designed how to present the AI model to them. We required the students to decide between one of two pathways, which also would determine the type of artifact the teams would have to deliver. Based on their assessment of the use case and problem definition, the teams needed to decide if they would follow the prototyping- or the provotyping-pathway. We added this step to the protocol to motivate the students to reflect on their findings early in the process and evaluate which path might be more appropriate for their goal and initial findings. Prototypes would be the tool of choice if the students were already confident in their problem definition based on their initial analysis and data collection and wanted to engage in "problem-solving". In contrast, a provotype would be chosen if the students decided to explore the problem space more and engage in conversations with users about their perspectives and attitudes towards responsible artificial intelligence. These conversations could then help the teams to define their hypothesis based on a deeper understanding of their users - which also affects reflection processes within the students.

A prototype can be described as an early-stage model, illustration, or manifestation of a concept, a product, or a service that allows users to interact with it (Preece et al., 2015). It is used to validate or falsify assumptions, test ideas and collect feedback. Prototypes can have various levels of fidelity, ranging from low-fidelity paper prototypes to high-fidelity prototypes with an extensive degree of interaction and a form factor close to a finished product. Prototypes can also take various shapes, from pen & paper versions to impersonation prototypes (e.g., Wizard of Oz) to websites or videos. A deep and comprehensive definition of prototypes' different types and variations would go beyond this work, but a critical analysis of prototypes can be found in Houde & Hills work "What do prototypes prototype?" (1997). For this paper, we will specifically focus on the educational aspects of prototyping, and the difference between prototypes and provotypes.

Prototypes possess an inherent quality of reflection that can be leveraged as a vehicle for education. In the "Field Guide to Human-Centered Design" prototyping is characterized as learning through making (IDEO, 2016, p.119). Schön (1983) describes the activity of building prototypes as an encouragement for reflection in design and it is recognized by designers from many disciplines as an important aspect of design (Preece et al., 2015). Dörner (2011) states that reflection improves the outcome of thinking and decision-making processes, which is in line with the findings by Mann et al. (2009), who found that reflection positively influences process and outcome across different disciplines. Jobst et al. (2020) even created a reflection tool based on this mediative virtue of prototyping. In our work of educating students about the concept of responsible AI, we utilized these characteristics of creating and testing tangible artifacts to enable learning processes through engagement and reflection.

Besides the described form of prototyping to create and test a specific solution to an identified problem, we also provided the students with a second variant of prototypes, so-called "provotypes". A provotype describes a provocative prototype that is used to cause a reaction — to provoke and engage people by imagining possible futures. On the one hand, provotyping resembles prototyping with respect to the need for concrete experience by working with concrete 'types' (Mogensen, 1994). On the other hand, however, the driving intention of provotypes is - as the name suggests - to provoke. Provotypes are used as a tool to expose and embody tensions about an area of interest to support collaborative analysis and collaborative design explorations across stakeholders (Boer & Donovan, 2012). By provoking those who engage with it, provotypes embody a learning tool to open a problem space for fruitful discussion and uncover deeper insights and viewpoints.

Finally, the three-week intensive course outcomes resulted in a presentation to the customer of the teams´ provo- or prototypes, and their evaluation results.

Research methods

Techniques based on close and extended interaction between the researcher and the setting, such as action research, are particularly well suited to promote experiential learning (Kolb 2014). To support and enable the students' educational learning while performing their research on the practical issue regarding user trust, our study followed an *action research* strategy, an accepted method within IS research in education that is concerned with solving practical problems in collaboration with the research subjects while creating change and consists of an iterative cycle of planning, acting, and reflecting (Baskerville and Myers, 2004). Furthermore, it is well suited for research projects involving several stakeholders with heterogeneous agendas as it is flexible and adaptable (ibid).

The village leader acted as a facilitator in the overall process for the course, while the challenge owner acted as an advisor providing the students' feedback and directions on the different tasks. The study started with the challenge owner presenting three topics to the student groups, where each group selected the topic they preferred. Following the action research strategy, they were encouraged to make a *plan* for how to conduct their research, and how to interact with the users. Subsequent *actions* were drawn from the findings when the students interacted with the users, leading to the development of changes in the form of proto- or provotypes, emphasizing change (ibid). After observing the user interactions with the proto- and provotypes, the students *reflected* on the process. Their reflections would then be used to revise the plan, which would be the starting point of a new cycle. Because of the time perspective for this project, the students only did the action research cycle once. The developed protocol is visualized in Figure 2.

