The Playbook on participation and accountability in City Challenges

Adapted for the AI HOKS project of City of Helsinki Education Division

Version 1, 9.08.2021

Prepared by:
Karolina Drobotowicz[1]
Vera Vidal
Marta Ziosi
Yung Au
Kaivalya Rawal
Giulia Schneider

As a part of: Summer 2021 AI Policy Research Clinic of The Berkman Klein Center for Internet & Society at Harvard University

Icons used in this document are by Freepik from Flaticon

[1] project leader, inquires about the playbook can be sent to: drobotowicz.karolina@aalto.fi

1. Introduction

This playbook presents the results of working group 2 during the Berkman Klein Center For Internet & Society research clinic in July 2021. The main task was to help the city of Helsinki with the design and deployment of Learning Analytics in its Education Department, and more specifically the AIHOKS tool currently under development. Within this task, the specific goals of our work were as following:

Goal 1: Assist the City of Helsinki's Education Division to create an inclusive, participatory, and sustainable strategy for stakeholder engagement throughout the design, development, deployment, and assessment stages of new classroom technologies.

Goal 2: Assist the City of Helsinki's Education Division to develop accountability mechanisms and practical implementation strategies.

1.1 Motivation

This playbook is thought of as a companion to the "Method for tackling City Challenges", by Coboi Lab (in this document referred to as **the worksheet**). Supported by the Diputació de Barcelona, an actor at the forefront of digital democracy, the Coboi Lab has already implemented the method in the city of Sant Boi de Llobregat, to successfully respond to specific city challenges. The fact that the method has been designed as transversal and transferable makes it suitable to be applied to other city challenges with a specific focus on collaboration and multi-actor participation. We believe that is the case for the City of Helsinki and this is why we elected it.

On one hand, the suggested method is fitting to Goal 1 as it takes a holistic approach to participation, by structurally plugging it into the design, development, deployment, and assessment phases. On the other hand, it lends itself to Goal 2 as it sets these phases in a loop, where closing this loop entails developing practical implementation strategies, such as shared evaluation measures, to ensure accountability.

1.2 What to expect

In this document, we personalized the tools and processes suggested by the Coboi Lab worksheet to best suit the context of the City of Helsinki and AIHOKS tool. Along with the document, this has been achieved by presenting a four-stage process where the motivations and introductions to each phase have been tailored to the information and the resources that have been provided by the Helsinki team. These are complemented by links to the frameworks from Coboi Lab worksheet and our recommendations for specific activities based on our knowledge and what we find important for this case.

Within this process, the elements of participation and accountability ought to be thought of holistically, and are to be integrated throughout. However, this method also envisages steps that are specifically aimed at ensuring participation and accountability. In the case of **participation**, this step is the Co-creation workshop. This workshop entails a collective process of co-education, solution ideation & validation, and the creation of shared impact indicators to monitor and evaluate the impact of the system on all of the actors involved. In the case of **accountability**, this step entails Closing the loop from the creation of shared impact indicators to the Evaluation and Communication of the project to the public. Accountability is here specifically understood as "Human Oversight". In line with the HLEG Ethics Guidelines for Trustworthy AI, in this playbook, "Human Oversight" will be approached on three different levels; Human In The Loop (HITL), Human On The Loop (HOTL), and Human In Command (HIC).

Throughout this playbook, there are text boxes with Scenarios and Parables about AI, smart cities, and policies from the real world which we hope will help concertize ethical issues and pitfalls of these types of projects. We have compiled these and some further examples in Appendix C (The Scenarios & Parables Worksheet) to help illustrate stories around the loop. This can be used to onboard team members, used in workshops to think through these issues with participants, and used as an educational resource to inform the public about the tricky landscape of AI and smart cities.

1.3 Playbook Outline

The outline of this playbook is structured so that the Introduction is followed by the Participation Process section. This section elaborates on the process which we propose and it stands as a key to the reading of each phase of the process. Subsequently, the Observation, Exploration, Action Plan, and Implementation phases are presented. These are followed by specific sections on the strategy for Closing the Loop. The main part of the playbook ends with a section on Limitations. The appendices with the Co-creation Workshop and Scenarios map are added.

2. The guide to the process

The proposed participation process consists of four main phases:

1. The Observation phase to identify the problem the city wants to tackle and analyse the city's capacity to solve it collaboratively;
2. The Exploration phase to identify and integrate the perspectives from the different actors involved and define collectively a vision for the future;
3. The Action Plan phase where specific actions are devised and designed collaboratively;
4. The Implementation phase to design and test the prototypes of the prioritised actions, and monitor the results.

Each of the phases consists of actions of three levels of priorities:

- essential presents the actions that we find necessary for completing the participation cycle,
- recommended parts can enhance the understanding of the problem and participation of stakeholders,
- possible are the activities that can be helpful if the team hadn't done those already or if they don't have organization-specific frameworks for them.

