SHERPA

Shaping the ethical dimensions of smart information systems– a European perspective (SHERPA)

D5.6

AI in Education Report

28.10.2021

This project has received funding from the
European Union's Horizon 2020 Research and Innovation Programme
Under Grant Agreement no. 786641

Document Control

Deliverable	D5.6
WP/Task Related	WP5
Delivery Date	28.10.2021
Dissemination Level	PUBLIC
Lead Partner	UCLan Cyprus
Contributors	
Reviewers	
Abstract	
Key Words	

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
0.1	11.10.21	Josephina Antoniou, Mark Ryan, Eleni Christodoulou		First Draft
0.2	19.10.21	Kalypso Iordanou		Second Draft
0.3	24.10.21		Doris Schroeder	Final Draft

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Table of Contents

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Executive Summary

The present deliverable addresses a study conducted within the *impact acceleration* project phase, which investigates AI education for professionals working with AI design and development. The deliverable describes the work, which focuses on promoting AI Ethics through workshops aiming to support consideration of multiple values (e.g. privacy; individual, societal, and environmental wellbeing) when designing smart information systems (SIS) to ensure development of ethical SIS. In particular, three 3-hour workshops were organized, where participants were engaged in scenario-based discussions and reflective activities. The underlying hypothesis was that engagement in group discussion and self-reflection on one's own values and of whether values have been considered during the AI design process, will promote higher self-awareness and self-regulation, and will support value-based AI design. Through this short exploratory research study, SHERPA aimed to better understand how professionals working with AI design and development approach their everyday tasks in the context of ethical values, and to help identify the needs and dynamics of a better provision of AI ethics education.

The workshops were conducted with different groups of stakeholders, and the workshop discussions and activities focused on the topic of ethics and how ethical action is implemented in organisations. The stakeholders were engaged in scenario-based discussions, aiming to better understand if and how AI professionals incorporate ethical reflection in their day-to-day activities, how ethical values and practices relate to their organisations' values, and when these clash, how would/do the participants respond.

In particular, participants had the opportunity to undertake activities such as to reflect on their most important professional value and to relate this, if possible, to a specific incident, event or issue from their professional life. The rationale was for participants to think of not just the value(s) but also reflect on whether and how the value(s) informs their practice, so for instance, how they responded or resolved an issue at work having this value in mind. Findings from the data analysis indicated that if more ethics education and training were initiated within these organisations and provided to the AI practitioners, the latter may be more willing, confident and able to implement them in practice.

List of figures

List of tables

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

1. Introduction

SHERPA investigates, analyses and synthesises our understanding of the ways in which Smart Information Systems (SIS; the combination of artificial intelligence and big data analytics) impact ethics and human rights issues. The project aims to develop novel ways of understanding and addressing SIS challenges.

This deliverable describes a study conducted within the *impact acceleration* project phase, which investigates AI education and in particular, aims to examine whether an intervention with professionals in AI-related posts is effective in promoting value-based design. As an extension of the SHERPA project, the study is based on experiences gathered from previous tasks and deliverables of the project. In particular, stakeholder input highlighted the importance of providing education for addressing the ethical challenges of AI for IT professionals in AI-related posts, in particular professionals working with AI design and development.

by Unknown Author is licensed under CC BY-NC-ND

The present deliverable describes the work, which focuses on promoting AI Ethics through workshops by taking multiple values into consideration (e.g. privacy; individual, societal, and environmental wellbeing) when designing smart information systems to ensure the development of ethical SIS. In particular, three 3-hour workshops were organized, where participants were engaged in scenario-based discussion and reflective activities.

The intervention was based on social identity and self-regulation theories (Markus & Nurius, 1986; Oyserman, 2007), self-affirmation theory (Čehajić-Clancy et al., 2011; Badea & Sherman, 2019; Crocker, Niiya, & Mischkowski, 2008) and the MAPS model of metacognition and regulation (Frazier Schwartz & Metcalfe, 2021). In particular, the design of the present study was built on the study of Crocker et al. (2008) which shows that reflecting on and writing about important values reminds people about things they care about that transcend the self and may trigger positive other-directed feelings whilst reducing defensiveness to potentially controversial/threatening/hostile information.

Based on the Frazier et al. (2021) model, we asked participants to engage in reflection and self-monitoring for considering multiple values when designing AI. The underlying hypothesis was that engagement in group discussion and self-reflection whilst considering different values in the AI design process, will promote higher self-awareness and self-regulation, and will support value-based AI design. Through this short exploratory research study, SHERPA aimed to better understand how professionals approach their everyday tasks in the context of ethical values, and to help identify the needs and dynamics of a better provision of AI ethics education.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

2. Methodology

2.1. Ethics Approval and Data Management

Initially, the task-leading team from UCLan Cyprus secured ethical clearance from the Cyprus National Bioethics Committee. Prior to the online workshops the participants' written consent was secured, using the information sheet (Appendix A) and the consent form (Appendix B) released by the task team to the participants. The workshops took place online (via Microsoft Teams) and were recorded and transcribed. To preserve the participants' anonymity, pseudonyms were used. The audio data of the study were saved in password-protected laptops to which only the researchers of the study had access to. All the data will be deleted 5 years after publication of this report.

2.2. Data collection

In terms of data collection for the empirical study, the task leading team organised three digital workshops in the early summer of 2021: the first one on the 29th of June, the second on the 2nd of July and the third one on the 8th of July. The workshops took place via Microsoft Teams and the organisers and task leaders recorded and transcribed these discussions (with the written and oral consent of the participants).[1]

The workshops were conducted with different groups of stakeholders, and the workshop discussions and activities focused on the topic of ethics and how ethical actions are implemented in organisations. The stakeholders were engaged in scenario-based discussions, aiming to better understand if and how AI professionals incorporate ethical reflection in their day-to-day activities, how ethical values and practices relate to their organisations' values, and when these clash, how would/do the participants respond.

The workshops were conducted in English, and each workshop had 2-3 facilitators. The workshops began with a brief introduction about the SHERPA project itself, the aims of the workshops, and an overview of values identified through the course of SHERPA, which strongly align with the EU Higher-level Expert Group's Guidelines for Trustworthy AI (HLEG, 2019) (see Table 1).

List of Values Shown to the Workshop Participants
1. Human agency, liberty, and dignity (positive/negative liberty, human dignity)
2. Technical robustness and safety (resilience, accuracy, fall-back plan)
3. Privacy and data governance (respect for privacy, data integrity, data access)
4. Transparency (traceability, explainability, communication)
5. Diversity, non-discrimination, and fairness (reduction of bias, fairness, avoidance of discrimination)
6. Individual, societal, and environmental wellbeing (sustainability, social relationships, democracy)
7. Accountability (auditability, human oversight)

Table 1 List of Values Shown to the Workshop Participants

[1] Furthermore, during the transcription and analysis, the participants' identities were pseudonymised for greater privacy protection. The audio data of the study was saved on password-protected computers, which only the researchers have access to. The data will be deleted after 5 years of publication of the project.

2.2.1. Participants

A total of 19 individuals participated in the workshops, 15 men and 4 women. Participants were recruited based on extended networks of contacts with interest and expertise in AI and Big Data. The task leader reached out to potential participants via email and those who were interested were sent the relevant information and invited to the online workshop. Every effort was made to have a gender balance in the sample but this was not achievable, reflecting also the wider, structural gender imbalance that exists in the field and the low response rate of the female participants invited to be involved in the workshops.

Two participants were research-active educators of AI design and development, while most (17 out of 19) of the participants were software designers and developers. Within this groups of stakeholders, 2 worked directly with AI applications development, 7 in AI cybersecurity, 5 in AI media, 1 in healthcare AI and 2 in AI information security/network management. The range of countries the participants were based in, ranged from the UK, Ireland, Poland, Ukraine, Italy, Cyprus, USA, Czech Republic, Russia, and Finland.

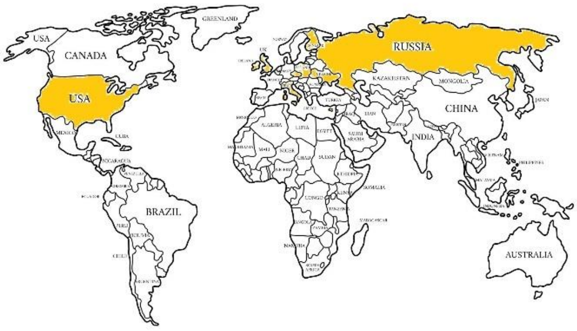

Figure 1: Range of countries represented by the participants

2.2.2 How were the workshops conducted?

All workshops were conducted in English and each workshop lasted for three hours. Each workshop had 2 or 3 facilitators depending on the size of each group.

The first part of the workshop consisted of splitting the participants into groups and relocating them to virtual breakout rooms. In these rooms, the participants worked in pairs, describing their most important professional value and an event from their professional life that demonstrated how they responded to, implemented, or reflected about, this value in practice.