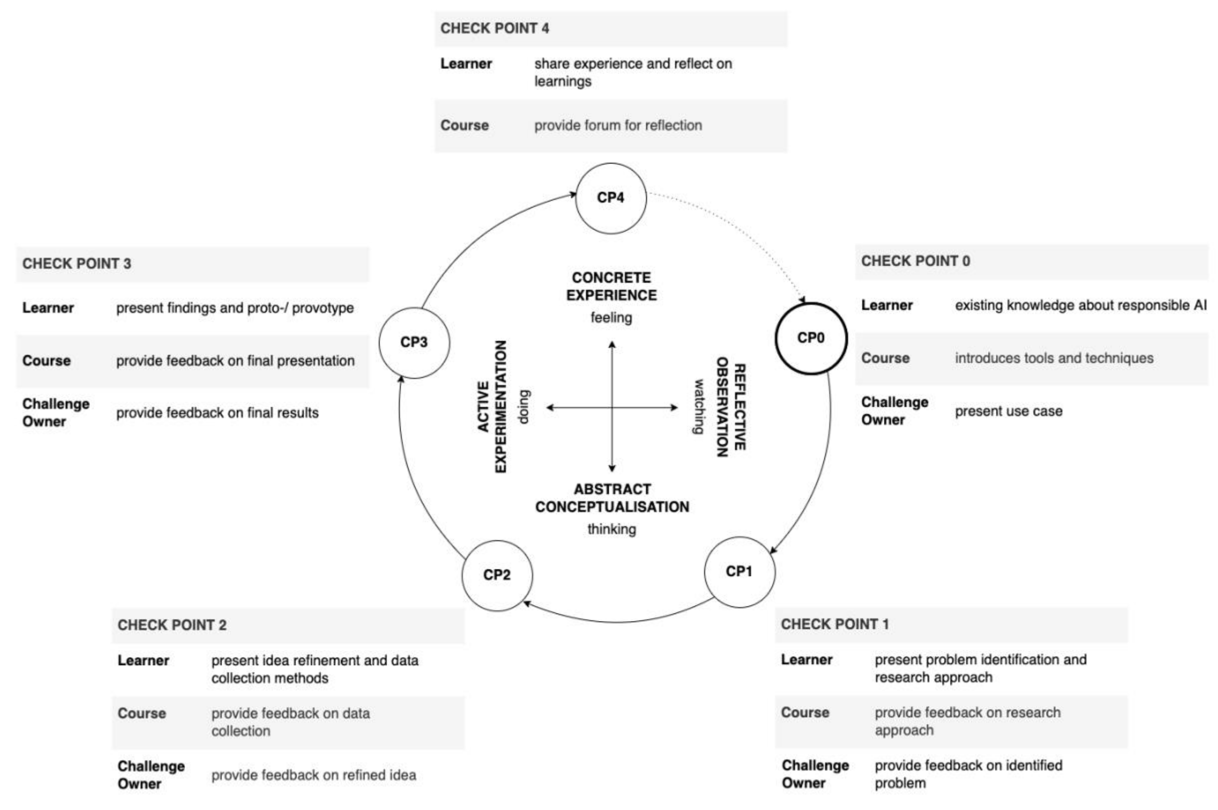

Figure 2. The protocol developed to support the students learning process about responsible AI. The protocol describes the course process including checkpoints and their corresponding experiential learning phases

Data collection

Our data collection was designed in two stages. The first stage lasted about four hours in total, where we performed participatory observations of the teamwork. We visited each team in their breakout rooms in Zoom for about five minutes, approximately two to three times per week. There were two reasons for these participatory observations, first, to be available for the students if they had any questions regarding the project; second, to observe the process where the students worked in teams, and to discover potential interesting discussions and ideas. The experiences from the participatory observations were used to form the interview guide in stage two of the data collection.

The second stage of our data collection was conducted after the students' final presentations but before they had to submit their final reports. Each team was offered the opportunity for a group interview to reflect on their experiences and learnings from the course and the insights they had collected. We deliberately chose this format since group interviews can result in more natural conversations with participants reminding and challenging each other on individual perceptions or details (Martin & Hanington, 2012, p.102). Furthermore, as the course structure was focused on collaborative group work, we also wanted to retain the same setting and dynamics during the group interviews. A total of five group interviews were conducted, with four to five students participating in each interview. Each group interview lasted for approximately one hour. These interviews served a dual purpose. First, the group interviews gave the students a forum for critical reflection on their research approach, process, and learning. Further, qualitative reflection from students who had spent a significant amount of time and analysis on the topic of responsible AI in the public service could potentially provide a wealth of input and perspectives and benefit the overall research in the field. Especially interviews are a suitable research method for collecting firsthand personal accounts of opinions, attitudes, and perceptions (Martin & Hanington, 2012, p. 102).