Before starting the actions we invite you to list out your constraints and resources: how much time, money and people's work time can you spend on this loop? What kind of skills, frameworks and activities do you have? Regarding people skills, we encourage you to onboard one person experienced in participation techniques (eg. service designer) to help you in shaping and leading the participation process.

Once you have completed the above, choose the actions you will have time to accomplish. For your convenience we included the time estimates for each activity, however, beware that those don't include the individual research and preparation time. Two shapes also indicate who is taking part in the actions: the rounded rectangle is for the lead team (so Helsinki team with a facilitator) and the parallelogram is for actions done with the impact team with chosen stakeholders (see the Exploration phase for more information). Once you choose your path, follow the colour coding in the playbook for chosen activities or check the page numbers in the figure.

If successful, this process can oversee that the introduction of Learning Analytics in the city of Helsinki stems from a problem that is defined collectively with and which is accountable towards the actors most affected by its future implementation. This is why we stress the importance of including students within the process, and more specifically students from vulnerable groups. We understand that the city is already in the last

phase. Nevertheless, in the spirit of iteration, we invite you to revisit the first, second, and third phases to ensure the success of the strategy.

Importantly, whatever action the city will take should be guided by a why and a who question: why is this thesis a “good idea” and who decides this thesis is a ”good idea”. Only a full multi-stakeholder participatory process from the very beginning of defining the why, the problem being tackled can provide relevant answers.