This was followed by a presentation of a scenario for the participants to discuss (see Appendix C). Pariticipants were asked to brainstorm different values that they could identify from the scenario, and why / how these values were important to them. Participants were divided into two smaller groups and were tasked to work as a team to respond to the scenario. They used a digital file to keep track of their discussions and were given an hour to discuss how they would approach the assignment. Within these breakout groups, the team used prearranged questions to enhance discussion and to encourage participants to discuss issues regarding the causes of disagreement for the design and use of advertisements, issues of responsibility, transparency and accountability, and their views on negative impacts of their work scenario, for instance, on society, environmental sustainability and human rights and liberties (see Appendix C).

Finally, the participants were invited back to their original pairs (first task) and asked to participate in a short 10 minute reflective exercise. They were asked to reflect on whether and how the design decisions of the project were consistent with their most important professional value and also to give us their feedback, thoughts and views on how the workshop affected their way of thinking regarding ethical values and their professional work.

Afterwards, the workshops were analysed using a thematic analysis (Braun and Clarke, 2006), which can be understood as 'a method for identifying, analysing, and reporting patterns (themes) within data. It minimally organizes and describes [the] data set in (rich) detail' (Braun and Clarke, 2006, p. 79). The codes created were based on an analysis of the transcripts from the workshops, and overall, we followed Braun and Clarke's six stages of thematic analysis (Braun and Clarke, 2006, p. 87): (1) Initial data familiarisation; (2) Generation of initial codes; (3) Search for themes; (4) Review of themes in relation to coded extracts; (5) Definition and final naming of themes; (6) Production of the report (see Figure 2).

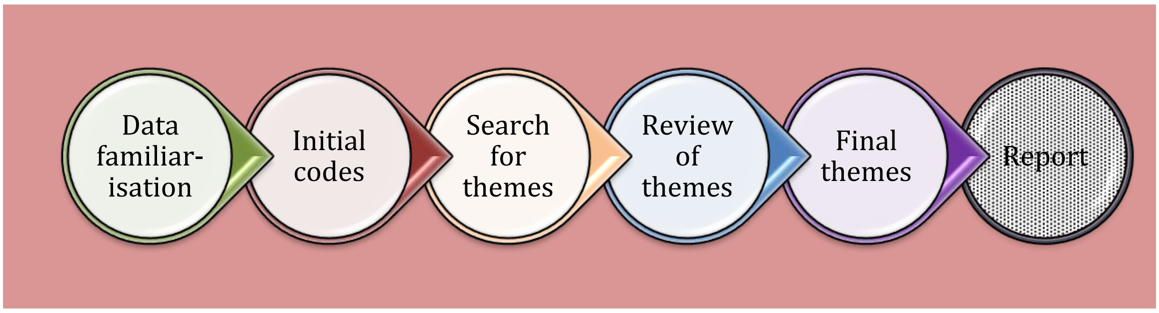

Figure 2: Braun and Clarke's (2006) Six Stages of Thematic Analysis

2.3. Data analysis

The data was analysed using the data analysis software NVivo (Version 2020). Only one researcher was involved in the data analysis to ensure consistency throughout the coding, but the team discussed all the findings, concerns, and ideas. The coding of text segments was not mutually exclusive, for instance one paragraph could include up to 5-6 different codes. For the coding process, the team followed closely the instructions and advice of Braun and Clarke (2006). Specifically, the team took a qualitative approach that focused on identifying emerging themes, understanding the 'what and how' of what was being said and utilising the insights and experiences in order to answer our research question which was how individuals with professional AI-related experience perceive values as they relate to software design and ethical concerns related to AI (and Big Data) technology and the effect that this educational intervention had on their understandings and perceptions of the importance of such values in their professional everyday life.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

3. Summary of Findings

3.1. Professional values – Pair exercise

In the first part of the workshop, as outlined above, participants had the opportunity to reflect in pairs on their most important professional value and asked to relate this, if possible, to a specific incident, event or issue from their professional life. The rationale was for participants to think of not just the value but also reflect on whether and how this value informs their practice, so for instance, how they responded or resolved an issue at work having this value in mind.

Empathy

This was a value mentioned several times by a female participant, Eva[2], which was not included in the initial list of values. She discussed the importance of empathy in the particular scenario given to her. She also gave a personal example of how she uses empathy in her classroom. When her students - who are not native speakers of English - are struggling to present their assessments, she tries to step in their shoes and appreciate that they may be speaking 3 other languages (Russian, Polish and Ukraininan) and also to positively encourage them in their efforts:

> "I think that would be **cognitive empathy or the ability to see different perspectives.** And I'm trying to work on this myself and with my students. So to be able to see different people and what they may need, for example from the software. Yes. That would be in the context… the scenario here. Looking at, for example, students and how they work and what they may struggle with, trying to understand where they're coming from, the background and appreciate the effort… and why, possibly there were some flaws in their presentation." (Eva, W1)

by Unknown Author is licensed under CC BY-NC-ND

Eva believed empathy was important both from her side as an instructor but also as relevant to the dynamics of a group, in this case her students. She argued that this value is important to be taught to students so that they understand that empathy is not just something that is 'given to you', you need to be able **to make an effort** in order to see someone else's perspective. Eva also related this to '**confidence'** and its importance in order for individuals in group settings to be able to function well. Students should be able to recognise, she said, that if their fellow students:

> "are making mistakes, making grammar mistakes, they don't sound perfect, it's because this is not their first language and you need to appreciate this. It's the same like you would be asked to speak Russian or German or any other language. You would probably not be so fluent and confident, and the confidence, the confidence as well, building up the confidence that the effort is important and I can see that the **starting point** was different compared to others". (Eva, W1)

In other words, empathy would help in better communication, would empower individuals to reach their maximum potential and avoid being overly critical of themselves and of others by engaging in inaccurate and unfair comparisons.

[2] Note that all names used in this report are pseudonyms and do not necessarily represent the participants' real names.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Persistence [3]

Persistence was also presented as a value, which was not included in the original list. This was seen as central when faced with new software and other challenges; to be able to not easily give up but rather be resilient:

> “one of the most important values at least that helped me in real life situations was persistence. So I was introduced to this new software that I haven't encountered before and I had to figure out how it works and I didn't have an option to ask anyone around me and none of my colleagues didn't know how it works so by being persistent, I was able to look through the Internet and get in contact with other people or from support from different communities that actually helped me figure out a lot of stuff” (Sherry, W2).

Creativity

Another interesting aspect presented as an important value, and again not in the original list, was creativity. The participant who suggested this as their most important value explained how they felt that this was really missing from everyday professional contexts as people tend to get lost in just completing their tasks, and yet sometimes the key to success is to be imaginative and think outside the box:

> “Creativity is something that is forgotten a little bit. A lot of people get the tied up with their work, tied up with the task they have to do and they don't stop and think about how they you know how they might do something differently” (Simon, W3).

The participant then gave a concrete example where they practiced creativity and then successfully solved a problem:

> “Recently I started playing around with genetic algorithms. Because I've been largely trying to solve like toy problems with reinforcement learning and I applied genetic algorithms to some of these toy problems. And this is like so, swarm systems, so multiple agents all independently training to work together to solve a task, and they weren't training with any of the reinforcement learning approaches that I tried. But actually when I tried genetic algorithm so I got a, like almost immediate success. It was like magic. I'm not quite sure how it works, but these agents were actually evolving optimal or somewhat optimal policies in a rather short amount of compute time. So, I would say that that was like a recent success in creativity. It's just, approaching the problem from a different angle” (Simon, W3).

This Photo by Unknown Author is licensed under CC BY-NC

This approach regarding creativity was also espoused by another participant who mentioned creativity as one among other important professional values. Their position was that given that people working in “the data science domain…are struggling with pretty hard problems”, creativity was crucial and ideally this should be done collectively through brainstorming sessions that would involve various ideas on how to solve an issue (Jeremy, W3).

Human dignity, agency and liberty (freedom)

[3] Persistence is not an ethical value, but since the educational intervention began by enquiring participants about their professional values, it was expected that the values that initially came up did not only include ethical values, but in this case other professional values such as persistence [and creativity]. The educational intervention will highlight the ethical values and record participants’ potential learning or change of perspective.

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Some participants chose human dignity, agency and liberty as their most important values. This set of values was included in the original list given to participants and it seemed to resonate with several of them. Alex saw this set of values challenged due to assymetrical power relations in companies and discussed its importance. Lack of this set of values related to human agency was seen as having negative consequences on transparency, informed consent, personal growth and intellectual curiosity. In other words, it did not lead to human flourishing and digital well-being.

> " five or six big companies in California are starting to have consolidate a lot of **power**, and they pretty much analyze everything we do [...]. You know.. it's all legal, but a lot of people just don't have any idea of just how much data is going towards them and all to sell ads. And then we have services on devices that are **locked down** and you know users nowadays have much less freedom than they did, like even 10 or 20 years ago. And I think that's **a pretty bad thing for overall development of humans themselves in society**. It's giving us less room to **grow as individuals** and to be intellectually curious…just making us more of ad driven consumer culture rather than creators or things like that. So I think for the future, that would be the most, the biggest thing to strive towards, and you know personally I do try to promote **free open source software** and you know, try to promote even even for stuff that the company I'm not now trying to support support for open operating systems such as Linux, because there's such a heavy focus on like lock down things, like Windows and the two mobile operating systems right now" (Alex, W1).