The interviews were run by the first and the second author acting as facilitator and moderator, respectively. In preparation for the interviews, a semi-structured interview protocol containing discussion prompts and guiding topics was prepared. The open-ended questions were enquiring about how the groups decided on their research process, their most surprising insights, and reflections on how their own understanding of responsible and trustworthy AI was affected by conducting the research. The moderator drove the conversation by asking the prepared discussion prompts and follow-up and elaboration questions. The facilitator took detailed notes of the conversation and oversaw technical peculiarities, e.g., starting and stopping a potential recording. Before the reflective part of the group interviews started, the students were provided with a brief introduction to the purpose as well as the format of the interviews, and any clarification questions could be addressed. Further, the students were asked if they would agree to the recording of the interviews, which would help the interviewers compare notes and avoid missing important parts.

The interviews were set up in a semi-structured format, providing loose directions for reflection but allowing for flexible detours in a conversational format. An open-ended question format encourages more detailed responses from the interviewees, which helps to elicit more detail about a topic (Cooper et al., 2014, p. 53). It was important to avoid pre-supposing any answers with the way the questions were phrased. Questions like *"Why did they choose this specific village /challenge?* or *"What are your thoughts about responsible AI after this project?* were asked to elicit a pre-post comparison in understanding the concept of responsible AI. Right after the interviews, the moderator and the facilitator compared their notes and highlighted any fascinating observations or clarified misunderstandings. All interviews were transcribed and summarized into themes and key findings, as well as specific relevant statements, were highlighted.

Data analysis

Eisenhardt (1989) argues the importance of defining the research question broadly to ease the data analysis. Consequently, the data were analyzed with the research question in mind following a stepwise deductive inductive (SDI) approach for qualitative research (Tjora, 2019), using the computer-assisted qualitative data analysis software NVIVO. The unit of analysis was the practice of learning about responsible AI through experience.

The inductive processing of the qualitative data started from the raw data collected from interviews and resulted in a manageable, quality-assured analysis of the data used to develop empirical arguments. The process was iterative (Eisenhardt, 1989; Tjora, 2019), with iterations back and forth. Codes were identified by analyzing the raw empirical data from interviews in a two-step iterative process of coding that preserved the details in the material. For example, when observing an interesting excerpt from the raw data such as "we quickly came to an agreement to mix approaches (...)" this could result in a code such as "we quickly came to an agreement to mix approaches" to preserve the details in the empirical material. The codes were then used to develop conceptual categories; for example, since the code "we quickly came to an agreement to mix approaches" is a reasoning about research approaches, it could belong to the conceptual category "considering research approaches". Patterns emerged from labeling codes into conceptual categories, which were used to determine overall themes. Conceptual categories such as "considering research approaches", "performing data collections", and "deciding on creating a provo- or prototype" could all belong under the umbrella of the overall theme "eliciting concrete processes" as they relate to students' thoughts on different aspects of the course. Furthermore, students' thoughts on aspects of the course can be helpful for the development of future courses on AI, relating to the research question on how we can educate about AI. The themes, conceptual categories, and excerpts found in this project are presented in Table 1.

Theme	Conceptual Category	Excerpt
Eliciting concrete processes	Considering research approaches	"I think we quickly came to an agreement to mix approaches because we felt that a survey was nice to get a lot of information from different people, and then the interviews gave us an opportunity to get in dept from users." P2, *Group 2*
	Performing data collections	"We wanted to formulate the questions so that they [the users] were not affected or leading the questions in a specific direction." *P2, Group 3*
	Deciding on creating a provo- or prototype	"We assumed keeping the attention of participants for a longer time and explaining the AI to them would be more difficult to do with a prototype." *P1, Group 4*
Learning about the use of RAI	Different impressions of the AI model	"I was surprised about the model itself and the estimate it used. I find it hard to believe that they are the only estimates they do. (...)" P4, *Group 2*
	Project contributing awareness	"During this course, I saw cases that the government can use it for. I did not think much about AI in a governmental context before." P2, *Group 3*
Revisiting the learning process	Reflections on the village	"I thought it was exciting to learn more about how AI should be used in welfare and also the general public views on AI to get a further insight on that." *P1, Group 2*
	Changing perceptions of AI at the end of the learning process	"In general, I like to learn more about systems, but I am also skeptical. I have learned more now, but I am also more skeptical now than I was at the start." P5, *Group 5*
	Experiencing time pressure	"[We] would have liked to have one additional week, which would have allowed us to get more responses from the prototype, be able to make it better" P2, *Group 1*

Table 1. Our analytical framework resulting from our data analysis

Research findings

This section presents the findings from our interviews with the student groups after the end of the project. Based on the analytical framework in Table 1, we divide our findings into the following three categories (1) eliciting concrete processes, (2) learning outcomes related to RAI, and (3) revisiting the overall process.

Eliciting concrete processes

To identify the students' experiences on the course, we asked if they could reflect on various aspects of their work and their choices during the project. As part of this, they were first asked to reflect on their chosen *research approaches* and methodologies.

All the groups reported that they used qualitative research approaches to understand the topic, specifically surveys and interviews as the primary data generation methods to reach out to many people and thus get information fast, and identify concerns in-depth, respectively. A student elaborated:

"*I think we quickly came to an agreement to mix approaches because we felt that a survey was nice to get a lot of information from different people, and then the interviews gave us an opportunity to get in-depth from users."* (P2, Group 2).

The students were also asked if they could elaborate on how they decided which questions to include in the questionnaires and interview guides as part of how they *performed the data collection*. Two groups explained that they had brainstorming sessions with individual thinking before sharing their questions and discussed which questions to include within the group. One group started from the assumption that people, in general, were skeptical towards AI, while another started asking general questions about AI, then focused on questions related to the specific case: *"We asked general questions in the beginning, but we quickly realized that we must have a specific focus of what we want to gain from this information"* (P4, Group 2). Moreover, all the groups avoided asking personal questions, arguing that it was irrelevant to the specific case.

To explore how the groups *decided on creating a proto- or provotype* for model explainability to users, the students were asked if they could elaborate on their choices for developing one over the other. Three groups chose to make a provotype, and two groups made a prototype. One of the groups that created a provotype explained that they were initially unsure what to make. However, they identified a provotype as the easiest to make and agreed on making one after a voting session in the group. Similarly, the second group creating a provotype explained that *"we thought it would be cool to make one without actually making a real prototype. (...) we discussed that for quite some time before concluding with the provotype"* (P3, Group 2). The group argued that a prototype is more challenging to make as a prototype needs to be a more finished model than a provotype. Figure 3a shows an example from one of the groups' provotypes.

Another group explained that they thought about the pros and cons of both, but they explained that they did not have any provocative ideas. Due to the short time frame, creating a prototype seemed more feasible for them. Figure 3b shows an example from one of the groups' prototypes. Similarly, another group also expressed how the time constraints made them decide to create a prototype while it seemed more straightforward than creating a provotype. A student explained: "*we did not know how to provoke someone"* (P1, Group 1). Also, while they wanted to build trust, they explained that a prototype would be more helpful than a provotype as there is already much bad sentiment about AI, and they did not want to provoke further – but rather make something positive.

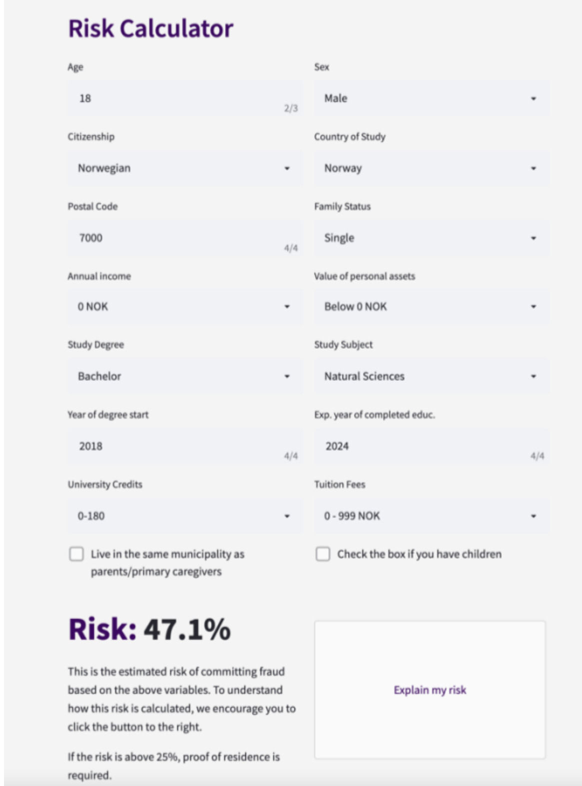

Figure 3a. Example of a high-fidelity interactive HTML provotype.

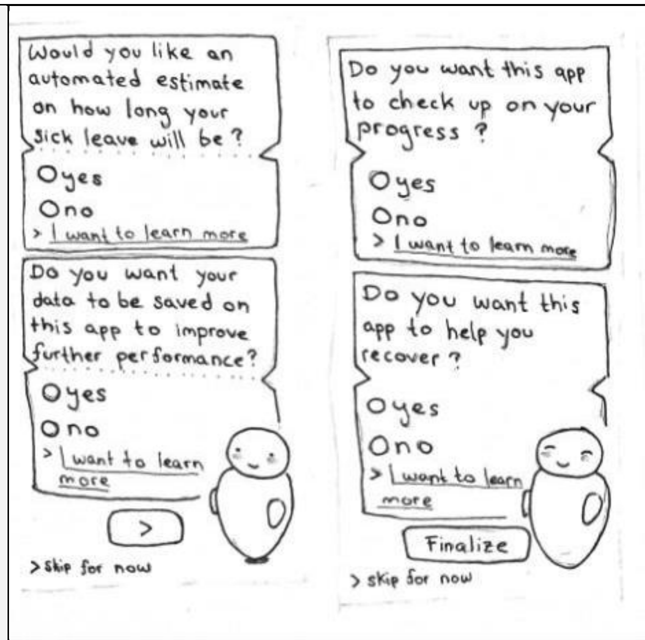

Figure 3b. Example of a low-fidelity paper prototype

Learning about the use of RAI

Based on the students' data collections, where they interacted with users through online surveys and interviews to identify their perceptions of AI in public services, we wanted to investigate different perceptions of what it was like learning about the use of RAI. First, we wanted to explore the *different impressions of the AI model* that was provided to the students.