As we can see, in the quote, what is seen as a challenge to human freedom and agency is not just the inequality in terms of power but also the asymmetry in terms of **access and openness of information**. So we can discern that one associated value is the openness of information and the ability for everyone to access it freely.

Inclusiveness and bottom-up approaches

The value of inclusive participation when making decisions or designing software can be seen as related also to the openness discussed above and being open to listening to other opinions and even revising the project priorities. It can also be seen as related to empathy i.e. to be able to understand the negative impact of exclusion, to consciously adopt participatory decision-making and co-creation.

> "so for me… in the team it's mostly to do with **co-creating and participatory** discussions. So how we can bring in communities for which these projects are being developed as part of them, to make them work **as part** of the developments. So for me in terms of professional value, it's more important to have **inclusivity** in mind, so having a more open approach to everything so that comes with having more, like keeping **an open mind to making changes** and, maybe changing some of the priorities that we had for project" (Lara, W2).

In this case the participant gave her experience from working with a marginalised group in London and talked about the importance of including the communities one is supposed to be working for and start working 'with' them rather than 'for' them:

> "the idea was to create the project for them and we eventually ended up bringing them in and listening to their ideas and their concerns and including their concerns in our design of the project. And that was basically what changed my perspective into bringing in those external, entirely external, but not professional ideas within the project, while before it was more to do with just having professionals included in the project"(Lara, W2).

Responsibility

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

The value of responsibility was interpreted as dealing with ownership issues and liability 'for your own actions' and for 'who you are' (Larry, W1). For this to happen it was seen as important for the individual to feel that they are in control or that they have certain control over ownership and liability. Responsibility was presented as being important in a professional context not only for the technology producer but also as responsibility as a user:

> " And we talk about responsibility in terms of software development and the possible implications of technologies, but also [it] is about our responsibility for ourselves and our actions and how we use technology".

Nevertheless, it was also recognised that taking responsibility was not an easy task:

> "that's actually also a really difficult thing knowing how to take responsibility for your own actions for your own behaviors. You know there are small things from you know, how you behave using things like technology, maybe via social media and taking responsibility for the actions, maybe if you comment on something or put out a posting through, you know, how your how you kind of interact with someone in a kind of direct manner and how you kind of **take responsibility for not just what you do, but how that might then be reflected in them**. **In other words, how it might make them feel…** There are times when students send me emails and you kind of think 'you really have not thought about what you're saying and how that's gonna come across'. You know all you're really thinking about is I want this, rather than actually I need to get someone to help me in order to be able to achieve that. And equally in my response as a professional academic, I could quite easily go 'Oh my God, you're just being horrible. You're being incredibly demanding and thoughtless'. I mean I actually had a student who phoned me I think it was about 7:30 in the evening when I was sitting down for dinner with my family and it just so happened it was also my birthday and she was, well, I I just need to speak to you now. And I just had to say look, I really can't, you know this?" (Larry, W1)

Interestingly, in this quote, responsibility is related to **respect**, to **accountability** and also to **empathy**. Although the latter is not explicitly stated, the phrase "you really have not thought about what you're saying and how that's gonna come across" is strongly related to the skill of empathy i.e. to be able to step in someone else's shoes.

Technical robustness, transparency, and being in a company that 'does technology right'

Regarding technical robustness, participants varied in terms of their reasons for preferring it. Some preferred it for **practical reasons** related to the success of a project and related it to issues to do with privacy legislation and how not doing so could lead to fines to the company or the client, while others took a more **ethical and personal stance** and emphasised the moral reasons for 'doing technology right' rather than focusing on the 'stick' or punishment that not doing so would entail.

One participant argued that technical robustness was important as it determines the overall success of a project. They argued that the technical aspects go hand in hand with privacy. They related this to the General Data Protection Regulation (GDPR) and how this 'can't be overlooked, on the one hand 'due to very high fees and punishments that can come not only for the designer and the development team, but also for the client' and on the other hand because this 'could negatively affect the design' and if a product is produced without taking that into consideration, it could completely fail' (Mary, W3). In other words, the emphasis was on the **success** of the project and **avoiding fines** rather than the ethical reasons for technical robustness and privacy. When participants further expanded on the reasons that the designer should have the data safely stored the rationale was again practical and going back to **EU policies and legislation**:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

> "... to avoid the system being hacked or even somebody on the team misusing the the data. Because all of these things will and violate all these policies that are in place that as a designer, we have to uphold in the EU" (Mary, W3).

Mary also gave a specific scenario in order to explain what she meant regarding technical robustness affecting the success of a project:

> "I'll use a scenario, for example of an E-shop, for customer web application. If it's designed for a normal Western market, then having something, some issue that doesn't have proper functionality, such as not really adding correctly on items into the cart or adding numerous quantities. This could lead to the user not only being unhappy and not wanting to use the product, but potentially paying too much or too little for the product by accident and thus the client will have to reimburse or have to really look into it. And these could be defects that could badly reflect on the designer, that the designer didn't take this into account and as well with the display of items on web applications. This can contribute to the users not using the product… or if the fonts are even incorrect or too small too big, these all influence the overall success of a product" (Mary, W3).

Other participants took a less 'legalistic' view and focused more on the ethical values and imperatives. One participant provided an example from their professional life where he was at an early stage of his career and was pondering whether to leave the software consulting firm he was working for or not. Even though he was enjoying his job, had positive relationships with his colleagues and the company was treating him in a really nice way, one of the reasons he decided to leave was the lack of transparency:

> "I think, we were basically building a business on the on top of the **ignorance** of our customers. Which were **not really understanding** what we were selling them. So yeah, I decided to change something." (Thomas, W1)

His decision to leave was also related to his belief that the company was not creating something 'good for me or society' and it was also not doing it in the **best technical** way possible, thus fitting also with **'technical robustness'** from the fixed set of values that was given to participants. Knowing that one is producing something good for society was seen as a strong motivation to continue working in that job:

"… I want to work in a company that does technology right …"

> "I think technology should like improve society and it's not just a way to make money… I feel it's really important for me to work in a company that produces a **positive societal impact.** And it really motivates me to do that. So, at that, for example, now I'm working a cyber security company that even though our services are expensive, we kind of fight the bad guys in a way. So, **I feel really motivated to wake up in the morning and go to work**." (Thomas, W1).
>
> "I think we were selling solutions that were **subpar from an engineering point of view.** I would solve the problems that we were selling, the solution that we were selling. I would have solved them with other things and it would be better for the client. But since we were backing up this product, **we had to sell the product and not the right solution**. So, both transparency and, **I want to work in a company that does technology right.** So, I want to sell the best product available. **Or develop the best solutions, without compromises**." (Thomas, W1)

Technical robustness was also discussed in terms of methodology and how this affects research outcomes. In one case a participant also left his job in a similar way to the participant above, precisely because he felt that behind the strategy that the medical charity was espousing was a faulty methodological approach that meant that what they were promoting was essentially flawed:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

> “I was brought into a health charity to create a research strategy and develop their research program. And that, the charity had at the core of this a piece of research that it was well known for and what the charity did, was based around this research. But **it was actually a very flawed piece of research**, and I'd come across this, you know, many years before or a few years before I started working for them. So, when I got brought into to work for them, I tried to sort of say to them look, this is not a robust piece of research, and on what we should try to do is either make it stronger, you know, or let's go at it again or trying to develop it out or move away from it. One or the other, but we can't really, I couldn't, **I didn't feel from an ethics point of view that I could stand over this piece of research.** So, there was an ongoing tussle about this over about six months where I kept trying to, you know, bring in different perspectives around how they could, you know, how we could improve the situation and eventually that just didn't work. I couldn't, I guess…it was too embedded in their culture too for me to be able to change it. So, I ended up leaving the company or the charity” (George, W1).

It can therefore be seen how for some participants it was crucial for the company’s values to be morally aligned with one’s individual values. Not just **ethics** but also **emotions** – which are rarely discussed in the context of Big Data and AI – play a central role in everyday decision-making of the software developers.

It is also interesting to note that in both cases discussed above, the participants felt that the situation was not inevitable, that the companies *could* and should have done something to improve technical and methodological robustness.