A recurring statement that several student groups mentioned based on their results from their data collections was trust in Scandinavian governments. One student explained: *"[Scandinavian] government[s] are doing a great job, and we believe that they will do it well"* (P2, Group 3). Despite Scandinavian trust in the government, some students were surprised that some informants did not find the use of AI in Scandinavian government services concerning, despite media, science fiction, and dystopian stories presenting AI implementation in negative ways: *"It is often presented as something provoking and something humans should be worried about"* (P5, Group 2). Students from one group found it surprising that people generally were more optimistic about AI implementation as they expected people to be more skeptical and afraid. However, students in another group explained that the majority of their informants were skeptical towards AI in general and that they further found it surprising that some people were more skeptical of the AI developers' *intention of the application* rather than the AI model itself. A student elaborated:

"I was surprised about the model itself and the estimate it used. I find it hard to believe that they are the only estimates they do. I would think that the AI models will use more provoking variables. So, I was shocked that the very stereotypical elements ... they were not big of an element as I would have assumed." (P1, Group 2)

One group elaborated on an interesting finding that did not align with what the group thought at the beginning: Trust in the public sector and trust in AI are inversional proportional. Therefore, the ones that trusted the public sector the least could very well imagine a new AI-based solution to be used in the public sector.

Based on the students' reflections on their findings from their interactions with users, we wanted to investigate to what extent working with and learning more about the use of AI *contributed to an awareness of AI*. *S*pecifically, we wanted to investigate if working with the project and the user interactions changed or affected the students' attitudes towards using AI for public welfare services and, if so, how their attitudes and perceptions were influenced.

Some students did not experience any specific change in awareness about AI, a student elaborated: *"It is more or less the same, [I] learned more about the dangers, but also the possibilities"* (P1, Group 5), while another student explained: *"The use case of the project was limited in scope, [it] cannot go "wild" with the cases in this project. But [I] definitely worry about broader AIs or other wider cases, that can have an impact on [our] lives."* (P3, Group 3).

A few students explained that they became more skeptical about using AI after learning more about it. For example, one student explained the use of AI as scary if it is left alone. Controversially, another student elaborated on how one should trust not the AI itself but rather the organization behind the AI. For example, one group found that trust in the public welfare sector and trust in AI is inversional proportional – meaning that the less trust people had in the public authority, the more trust they had in AI, and conversely. The reason for this seemingly contradicting statement was the assumption that an AI system would be more impartial than a human employee.

Moreover, there were also optimistic students that were positive about the use of AI for public services. One student described how they got a stronger feeling throughout the course when learning about the different use cases of AI about how the "good" outweighs the "bad". Similarly, another student explained that AI was less intrusive than first thought: *"I was surprised of the model; I would have thought it was more intrusive. (...) I learned that most of the models are not necessarily like that"* (P4, Group 2). Several students mentioned how science fiction and media have contributed to prior skeptics towards AI by showing bad scenarios of AI. However, the students explained that they became more tolerant of the use of AI during the course as they learned more about AI and got a better understanding of how AI works. For example, a student discussed the advantage of thinking about AI as any tool used for decision making and furthermore that *"if it is to be used by humans, then it is useful to understand what's happening so it can actually aid your decision making rather than make the decisions for you"* (P5, Group 2).

Revisiting the learning process

To identify how the students perceived the course and the overall project, we were curious about their *reflections on the village*. Due to the nature of the course, which is based on the development of multidisciplinary teamwork skills, the students came from different study programs and thus entered the project with different expectations and experiences with technology generally and AI specifically. Two arguments emerged among the students as reasons why they chose the village: (1) general interest in AI and (2) a desire to learn more about AI. Some students explained that they found AI-related questions and topics interesting, such as discussions, problems, and general ethics around AI, and that they chose the village to learn more and hopefully get more insights into AI. AI's rapid growth and *golden age* were also mentioned as compelling for learning more about AI and the benefits of using AI in welfare – for example, one student explained that it would be interesting to understand how AI can benefit society. Along the same lines, two students from the medical- and electrical engineering fields discussed that they wanted to learn more about the benefits of AI as it will become a part of their everyday life.

Second, we wanted to investigate whether the students found the course helpful in learning about AI to find out if students experienced any *changed perceptions of AI after the project*. Two students who explained that they entered the course with technology interest and with a good level of previous knowledge and understanding of AI discussed respectively: *"I learned a bit more, [it was] interesting to investigate further about AIs possibilities and dangers"* (P5, Group 5), and *"the course increased the understanding of how it [AI] works"* (P3, Group 5). Another student in the same group was more skeptical and explained: *"I need to learn more about technologies, I need more knowledge before I trust it"* (P2, Group 5). Along the same lines, one student discussed:

"AI is so huge and there is so much that you cannot know. (...) Even if you learn a lot about AI, you still do not really know anything about AI, I think. I'm kind of a pessimist when it comes to technology. In a way, I want to know more about it because I'm critical towards it. But you changed my way and my perceptions about it." (P2, Group 2).

To identify potential learning outcomes that the students had from the project, they were asked if they could elaborate and reflect on the process. The students mentioned several interesting areas they could have improved- or would have done differently. For example, several students *experienced a time pressure,* elaborating that they experienced a lack of time and would like to have an additional week for the data collection to strengthen it by collecting several responses from the survey and performing more interviews: *"[We] would have liked to have more time finding and reading more source material. Because of the lack of time, and the fact that we had to be in Zoom the whole day, prevented any deeper research."* (P1, Group 1).

Additionally, a student explained that they should have asked more questions. One additional week would also have contributed to finding and reading more source material: One student elaborated that they would have looked for user behavior and patterns within their collected data. At the same time, another student explained that it would have been interesting to investigate how the provotype could have worked as a prototype. The students argued that since the course was online, they were prevented from doing deeper research while they had "*to go online the whole day"*. Consequently, some students informed us that the results and outcomes of the project could have been different if the course had been run in person and with face-to-face interviews.

Discussion

Rapid developments in technology have transformed previous manual tasks with digital technologies in the private and public sectors (Wessel et al., 2021). AI techniques have a strong potential to increase the efficiency of previously manual tasks, but they also tend to black-box their inner decision-making processes. As these techniques enter several arenas of our day-to-day life, questions arise as to how we can promote a critical learning process for students. In particular, the increasing adoption of AI for decision support in the public realm opens questions on how we should educate students at the university level to think critically about AI technologies and learn to understand AI through a responsible lens (Vassilakopoulou, 2020). The implications of this question are wide, as it involves educating tomorrow's professionals and citizens toward taking a human-, as opposed to technology-, centric perspective to development.

RAI is an emerging phenomenon for IS research that has unfortunately so far received little attention from an education perspective (Matthee and Turpin, 2019). New course designs on responsible AI are emerging outside of IS to tackle collaboration challenges between different disciplines (Hod et al., 2022). It is crucial that education keeps up with the social transformation brought by AI (Dignum, 2021). To address this gap, we respond to calls for further work on how to teach foundational skills (Matthee and Turpin, 2019) through a research-based approach (Topi, 2019). In doing so, we proposed a course protocol (Figure 2) building on and extending previous research in education stressing the importance of promoting an experiential learning process (Kolb, 2014) based on collaboration and active learning through student engagement to promote individual learning (e.g., Goh, Gangi and Gunnells, 2020; Shahrasbi, Jin and Zheng, 2021; Hod et al., 2022).

The proposed protocol contributes to an improved learning process about RAI by providing methods and tools to facilitate critical discourse (Figure 1). It is inspired by design thinking, a well-established methodology to empathize and understand needs and perceptions (Dunne & Martin, 2006). The protocol integrates into the overarching scaffolding of the Expert in Teamwork framework (Sortland, 2001), alongside existing entities like roles, challenges, and tasks. To enhance students' learning, the protocol demands data collection, e.g., through surveys and interviews to promote awareness and generate insights. It enables reflection and analysis by the students through working with the collected insights to create personas, scenarios, and proto- or provotypes. Furthermore, it offers reflective techniques, e.g., closing interviews to provide additional spaces for zooming out and evaluating one's own path and evolution during the project. Through in-depth engagement with the topic and discussions with potential users, experts, or other stakeholders, the protocol provides extensive learning capabilities. A potential risk is a dependency on the characteristics of study participants. It is therefore crucial to ensure a balanced participant pool to mitigate the risk of influencing the learning by propagating biases.

In the following, we elaborate on the implications of our work.