> “I just thought that research in this area of practice was so important that it couldn't have these kind of things happening. And also they didn't need to do it as well, **it was unnecessary**. They were well supported charity so…they could have moved away from it. **We could have moved into a more positive future about…**I just felt that I couldn't stand over what they were. (George, W1)

Also related to the value of technical robustness and doing something that has a positive societal impact was an example of top-down pressure that some developers have to choose speed over quality and to ‘do things fast in a bad way’. In this case, the participant acknowledged that it was important on the one hand to be ethical, especially if you are conscious of doing technology in a way that can create more harm than good, but on the other hand was cognisant of the structural difficulties that individual and especially young developers may face when they are not only lower in the hierarchy but also need to keep their job and don’t have support from their peers or superiors:

> “it's actually a huge problem in this in the modern software development, especially for junior developers, you come to work, you do something and then you have a pressure from the business **to do things fast in a bad way**. And, it's a lot of luck when such people can get support from, let's say, more senior members of the team. But it doesn't always happen like that. So I don't know if I manage to explain what's my important way most important value: but not doing bad things knowingly” (Chris, W2).

Ultimately, the participant argued that the most important value here is to have ‘respect to software development’ meaning technical robustness and integrity and not taking ‘shortcuts’:

> “There is a culture in the modern software development of **doing things fast** and I think it's **very, very misused**. Doing things fast is extremely valuable when we do the research or when we do the pilot project and so on, but quite a lot of people, especially somehow business people, expect doing things fast and **taking shortcuts** in the production software no matter what you are doing. And very often very often there are things in the software development **you just have to do well to not to cause any potential harm**. Meaning things like privacy protection or authentication or safety or whatnot” (Chris, W2).

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Finally, in terms of **transparency** one participant discussed this value but initially not in the context of the company being transparent to the user but for developers to be transparent within the company, so with their colleagues:

This Photo by Unknown Author is licensed under CC BY-NC-ND

> "let's say a new developer and by accident he stores the credentials in the repository. If no one sees it and no one is alerted, then this is the real problem. But you know, everyone can make mistakes, but we need to share that it happened actually" (Jeremy, W3).

However, after being probed by the facilitator, the participant then also added the angle of also being transparent towards the client:

Facilitator: "You mentioned transparency within the company, but does this mean that also you have transparency towards your customers or any other stakeholders?"

Jeremy: "Yeah it is a crucial part as well, because when something happens we need to contact, the clients as fast as we can actually. So, yeah…for example if the hacker is successful in getting into the server of the company, it is super important to let the company know as fast as we can to somehow lower the impact of that breach…so yeah, so that's why it is important also to be transparent to clients."

One may argue that Jeremy's approach was more aligned with Mary's approach above, where the reason for this value was for practical/technical reasons – in this case to "lower the impact of the breach" without any explicit reference to ethical reasons. In fact, towards the end of the discussion Mary pointed out that in her field transparency was important but *not* the most important value:

> "in my field **transparency** is very dependent on who the client is, it's not the highest value completely, even though it is still important, but it's not at the complete top." (Mary, W3).

Trustworthiness

This value came up once and was discussed both in terms of trust in humans and in technology. It was seen as an essential part of having functional everyday professional environments as the lack of it meant adding extra layers of unnecessary stress and wasting time and energy. In other words, time was of the essence and trust was something that was seen as important but at the same time the participant seems to prefer to 'let go' of this and focus on time.

> "My main value that I posted was **trustworthiness**. That goes in my life right now. And that goes to different systems that we built and use. Because to me that is a major value because you save a lot of time even. At my work, if I assign it to somebody and I'm **worrying** about him, is he going to do it or not? Is she going to get lazy or something? So, I'm **losing my trust** into that person. Then I'm losing valuable **time** from my time because I'm going to have to go back and check on that person and worrying if this is… I mean, the task is going to be done…
>
> "I use it the same as when building a system like, you know, an artificial agent or any other software. **If we have to worry that, if that system is, it's not 100% trustworthy, then we are losing lots of time**… If you have a system that is not giving you the correct values or responses and you have to have somebody to work, check the values or the data collected all the time if it's correct, if it, was collected, you know, in a timely manner or whatever you know. I believe it's the same thing." (Mike, W3)

3.2. Group task

The group task features debates over responsibility and human agency, responsibility to educate, individual vs. societal well-being, democracy, role of the state, regulation and the nature of advertisement.

During the group task, the nature of the scenario and the fact that there were two different core positions regarding the specific task assigned to the imagined company triggered discussion. Participants were asked to work as a team to clarify and decide upon the key design and development objectives and requirements of developing a new social platform, and this naturally meant that there were often debates, agreements and disagreements over a range of issues. These were related to which values should be prioritised but also regarding the nature of **responsibility**: who should be responsible for what, and to what extent should the public or the state shoulder responsibility versus the consumer or the company. Some participants argued that responsibility should be given to the state (or political forces), others to the wider public in terms of pushing for more accountability through a social movement and others emphasised the role of the individual (or a combination of the three).

The fact that responsibility emerged as a prevalent theme in the findings can partly be explained due to the fact that after the group finished their task there were specific questions posed to the group (see Appendix C) asking participants who they think should be responsible for important decisions regarding the design process or who should be held responsible if personal data were leaked etc.

In order to ensure that the diversity and the richness of the discussions is reflected well within the limited space available, each workshop is discussed individually (but not exhaustively), focusing on the more prominent areas of debate and avoiding repeating the same arguments if they were discussed across all the workshops.

Workshop I

During the first workshop there were debates around the following themes: responsibility; human agency; over-regulation; individual vs. societal well-being; democracy and the role of the state; the demonised nature (or not) of advertisements; and environmental sustainability.

In the specific extract that follows, we see that the first two participants reach a common ground regarding the importance of human agency, although there is a difference as to the order of other values. For instance one participant has privacy and data governance as a higher order value whereas the other one believes accountability is at a higher level in the value hierarchy. The third participant who enters the dialogue questions the assumption that people want to be placed in a position of responsibility and argues that perhaps these values including human agency, privacy and data governance, transparency etc. are "too much" for the public and that in fact they prefer it when the responsibility is taken by the developers or by technology itself: "they want other people to look after them and to protect them" (Larry, W1).

George: "I think **human agency** is my most important value and then privacy and data governance and then transparency. So I'd be in the kind of group that suggests a registration fee or a subscription fee and full transparency. And…also.. I know that there's business problems sometimes around that kind of thing, but I wonder if there's kind of **creative** ways to be very **upfront** with our customers and give them the **choice** about how they want to…you know, if they want to sell their data, or if they want to just pay a 5 or a month or whatever and have access to the services...so that's what I would be kind of putting forward" (George, W1)

Eva: "Yeah, I completely agree with this **human agency**. I think for me this is where everything starts and I would add to this **accountability**. So if we have those two, we already take care about others. Yeah, because we are responsible for our work and we respect other people. So transparency, privacy and

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

data protection and everything else for me comes as a result of that, uh, that yeah, it's just basically being more specific. But agency and accountability. These are the higher level ones." (Eva, W1)

Larry: "OK, so to take what they call the devil's advocate position, maybe people don't want to be put in that position where they have to be responsible. Maybe people don't like having to make those decisions. **So maybe human agency is actually too much. And really, what people want is that they want other people to look after them and to protect them**. So things like privacy and data governance and the sort of transparency need to be taken care of by others, in other words by the technology by the system, by people who are developing the system. Just because people are like, well, you know. I mean, how many people actually do read the terms and conditions rather than just accept them and go well, "I'm sure it's OK. Someone else will be sort of taking care of this", so I I'm not, you know, I'm not sure that human agency can be put at the top of that list. I think we, **we allow people too much agency in some ways and they don't really want it. They actually want other people to look after them**". (Larry, W1)

The difference between respecting and protecting was also raised by one participant, who argued for at least aiming for the latter in terms of privacy and data governance. Another debate that emerged was regarding individual vs societal well-being. One participant stated that their most important value is individual and societal well-being. This debate reflected a wider disagreement of individual interests vs the common good in the context of a **democracy**.

During the discussion Larry referred to empathy which Eva had originally mentioned, showing how workshops like these offer the space for participants to brainstorm regarding the importance of values but also be influenced by each other's views and think of values in ways they had not thought of before. In this context, Eva stepped in to clarify further what she meant by empathy and how she sees it being useful for the developer. She clarified that for her empathy included "perspective-thinking" and "perspective-taking" and to better allow others to understand your position she suggested 'breaking down' the information into smaller parts:

> "From perspective thinking where we are breaking down information, people can see, oh, this applies here. This applies here or this is important to take into consideration when we are taking care about this small bits of information." (Eva, W1)

In this workshop the issue of environmental sustainability was brought up in passing by one participant, the youngest of the group, *before* the specific environmental question (see Q8, Appendix C) was posed by the facilitator, which was evidence that this was indeed an important aspect for them, or at least one that they had considered in their work:

> "we should strive thrive to build a system that should do no harm or the least amount of harm possible…It should be sustainable both economically from the company side and **environmentally**. This could be achieved, for example, by funding some sustainable program, maybe like a compensating carbon emission one…" (Thomas, W1)

When specifically probed about potential implications regarding environmental sustainability that the features of the platform they were creating could have and if they thought this was an important aspect that should be addressed more a debate ensued where on the one hand participants argued that environmental aspects were not adequately considered and "they should be taken into account more than they are now" (Thomas, W1) and on the other hand there was a counter-position to this by another participant who basically questioned how realistic or feasible this was in practice.