First, our findings illustrate that it is important to educate students about RAI as a way to overcome the black-box problem in the future by allowing them to learn to ask critical questions and make informed decisions about AI (see also Hod et al., 2022). IS literature on AI proposed to provide a safe space for organizations to test and reflect on the interaction between social and technical elements in AI models. We believe that such an approach could also be adapted to an educational context through the design of courses – like the one we proposed – that allow students to learn through experience and empirical data about the socio-technical aspects of AI with a responsible mindset. This has both a short-term and a long-term perspective. In the short term, educating students on RAI can facilitate the advancement of AI-enabled learning systems, which are gaining traction due to their ability to deliver learning content and adapt to students' needs (Kabudi et al., 2021). Students can become increasingly aware of the importance and potential of RAI, thus influencing their overall learning experience as well as the successful implementation of AI systems in education. In the longer term, students are tomorrow's users and decision-makers of AI in the public and private sectors (Topi, 2019). Students' reflections and learning will be important to support higher-level thinking and career readiness, skills that later can be important to communicate to the industry (Hall, 2018).

Second, our proposed course protocol supports the student's learning process through reflection by leveraging team-based collaboration and by interacting with citizens (cf. Goh, Gangi, and Gunnells, 2020). According to the final evaluation, the students described the multidisciplinary nature of the course as a positive trigger in learning about RAI, because it forced students from heterogeneous fields (technical, social, profession-oriented) to collaborate. Moreover, this arrangement enabled them to highlight and problematize distinct aspects of a complex problem area. In general, this process has raised students' awareness of RAI as being in a team allowed them to continuously reflect on the process and thus skills through practice (cf. Connolly, Rush, and Mutchler, 2020).

A final implication of our findings supports calls for problem-based teaching and learning. The course we developed was based on an ongoing research project funded by the National Research Council. As a result, it was directly informed by research (Topi, 2019), thus allowing students to experience the study of a contemporary setting and technology and to deal with real-world problems.

Conclusions

This paper addressed the following research question; *how can we design a course to educate students about responsible AI?* We contribute to the literature in IS and education with a course design (a protocol) that is aimed at promoting university students' collaborative, experiential learning process about responsible AI (Figure 2). The protocol was developed through an action research approach and validated by interviewing student groups to obtain feedback about the course. Other researchers can adopt this protocol to educate students to learn critically about emerging sociotechnical phenomena in the IS domain.

This work has limitations. First, this is the first time running this course, making it difficult to generalize the findings. Second, the study was conducted in a specific context in a course that already had an infrastructure in place providing tools and activities for the students to perform team-based experiential learning. While this proved useful for us, it is not given that similar infrastructure is available at other educational institutions. Third, this study relies entirely on empirical data collected through student group interviews. The students may have been influenced by each other´s answers and we could have missed potential interesting points of view if some students avoided talking because of other students dominating the discussion. However, we hope that the protocol and the findings from the group interviews can inspire new projects on education in IS. An exciting path to investigate further would be to implement a similar protocol in other courses and investigate its usefulness.

References

Asatiani, A et al. (2021). "Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems". In: Journal of the Association for Information Systems (22) 2, pp. 325–352.

Baskerville, R., & Myers, M. D. (2004). Special Issue on Action Research in Information Systems: Making IS Research Relevant to Practice: Foreword. MIS Quarterly, 28(3), 329–335. https://doi.org/10.2307/25148642

Beard, C. (2010). The Experiential Learning Toolkit: Blending Practice with Concepts. Kogan Page Publishers.

Berente, N et al. (2021). "Managing Artificial Intelligence". In: MIS Quarterly 45(3), pp. 1433–1450.

Boer, L., & Donovan, J. (2012, June). Provotypes for participatory innovation. In /Proceedings of the designing interactive systems conference/ (pp. 388-397).

Bovens, M (2007). "Analysing and Assessing Accountability: A Conceptual Framework". In: European Law Journal (13) 4, pp. 447–468.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), pp. 1-12.

Connolly, Amy J.; Rush, Daniel; and Mutchler, Leigh, "Using Role Play to Teach Complexity in IT Project Management" (2020). AMCIS 2020 Proceedings. 3.

Cooper, A., Reimann, R., Cronin, D., & Noessel, C. (2014). About face: the essentials of interaction design. John Wiley & Sons.

Dignum, V. (2019). Responsible Artificial Intelligence. How to develop and use AI in a responsible way. Springer.

Dignum, V. (2021) 'The role and challenges of education for responsible AI'. London Review of Education, 19 (1), 1, 1–11. https://doi.org/10.14324/LRE.19.1.01

Dunne, D. and Martin, R., 2006. Design thinking and how it will change management education: An interview and discussion. *Academy of Management Learning & Education, 5(4)*, pp.512-523.

Dörner, D. (2011), Die Logik Des Misslingens: Strategisches Denken in Komplexen Situationen, Rowohlt Verlag GmbH, Hamburg

Eisenhardt, K. (1989). Building Theories from Case Study Research. The Academy of Management Review, Vol. 14, No. 4, pp. 532-550.

European Union (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence ACT) and amending certain legislative acts, no. COM/2021/206 final (2021). https://eur-lex.europa.eu/legalcontent/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U. and Rossi, F. (2018). 'AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations'. Minds and Machines, vol. 28, no. 4, pp. 689-707.