> "Let's see uhm do you every time you do an Internet search…do you think about this? Do you think: Oh, I won't do that search because..?" (Laurence, W1)

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

After a third participant posed the question to the team what the best way to move forward was, the issue of ethics washing was implicitly discussed – in this context how companies are now using the 'green' aspect as a marketing technique.

> "No, but I mean.. I mean some companies use it as a marketing, don't they? So they they will use their kind of green credentials as a way of kind of marketing themselves and trying to kind of, you know, put themselves into different market essentially. On the wider perspective, I think you have to look up, you know where, where is the key elements to the issue. So you know we can all turn off our TV's at night and save water. But if you know they're building power plants every day in China that are pumping out more greenhouse gases, then is our little bit gonna offset about that big bit?" (Larry, W1)

The counter-argument presented by Thomas was that at the end of the day unless bigger structural changes were made at a company level, government level or through a wider social movement, then individual actions will would ultimately lead to deep-rooted and substantial change:

Therefore, the suggestion by Thomas was that only *collective* political and social action would possibly lead to environmental change:

> "So you have got to get the kind of political forces together, but they will only do it if there's a political will to do that. And it's a bit like I'm old enough to remember the.. Uh, stop littering campaign and the way that you stop littering is you don't go out with a big brush and sweep it all up. You tell everyone it's their responsibility for their own little bit of litter, and then they pick that little bit up. Then together as a body. People stop littering, and in fact then there's a social pressure not to as well a bit like smoking became a social pressure to stop smoking. And then there's a political will to make that happen." (Thomas, W1)

Workshop II

During the second workshop there were less intense debates compared to the first one. The discussions (and often the consensus) were around ethical issues related to how profit is being made; demand, cost and profit; transparency to the user regarding data collection, inclusivity, responsibility, environmental issues and human rights.

One of the first questions asked was the extent to which there was enough demand to be able to sell the service if they opt for a scenario without advertisements. They discussed the fact that they would have to find a way to be financially sustainable given that keeping the servers up and running would cost them money.

Regarding demand, one of the participants argued that they would have to have a factor that made them stand out when compared to others, otherwise they would not be able to financially survive and the company would fail. This participant preferred the no registration fee model and suggested donations and crowd-funding as a source of keeping the company going, similar to that of Wikipedia:

> "I would much more prefer to do something for the common good based on donations like Wikipedia. Nobody pays for Wikipedia to access it, and nobody sees smart advertisements on Wikipedia, but it survives…it shows that community funded model is actually something that can survive. Because seriously, if we are building Facebook, , Twitter or LinkedIn, we already failed before we start. We won't be able to compete with those, so there should be some differentiating factor and we should know what is this different changing factor?" (Chris, W2)

One ‘differentiating factor’ that participants achieved consensus on was having a **niche market** and one that had an explicitly **ethical** orientation:

> “We'll have a differentiating factor of not being evil. So like, no smart advertisement, no stealing from the users and stealing…most of the social networks if not all they are stealing from users, they are stealing from users even when the users give their consent because everything is built in such a way that you don't understand what you actually give away for free” (Chris, W2).

In this model, there would not just be consent but properly informed consent so that the way the users’ data is used is adequately and properly communicated and will be understood by the user.

Participants sometimes were themselves unsure, brainstorming as a process of evaluation, raising their own dilemmas as to which of the two options in the scenarios they would choose. The point raised by the following participant is that whereas opting for registration fees would be more ethical in terms of privacy and transparency, this would not necessarily make data more secure.

> “The average user doesn't really know or doesn't really know the extent to which their data is being collected, so they don't really mind using another social media platform, but it's taking their data. But if we consider the ethical parameters of not using someone's data without their knowledge or with their minimal knowledge, then we could go with with Group B with that option of creating a more sort of ethical option of giving people the idea that this social media platform is… you pay for it, but you'll get a more safe environment or like more secure and let's say data focused platform. But at the same time it doesn't make it kind of more secure in terms of like privacy, so I don't know what would win in me…” (Lara, W2)

Another challenge that the group members discussed was the extent to which **diversity, equality and inclusiveness** could be achieved as objectives. It was argued that these need to be built into the system from the start:

Again, in this discussion agreement prevailed over disagreement. For example the reaction to the above comment by the participant who suggested Wikipedia was:

> “You make a lot of sense…unfortunately, the resources are like huge philosophical problems and I don't know how to tackle it… Quality is a process, it's not something that can be added on top later. It's not like we build something and then we add good values on. It’s not going to work like that. You're completely right.” (Chris, W2)

Here we see that the discussion of values has led to a consensus as to not only the ethical nature of the objectives but also that these ethical values should be built into the system from the beginning. The agreement was to try and address this issue by ensuring the team members included both technical people but also people who work in other disciplines that have more “qualitative” and “social approaches”, so teams should be more diverse “in order to create a more diverse product” (Lara, W2).

The team went into deeper discussions of how they could possibly “test” (Clark, W2) if their work process and product was indeed ethical, inclusive, transparent etc. One suggestion was to start by outlining what they do not want to do i.e. what **harms they want to avoid** such as avoiding bias. In other words in the design process they would consider what they “want[ed] to stay away from” (Chris, W2). One example of this was big data used for personalised advertisements and selling private data to third parties. Another was not operating in places where the government would have extensive involvement in accessing private data (originally China was mentioned but then other countries like the US were discussed who are still guilty of it with the only difference being that the latter do so behind the scenes and less explicitly). One

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

other thing the participants agreed on avoiding was “third-party locking with respect to software” (Chris, W2).

The team ended up mostly focusing on **positive-worded objectives** e.g. **the right to be forgotten, transparency regarding users’ rights, target non-technical users, security and privacy, addressing users’ needs and added value, inclusivity, portability, openness (open source code) and localization**, the latter being related to inclusivity in that it allows “people with different languages with different mother tongues [to] use the platform from the start” (Chris, W2). In general, the team concluded that inclusivity and the right to be forgotten were absolutely crucial in relation to the protection of **human rights**.

Interestingly, in terms of environmental implications, in one of the groups of this workshop one of the younger participants suggested for the team to add the importance of **“environmental** impacts and sustainability goals” and the rest of the team agreed, offering their own reasons and examples of why this was important. The younger participant argued that it was important to consider “server capacity and how much energy your servers are and these are sensors are needing” (Lara, W2).

> **Lara:** “I think it's not a priority, but it's something that we maybe consider.”
>
> **Clark:** “We should really consider…”

Chris: “Yeah, I agree completely .We definitely should consider such things because I can give you an example like there is a blockchain buzzword. And people like to put it like everywhere. That's a pretty much a solution without a problem. But energy consumption is horrible. Like all the all the things Bitcoin…Check the price of 1 Bitcoin transaction. This is a disaster like with respect to carbon footprint and it’s very hard to understand. I don't think there is even any good research how you do like sustainable software, because from one point of view you want the security and security in our world means pretty much like open public key security and those things are computationally very hungry.”

Despite not knowing exactly how to do this, this team agreed that they would put ‘thinking about sustainability’ as an objective before adopting a technology. This was in contrast to the second team of Workshop II. Whilst they agreed that it was an important issue, they reached a general consensus that it was not something that developers thought of everyday, or something that they *should* think of everyday as it fell **outside the scope** of their work:

> “Yeah, it's it's really relevant but nothing that I guess we think about on a daily basis… Not sure what to do about it because the trend is really going in the direction that there's only more computation being done already all the time and more data being stored. But it… we we can't neglect the fact that it is really, really consuming a lot of energy… usually companies rely on cloud platforms for computation and storage, and it's not really in the hands of the individual company to do anything about it…I don't think it can lie within the scope of the company.” (Patrick, W2)

On responsibility, when specifically asked by the facilitator regarding who they thought should be **responsible** for making the ultimate decision about whether advertisements will be used, the collective answers of Workshop II included regulation at a regional EU level, the government, stakeholders, company leaders or owners and multidisciplinary ethics boards. Regarding the latter, however, one participant raised the possibility that ethics boards may still not solve the issue of who is ultimately responsible. Instead, he suggested that there could be within the company:

> “a single person being ultimately responsible for ethical questions…because otherwise you are getting designed by committee and nobody is responsible. So there could be a committee for performing the work and finding a decision and so on but finally, there should be accountability, responsibility. There should be a single person like you have in a big company. You have a lot of

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

> people who work with finance, but ultimately you have CFO who is responsible for financial standings of the company" (Chris, W2).

Workshop III

Workshop III was also similar to workshop II in that there was more of a climate of consensus rather than intense debate. There was more or less consensus that objectives or values like technical robustness or privacy could be achieved as there was concrete guidance, especially on an EU level on how to follow this through. However, it was noted that other values required more careful consideration of the type of moderation mechanisms that should be put in place:

> "Technical robustness …privacy and data governance…There is plenty of literature and guidance about how to do these things properly, especially in the context of the EU. I think that the important things though, the ones like **human agency, liberty and dignity individual societal environmental well-being, diversity and non-discrimination, those ones require moderation like or ways of ways and mechanisms of achieving some sort of moderation** on the platform and then for the last the transparency and the accountability; so those things are either ways of defining like how this moderation happens or defining how decisions are made on the platform" (Simon, W3)

Although there were disagreements as to the nature of the platform with some participants preferring the platform to take a more general nature like Facebook or Twitter and others preferring to have something LinkedIn. They agreed though that this decision would ultimately determine also the nature of the problems they would need to address later on as a team, for example if it was misinformation in the case of the general platform or if it was a matter of fake accounts and cybercrime:

> "different social networks have different problems right, I mean, LinkedIn if we're going for a like a social network like LinkedIn, that's gonna be a different set of problems to something like Twitter or Facebook. Uhm so I mean on LinkedIn you still have people who create fake accounts in order to like phish people… in preparation for like cyber attacks, things like that, or just to **scam** people you know, so, so these are things you want to defend against. But then on other social networks where people are talking about politics and general things, then you have the problem of **disinformation** as well right so I think it it's sort of, I think that the problem statement is defined by what your social network is trying to do" (Simon, W3).

Another participant agreed and added to this the suggestion that in order to protect privacy and prevent discrimination it should be built in the design of the platform in an automated way. In other words the objective here would be automate consent or privacy rules so as to have an automated solution within the platform that does not discriminate from the start:

> "I also agree about the discrimination, but as well as important to note like with privacy policy, there are protocols and regulations and information to which as well. Companies can follow and there are rules and in reality when designing something like that. I think it's important, then from the side of the project or the developers to really implement that the social media platform **won't discriminate users based on the personal data they provide such as their race gender or other demographics that they have the same access to the data over the content of the social media platform like people of other preferences or aspects.** And that's how it can look like in the design in the reality of the design, how it can be implemented…all of these protocols with discrimination and the privacy and the data management has to be followed in that when designing such a social media platform. It's really important to have forums where users can consent …in reality can be done with forms , with registering and mainly not limiting the users to access of content based on individual aspects that they have." (Mary, W3)

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Collectively the objectives discussed in this last workshop included transparency, data ownership and security, upholding human dignity and preventing discrimination, avoiding personalised advertisement systems and technical robustness.

One of the difficulties raised by the team was the feasibility of achieving **individual and societal well-being** in the context of freedom of speech vs. hate speech/extremism or misinformation:

> “if the social media platform wants to allow freedom of speech and the control of the content will be done with some sort of automation, it can happen a lot that users have censored content or content is being erased or flagged as inappropriate as it's violating certain policies when in the end it might not. It might be completely legit, but because it's an automated process it's quite hard to detect if freedom of speech is being violated” (Mary, W3).

This difficulty was also cited as a reason as to why the specific value of individual and societal well-being was not prioritised in the previous exercises i.e. it was something not so simple or feasible to implement in practice.

3.3. Reflection exercise (in pairs of two or three)

One of the rationales of the reflection exercise was for the participants to be able to openly reflect on how the workshop impacted their way of thinking. During this exercise the participants were reminded of their most important professional value that they had originally identified and then they were asked whether this value changed at all after the group activity, or whether some other values that were discussed became more understandable or important to them.

From the analysis of the discussions we can see that sometimes the participants directly or explicitly thought it changed their views or that they thought about things differently as a result of the workshop.

For example, Christiana noted how the workshop, and the discussion questions related to the **environment**, had made her think about the environmental aspects as part of her everyday tasks in a way she had never thought of before:

> “…the environmental well-being was something I had previously thought of but not so much. So I think it was interesting that you raised it because it kind of puts things in a very different dimension I think, and it’s one you don't tend to think of in your everyday day-to-day tasks, despite increasing reliance on technology. So thanks, thanks for raising that. That was a good trigger, let’s say” (Christiana, W2)

The value related to the environment was something that also resonated with Jeremy. His original main value was transparency and although he admitted that won’t change much as he works in cybersecurity and it is rather significant in that field and for his specific project:

> “today's workshop really opened my eye to the importance of that **individual and social environmental being**, and that we need to assess the impact of our solutions on the society, and this is really, it really seems like the issue of today's systems. And yeah, this is really useful for me, thank you” (Jeremy, W3).

Similarly, Robert expressed how his original focus or priority value changed as a result of the workshop. For example, whereas originally responsibility and accountability were the priorities for him, now after

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

engaging with Scenario 2, he was "leaning more to **privacy** and **data governance** and **transparency**" because as he added he saw it now as **'the real issue'**:

> "That's the real issue of like using data and selling it of people. Be transparent and also **human agency and liberty** and give them a choice like we discussed before an try to **inform** them how the data is going to be used." (Robert, W2)

Another example of change was expressed by John, a participant who originally had presented privacy as his main value but after Scenario 2 he also noted how he would be giving 'some **extra value** to transparency':

> "yeah, now **I can see more clearly** why this is an issue. So if the user is more able to understand how their data are being processed and why they are seeing what they're seeing it's…it has the same importance with privacy. It clears out many issues of privacy. I think when we have transparency a lot of issues of privacy go away."

So the workshop not only enabled him to appreciate more the significance of transparency but also allowed him to see **more connections** with other values and their interdependencies, for instance with privacy.

Other times participants said that the discussions aligned with their own values, and they were 'happily pleasantly surprised' by this and that it was 'interesting' or useful to hear other peoples' perspectives but that it **did not change** anything in their thinking (George, W1).

> "I was really happily, pleasantly surprised. I thought everyone, it was really nice to hear everyone's perspective about it. And I think that sort of the kind of ethical or whatever, questions were quite to the front of everybody's thinking. And it didn't really change anything in my thinking, in terms of the discussion that happened today or, but it was nice to see that that was there. I think it does align with the values that I put down in my response" (George, W1).

Interestingly though, for example, one person who said this, in the previous discussion in the online google exercise with Group A and Group B, they had said several times in their responses that they were 'inspired' by a particular participant or that 'I wasn't thinking about it at the beginning' but then thought about it (this case 'standards') after a participant had said something. So it is important to consider that **engaging in group discussions during a workshop and hearing others perspectives and experiences may well trigger a change or a processing of new information** but without the participant being always conscious/aware of it.

Others also took this opportunity to express how some values that were discussed in the group **did not align** with their values. For example one participant disagreed with the position that "a tech company should protect vulnerable users" unless it was something more obvious "like not showing pornographic material to minors". As we can see in the dialogue that follows, his position was that as adults, people:

... engaging in group discussions during a workshop and hearing others perspectives and experiences may well trigger a change or a processing of new information ...

> "should be left free to make [their] choices. And so, I wouldn't, if I try to limit their ability because they have some personal issues or something like this, I would feel something like as a spiritual leader more than a tech company. I wouldn't put myself like making choices for other people."

To which the facilitator asked whether the participant would "prefer the individual to have more responsibility to protect themselves rather than structural institutions to promote this kind of culture of protection." The answer focused on the limits of the responsibility of the tech company itself.

> “I think like for example, a state should definitely look into protecting vulnerable people. **I don't think that's the purpose of a tech company**. Even though the technology should definitely preserve human rights and act ethically, I don't think it should, focus on, I don't know… Like it should give people the right to choose. I think there are two separate levels. I don't think a tech company should delve into the spiritual or political decision making.” (Thomas, W1)

The above discussion is related again to the theme of **responsibility**. It seems that in this case the participant despite having intense discussions with one other participant in the workshop, still felt strongly about the company being very limited in terms of its duties or responsibility to protect for example vulnerable groups or users in general – this according to him was the responsibility of the state, of the individual and of spiritual/political leaders. This again reflects a wider debate in terms of whether the users are educated enough in order to be able to protect themselves and whose duty it is to educate them. For this specific participant such issues were relegated to spiritual or political spheres and should be seen as separate to company issues.

4. Further discussion

While AI ethics is a blooming field, there has been little research conducted on how organisations and businesses integrate ethical practices or how AI practitioners negotiate/mediate ethical values and integrate these values in their workplace. There is even less research being conducted on what happens when the AI practitioner’s values clash with those of their organisation. Recently, we have witnessed a number of high-profile clashes between AI researchers and the organisations that they have work(ed) for, such as the much-publicised firing of Dr. Timnit Gebru and Dr. Margaret Mitchell[4], founders and leads of the AI ethics division at Google. From the discussion of the workshops (W3 in particular) the opinion that transparency within the tech industry is often not valued was highlighted, a point which was also reflected by Timnit Gebru (Tiku, 2020). AI companies may develop technologies that have specific functions, but how those technologies are used is not always clear. Organisations may not be transparent about how their AI will be used.

What became apparent during the course of the workshops was a strong emphasis to abide by what is legal. Many of the participants stated that their organisations only cared about what is legal, even if this explicitly contravenes what is (often, glaringly) ethical. This point can also be demonstrated in the Google case (Tiku, 2020), where Google was not doing anything illegal, however, the Google employees still deemed actions to be unethical.