Glikson, E., & Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals*.

Goh, S. H., Di Gangi, P. M., & Gunnells, K. (2020). Teaching Tip: Applying Team-Based Learning in Online Introductory Information Systems Courses. Journal of Information Systems Education, 31(1), 1-11.

Hall, M. (2018). Service Learning in Information Systems Education: Pedagogical Approaches to Support Experiential Learning and Higher-level Thinking. ICIS 2018 Proceedings. https://aisel.aisnet.org/icis2018/education/Presentations/8

Hanington, B., & Martin, B. (2019). Universal methods of design: 100 Ways to research complex problems, develop innovative ideas, and design effective solutions. Rockport publishers.

Hod, S et al. (2022) Data science meet law. Commun. ACM 65, 2 (February 2022), 35-39. https://doi.org/10.1145/3506575

Houde, S., & Hill, C. (1997). What do prototypes prototype?. In /Handbook of human-computer interaction/ (pp. 367-381). North-Holland.

IDEO.org. (2016). The Field Guide to Human-Centered Design (1st ed.). IDEO.org / Design Kit.

Jobst, B., Thoring, K., & Badke-Schaub, P. (2020, May). Introducing a tool to support reflection through sketching and prototyping during the design process. In /Proceedings of the Design Society: DESIGN Conference/ (Vol. 1, pp. 207-214). Cambridge University Press.

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017.

Kane, G. C., Young, A. G., Majchrzak, A., and Ransbotham, S. 2021. "Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants," *MIS Quarterly* (45:1), pp. 371–396. https://doi.org/10.25300/MISQ/2021/1578

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*.

Lennox-Chhugani, Dr Niamh (2020). Evaluation of EMRAD AI in Breast Screening Project: Final Report. URL: https://www.emrad.nhs.uk/images/AI_in_Breast_Screening_Final_Evaluation_Report_-_Summary.pdf

Norwegian Data Protection Authority (2018). Artificial intelligence and privacy. URL: www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf (visited on Oct. 15, 2021).
Mann, K., Gordon, J. and MacLeod, A. (2009), "Reflection and reflective practice in health professions education: a systematic review", Advances in Health Sciences Education, Vol. 14 No. 4, pp. 595-621.
Matthee, M. & Turpin, M. (2019). Invited Paper: Teaching Critical Thinking, Problem Solving, and Design Thinking: Preparing IS Students for the Future. Journal of Information Systems Education, 30(4), 242252.
Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *0*(0), 1–12. https://doi.org/10.1080/0960085X.2022.2026621
Mogensen, P. H. (1994). Challenging practice: An approach to cooperative analysis. /DAIMI Report Series/, (465).
Orr, Will and Jenny L. Davis (2020). "Attributions of ethical responsibility by Artificial Intelligence practitioners". In: Information, Communication Society (23) 5, pp. 719–735
Preece, J., Sharp, H., & Rogers, Y. (2015). /Interaction Design: Beyond Human-Computer Interaction/ (4th ed.). Wiley.
Schön, D. (1983). The Reflective Practitioner: How professionals think in action. Basic Books, New York.
Shahrasbi, N., Jin, L., & Zheng, W.-J. (2021). Teaching Tip: Design Thinking and Mobile App Development: A Teaching Protocol. Journal of Information Systems Education, 32(2), 92-105.
Sortland, Bjørn (2001). "Experts-in-team–multidisciplinary project". In: International Conference on Engineering Education. Citeseer, pp. 6–10.
Shankaranarayanan, G., Stoddard, D., Gilleran, R., Beitelspacher, L., & Bravo, S. (2021). Experiential Learning of Information Systems in Functional Contexts: The Digital Brand Strategy Project. *Communications of the Association for Information Systems*, 49(1). https://doi.org/10.17705/1CAIS.04902
Tjora, Aksel H. (2019). Qualitative research as stepwise-deductive induction. Routledge.
Topi, H. (2019). Invited Paper: Reflections on the Current State and Future of Information Systems Education. Journal of Information Systems Education, 30(1), 1-9.
van den Broek, Sergeeva, and Huysman (2021). "When the machine meets the expert: An ethnography of developing AI for hiring". In: MIS Quarterly, pp. 1–50.
Vassilakopoulou, P. (2020). Sociotechnical Approach for Accountability by Design in AI Systems. ECIS 2020 Research-in-Progress Papers. https://aisel.aisnet.org/ecis2020_rip/12
Wallin et al. (2017). "Experts in teamwork-A large scale course for interdisciplinary learning and collaboration". In: 13th International CDIO Conference. University of Calgary Calgary, Canada, pp. 1–11.
Wollowski, M. et al. "A survey of current practice and teaching of AI." *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No 1).