A participant (W1) did mention that discussing the responsibility of AI organisations to protect vulnerable people from AI can be too vague and open to interpretation. AI companies cannot hope to prevent *all* possible misuses and impacts of their AI and that more specific examples need to be given. Of course, there is also a flipside: if AI organisations require very specific evidence about what kinds of violations will be caused by their AI, then there is room for them to claim ignorance. For instance, a company can state that they adamantly protect human rights and implement ethical guidelines in their development of AI and are simply uncertain how others may use their AI in practice. Clearly, companies will not have full knowledge about how *all* of their customers will use *all* of their AI, but there are many instances where claiming ignorance is not ethical.

[4] Both women had campaigned for more diversity at Google and were fired (https://www.bbc.com/news/technology-56135817)

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

In fact, there is no silver bullet for organisations to ensure that all of their AI will be used in an ethical way. Nevertheless, ensuring inclusiveness, co-development, and participatory decision-making (W2), may help minimise some of harms. Still, this should be legitimate and thoughtful inclusiveness, rather than being implemented solely for appearance or 'participation washing' (Ayling and Chapman, 2021; Sloane et al., 2020).

Within a democratic context, AI practitioners should be able to express their concerns or even challenge the ethical decisions of their organisation. To do so, it is crucial to have avenues or forums to discuss these concerns with the company and be able to initiate change if they identify injustices and harms taking place internally. However, as we have seen from the workshops and the Google example, this is not always possible in real-life where tensions between financial interests and moral values ensue. Often, change requires a degree of public shaming or controversy to initiate action.

Furthermore, there is not always one clear way that employees can initiate collective action within their companies. AI practitioners are often left uncertain about a clear-cut way for implementing change. While AI ethics guidelines and regulation are a good starting point for ensuring that AI is ethical (Jordan, 2019), AI practitioners should also feel empowered to implement these values and not simply follow the directions of the client or company, which could lead to unethical practices, as identified in (Orr and Davis, 2020). Empowerment is 'important because it ensures that people feel a greater sense of ownership in the solutions that they are building. They become more capable of solving problems that really matter to their users and customers' (Gupta, 2021). Management should implement 'training programs for ethical awareness, ethical conduct, and competent execution of ethical policies and procedures, and these programs should cover the ethical deployment and use of the system' (Brey et al., 2021, p. 72). This should 'encourage a common culture of responsibility, integrating both bottom-up and top-down approaches to ethical adherence' (Brey et al., 2021, p. 72). In fact, it can be seen that, often, AI organisations implement internal ethics boards, ethics committees, and ethics officers, to deal with these concerns and challenges inclusively and transparently (Stahl et al., 2021).

There also needs to be a certain level of independence and freedom to challenge the norms of the organisation, internally. Individuals within these organisations should be protected to conduct their research to ensure that the AI is developed and deployed in an ethical way. These organisations need to act on the feedback and advice from their employees, rather than simply using AI ethics teams and responsible AI groups as a façade (Lazzaro, 2021). As Timnit Gebru stated in a recent interview, without labour protection, whistle-blower protection and anti-discrimination laws, anything that AI ethicists do within large organisations 'is fundamentally going to be superficial, because the moment you push a little bit, the company's going to come down hard' (Bass, 2021).

Therefore, there should be accessible routes for the AI practitioner to follow, externally, if they feel their concerns are not being listened to. For example, establishing independent AI ethics ombudsmen to investigate these matters, AI ethics bodies (nationally or internationally) where the AI practitioner can follow-up about these issues, and an AI ethics whistleblowing group to allow the general public insights about the nefarious practices taking place with the organisation. There should be 'a process that enables employees to anonymously inform relevant external parties about unfairness, discrimination, and harmful bias, as a result of the system; that individual whistle-blowers are not harmed (physically, emotionally, or financially) as a result of their actions' (Brey et al., 2021, p. 44).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

4. Conclusion

*… if more **ethics education** and training is initiated within these organisations and provided to the AI practitioners, they may be more **willing, confident and able to implement them** in practice …*

This deliverable D5.6 addressed a study conducted within the impact acceleration project phase, which aimed to investigate AI education for professionals and in particular, examined whether an intervention with professionals in AI-related posts might be effective in promoting value-based design. As an extension of the SHERPA project, the study was based on experience gathered from previous tasks and deliverables of the project and reaches out to more stakeholders in an attempt to accelerate the impact of the SHERPA recommendations, in particular the recommendation for AI Ethics education and training.

One of the interesting findings was that many of the participants in the workshops placed an emphasis on what is legal because it was the easiest approach to take and for fear of getting in trouble. This may be a sign that individuals are more likely to follow ethical principles if they were concrete and presented as ethical codes of conduct, just as they feel safer and more comfortable to follow legal rules because these are provided in clearer and more concrete terms. Workshop participants were also often unsure about how to implement ethics in practice.

*… more research into the particular **educational competencies required for revaluation and rethinking of values** in the context of everyday work practices is necessary …*

One could provisionally conclude from this small study therefore, that if more ethics education and training were initiated within AI organisations and provided to the AI practitioners, they may be more willing, confident and able to implement ethical action in practice. Others, were used to focus on 'getting the job done' and doing this in a technically efficient way. However, AI practitioners, e.g. software designers and developers, can act as 'agents' of change, in the sense that they are made aware of the ethical impact of technology, and consider this in addition to technical efficiency. This entails rethinking the 'quick and dirty' mindset and prioritising digital and ethical well-being (Burr and Floridi 2020) above speed and absolute profit.

Finally, the study concluded that more research into the particular educational competencies required for revaluation and rethinking of values in the context of everyday work practices in AI development is necessary. Ethical values are indeed a prerequisite to implement ethical-oriented goals and change is not easy, but there is still optimism to see opportunities for change within the wider dynamics of the AI industry.

References

Ayling, J., Chapman, A., 2021. Putting AI ethics to work: are the tools fit for purpose? AI Ethics 1–25.

Badea, C., & Sherman, D. K. (2019). Self-affirmation and prejudice reduction: When and why?. Current Directions in Psychological Science, 28(1), 40-46.

Bass, D., 2021. Google’s Former AI Ethics Chief Has a Plan to Rethink Big Tech. Bloomberg.com.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101.

Brey, P., Lundgren, B., Macnish, K., Ryan, M., Andreou, Brooks, L., Jiya, T., Klar, R., Lanzareth, D., Maas, J., Oluoch, I., Stahl, B., 2021. D3.2 Guidelines for the development and the use of SIS. https://doi.org/10.21253/DMU.11316833.v3

Burr, C., and Floridi, L., (2020). The Ethics of Digital Well-Being: A Multidisciplinary Perspective, in Ethics of Digital Well-Being, A Multidisciplinary Approach. Editors C. Burr, and L. Floridi (Cham: Philosophical Studies Series), 1–29. doi:10.1007/978-3-030-50585-1_1

Čehajić-Clancy, S., Effron, D. A., Halperin, E., Liberman, V., & Ross, L. D. (2011). Affirmation, acknowledgment of in-group responsibility, group-based guilt, and support for reparative measures. Journal of personality and social psychology, 101(2), 256.

Crocker, J., Niiya, Y., & Mischkowski, D. (2008). Why does writing about important values reduce defensiveness? Self-affirmation and the role of positive other-directed feelings. Psychological science, 19(7), 740-747.

Frazier, L. D., Schwartz, B. L., & Metcalfe, J. (2021). The MAPS model of self-regulation: Integrating metacognition, agency, and possible selves. Metacognition and Learning, 1-22.

Gupta, A., 2021. How to build an AI ethics team at your organization? [WWW Document]. Medium. URL https://towardsdatascience.com/how-to-build-an-ai-ethics-team-at-your-organization-373823b03293 (accessed 10.5.21)

Jordan, S.R., 2019. Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI. Presented at the International Symposium on Technology and Society, Proceedings. https://doi.org/10.1109/ISTAS48451.2019.8937942

Lazzaro, S., 2021. Are AI ethics teams doomed to be a facade? Women who pioneered them weigh in. VentureBeat. URL https://venturebeat.com/2021/09/30/are-ai-ethics-teams-doomed-to-be-a-facade-the-women-who-pioneered-them-weigh-in/ (accessed 10.5.21)

Markus, H., & Nurius, P. (1986). Possible selves. American Psychologist, 41(9), 954–969.

Orr, W., Davis, J.L., 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. Inf. Commun. Soc. 23, 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Oyserman, D. (2007). Social identity and self-regulation. In A. W. Kruglanski & E. T. Higgins (Eds.), Social psychology: Handbook of basic principles (pp. 432–453). The Guilford Press.

Sloane, M., Moss, E., Awomolo, O., Forlano, L., 2020. Participation is not a design fix for machine learning. ArXiv Prepr. ArXiv200702423.

Stahl, B.C., Antoniou, J., Ryan, M., Macnish, K., Jiya, T., 2021. Organisational responses to the ethical issues of artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-021-01148-6

Tiku, N., 2020. Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it. Wash. Post.

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

5. Appendices

5.1. Appendix A – Information Sheet

Information Sheet

SHERPA - Shaping the ethical dimensions of smart information systems (SIS) – a European perspective

Please take some time to read this information and ask questions if anything is unclear.

Contact details can be found at the end of this document.

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

What is the purpose of this study?

The SHERPA project investigates, analyses and synthesises our understanding of the ways in which smart information systems (SIS; the combination of artificial intelligence and big data analytics) impact ethics and human rights issues. The project aims to develop novel ways of understanding and addressing SIS challenges.

The current study is part of the SHERPA project and is based on experience gathered from previous tasks and deliverables of the project. In particular, stakeholder input highlighted the importance of providing **educational material for AI** on all levels of education. SHERPA is in a unique position of having done much fact-finding work and developing mitigation proposals which can be used to enrich educational material. SHERPA plans to contribute to AI education by a) providing teaching-oriented scenarios building on real-life case studies and b) training provision based on SHERPA-related outcomes underpinned by relevant theoretical perspectives. The specific study will focus on value-based design among professional IT officers/designers/developers.

Who is organising this research?

The research for this study is being undertaken by the EU-funded SHERPA project (SHERPA is the acronym for 'Shaping the ethical dimensions of information technologies – a European perspective' (https://www.project-sherpa.eu). A Research Ethics Committee has reviewed and approved this research.

Who is funding the research?

This research is funded by the European Commission's Horizon 2020 Research and Innovation Programme under grant agreement no. 786641.

Who is selected to take part in the study?

The project participants are professional IT officers/designers/developers or work alongside such team of professionals (e.g. ICT research, information security management).

Do I have to take part?

Participation in this study is voluntary and you may ask any questions before agreeing to participate. If you agree to participate, you will be asked to sign a consent form. However, at any time, you are free to withdraw from the study and if you choose to withdraw, you will not be asked to provide any reason for doing so.

What will I be asked to do if I agree to take part?

If you agree to take part in this study, you will participate in a workshop (3 sessions) involving group design tasks and discussions. The first session of the workshop (approximately 1 hour) will involve your input regarding an application/platform/algorithm. You will be given a short questionnaire and you will be also asked to write about your professional priorities and standards. The second session (approximately 3 hours) involves working on a scenario in groups of 3-5 people. You will also be asked to discuss the process both in pairs and in the larger group. The final session is similar to the first one (approximately 1 hour) and will include an individual interview. All sessions will be recorded and transcribed for analysis. The study will take place online at a time convenient for the participants.

What are the possible benefits of participating?

The study aims to improve your understanding of value-based design and develop the necessary skills. In addition to contributing to the process of finding answers to the research questions posed by the SHERPA project, you may personally and professionally benefit from the scenarios and the discussions taking place during the workshop. A certificate of participation will be provided at the end of the workshop to each participant.

What are the possible risks of participating?

There are no risks in taking part in this study. At any time during the research you can choose to withdraw. You may also choose to withdraw your data from being used in the project at any time until 1st July 2021.

How will the research data be used?

The data collected from the research study will be analysed by the SHERPA researchers. The recording of the discussions may be transcribed by parties outside of the consortium. If this happens, the transcription company will delete the recording and transcription after the transcription is approved. On the consent form we will ask you to confirm that you are happy for the SHERPA consortium to use and quote from your interview. Any such use will be anonymous unless you indicate otherwise on the consent form. Information which will identify your organisation will also be kept out of publications unless otherwise indicated on the consent form.

What will happen to the results of the project?

All the information that we collect about you during the course of the research will be kept strictly confidential. You will not be identified in any reports or publications and your name and other personal information will be anonymised unless you indicate otherwise on the consent form.

What happens to the data collected during the study?

The discussions from all three sessions will be transcribed by the interviewers or a designated, approved third-party agency. If we use a third-party transcription service, we will ensure that there is a signed data processing agreement in place. The audio files will be deleted, once the analysis of the data is complete. The transcriptions as well as the answers given to the short questionnaires will be analysed by the SHERPA researchers.

How can I access the research findings?

You may request a summary of the research findings by contacting Kalypso Iordanou, University of Central Lancashire Cyprus (KIordanou@uclan.ac.uk).

What about use of the data in future research?

If you agree to participate in this project, the research may be used by other researchers and regulatory authorities for future research. The transcript will be kept for five years after the publication of the findings of the study.

What should I do if I have any concerns or complaints?

If you have any concerns about the project, please speak to the researcher, who should acknowledge your concerns within ten (10) working days and give you an indication of how your concern will be addressed. If you remain unhappy or wish to make a formal complaint, please contact Dr Stephanie Shaelou, SLaulhe-Shaelou@uclan.ac.uk

Fair Processing Statement

The information collected will be processed in accordance with the provisions of the EU *General Data Protection Regulation* (*GDPR*)

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

5.2. Appendix B – Consent Form

Project SHERPA – Consent form

Statement	**Respondent's initials**
I have read the information presented in the information sheet.	
I have had the opportunity to ask any questions related to this study, and received satisfactory answers to my questions, and any additional details I wanted.	
I am also aware that excerpts from the three research sessions may be included in publications to come from this research. Quotations will be kept anonymous unless I give specific permission to the contrary (below).	
I give permission for my name to be associated with excerpts from the recorded transcripts which may be included in publications to come from this research.	
I give permission for my organisation to be identified in any final publications produced by SHERPA.	
I give permission for the three sessions (including the interview) to be recorded using audio recording equipment (if necessary).	
I understand that relevant sections of the data collected during the study may be looked at by individuals from or a project partner from SHERPA. I give permission for these individuals to have access to my responses.	
I understand that the audio recording may be given to a transcription service company to transcribe. I give permission for these organisations to have access to my audio files for transcription purposes.	

With full knowledge of all foregoing, I agree to participate in this study.

I agree to being contacted again by the researchers if my responses give rise to interesting findings or cross references.

□ No □ Yes

If yes, my preferred method of being contacted is:

□ Telephone: …………………………………………………..

□ Email: ………………………………………………………….

□ Other: …………………………………………………………..

Participant Name		Consent taken by	
Participant Signature		Signature	
Date		Date	

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

5.2. Appendix C – Discussion questions and Scenario

1. Why do you think there is disagreement about the use of advertisements as a source of income from this new platform?

2. Who should ultimately be responsible for making the decision about whether advertisements will be used?

3. Who should ultimately be responsible for making the decision about face recognition or job recommendation features are developed?

4. What are some potential vulnerabilities from developing face recognition or job recommendation features?

5. Consider the scenario that each one of the two different platform implementations are used widely within a community. What positive and negative societal impacts do you foresee?

6. Who should be held responsible if personal data is leaked in any of these situations? [accountability]

7. What possible implications regarding environmental sustainability could the features of this platform have? Is this an important aspect for you? Why? Do you think it should be addressed more in future work? Why/why not? [environmental sustainability]

8. What possible implications regarding human rights and liberties could the features of this platform have? Is this an important aspect for you? Why? Do you think it should be addressed more in future work? Why/why not? [human agency and human rights/diversity and fairness/inclusion and social justice]

9. What possible implications regarding transparency could the features of this platform have? Is this an important aspect for you? Why? Do you think it should be addressed more in future work? Why/why not? [transparency]

Scenario 2: A problem has come up. A new social media platform is going to be developed (similar to Facebook/Twitter/Linkedin).

Some of the developers (Group A) of the platform support that the platform should be freely accessible to the public and have advertisements as their source of revenue. To use advertisements the team will employ AI and Big Data to allow for automated social media posts, and optimisation of social media campaigns for the advertisers. AI and Big Data, also referred to as Smart Information Systems (SIS), will create an SIS that will be able to figure out what works best using advanced analytics, and also decode trends across social media to find the best target audience for each product. To do this an SIS-based social media monitoring mechanism will be developed. As a secondary feature of the platform, the developers wouldn't mind using SIS to also create some interesting features for the users at a later stage, however their main focus for the initial product is the use of SIS for smart advertising.

Other developers (Group B) within the company support that there should be a registration fee for users, for covering the revenue of the company, with no advertisements. They still feel that AI and Big Data should be used only to provide more services to the users and that the users should be able to at least consent to data collection, e.g. by actively selecting a specific service. For example, the developers will offer options to the users to use some features of the social media platform that are SIS-powered, such as face recognition, the platform will use AI and Big Data to recognise the users face in photos and based on that provide filter options, etc. Another feature of the social platform will be the option for companies to advertise their job posts, and AI and Big Data will be used to create a service to match the platform users with potential jobs. In both these examples, the user will be able to control whether they would like to have the use of filters or job recommendations as part of their profile.

The developers basically agree on the use of AI and Big Data but disagree on the emphasis they should place on using SIS for improving user-centered features, and they also disagree on the use of SIS for advertising.

Consider that this assignment has been assigned to your team. Work in your team to clarify the key design and development objectives of this task.