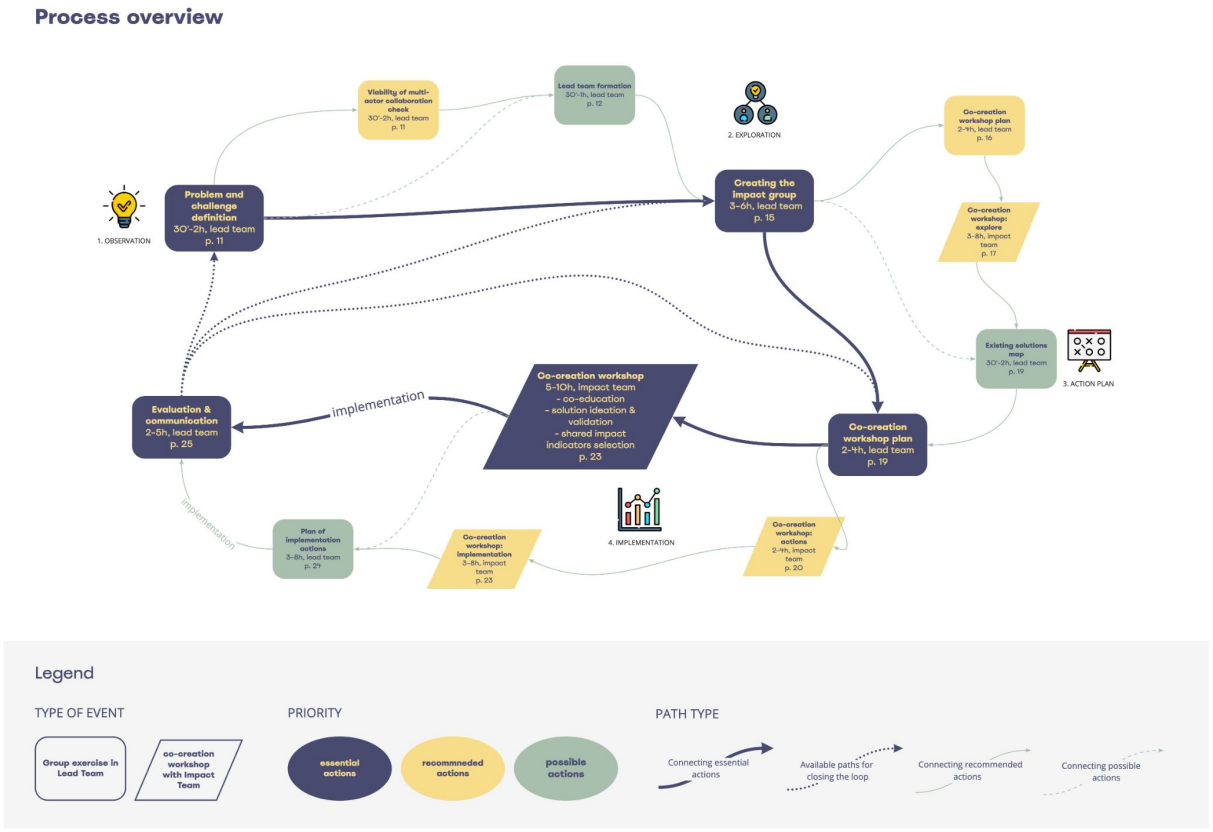


Fig. 1. The process overview - see appendix A for a better quality picture

2.1 Observation phase

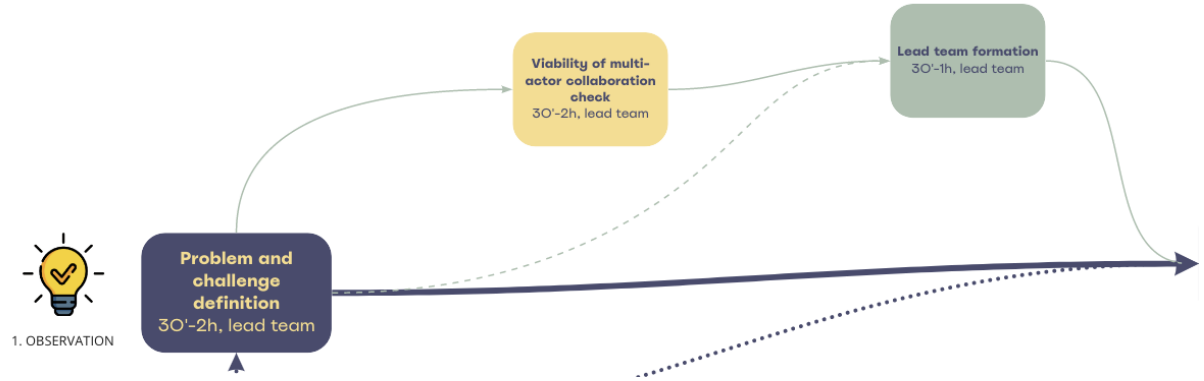


Fig. 2. The observation phase overview.

What is it about?

Before designing and developing any innovative project, it's important to identify and analyze the problem or need for change you want to address in the area of interest, be it social, cultural, economic, or environmental. The Observation phase will help you understand the problem and reformulate it as a city challenge that fits within the strategic vision for the city.

In this specific case, the Observation phase may help to identify a need for change, such as a high rate of student drop-out, in the area of interest of education within the city of Helsinki. The challenge will consist in turning the need for change into an opportunity for action, which in this case may be the devising of the AIHOKS tool.

In the spirit of participatory design, this phase would ideally include a step where the viability of multi-actor collaboration is assessed. The inclusion of a variety of actors willing to help find solutions to the challenge increases the capacity for action and its legitimacy in the eyes of citizens. This is key in achieving one of the Helsinki team's main goals: trust.

Once the collaboration among multiple actors is deemed viable, the next crucial step would be to form a Lead team to drive the subsequent phases of the process forward. Whether and how these two steps are carried out varies depending on factors such as funding, time, and at which stage of development the project is. The next lines will elaborate on that.

Overall, the Observation phase aims to set the stage for the design of new actions that can bring you closer to the desired change in the city, and within its educational ecosystem, in a participatory and yet actionable manner.

[text adapted from the worksheet pp. 16-17]

Clearly (re)defining the problem and city challenge Essential

Why? The goal of this step is to identify and understand the problem. By analyzing the relevant aspects of the problem, such as its incidence, the main actors affected, associated cultural and social factors, and the sense of urgency for a solution, we can start reflecting on the desired change. We can then define the city challenge we plan to tackle collaboratively. [text adapted from the worksheet pp. 20-21]

Who? The core team who has been working on the pilot

How long? 30'-2h.

Worksheet pp. 20-25

Our recommendations: This initial step is an opportunity to revisit internally the challenge the AI-education tool is aiming to tackle and integrate the learnings from the pilot into the understanding of the problem and the challenge. Guiding questions could be:

- What is the main problem we are addressing? Is it that of drop-out, or antiquated pedagogy innovation?
- What knowledge do we have on it based on studies, discussions with teachers and students, our pilot?
- What knowledge are we still missing?
- How is the AI-HOKS tool fitting with the wider strategy at the Helsinki level to deal with this problem?

The lead team can either move towards the creation of the impact group (essential route) or towards assessing the viability of a multi-actor collaboration (recommended route).

Assessing the viability of multi-actor collaboration Recommended

Why? To tackle a complex city challenge, a multi-actor collaboration enables to incorporate diverse perspectives when considering what actions to take. Coordinating efforts, experiences and learning helps develop innovative and systemic solutions. Nonetheless, it is a complex process. To ensure the viability of the collaboration, it is useful to verify which actors are affected by the challenge and how their level of interdependence and cohesion and their ability and willingness to work together. Three pre-conditions are key to the viability of the process: a sense of urgency for change, the existence of an influential champion, and the availability of adequate resources. [text adapted from the worksheet pp. 26-27]

Who? The core team who has been working on the pilot

How long? 30'-2h

Worksheet pp. 26-31

Our recommendations: This activity can help clarify some potential challenges of the collaboration.

⇨ The lead team can either move towards the creation of the impact group (essential route) or updating the lead team (possible route).

Updating the lead team Possible

Why? The role of the lead team is to facilitate alignment and collaboration between the multiple actors involved and guide the collective work process. It is made of people from the stakeholder organizations involved from the beginning of the process. Its composition can be updated at any time. They will be in charge of facilitating the creation of a collective vision, coordinating decisions of what actions to take, and ensuring the monitoring of those actions among other things. Roles and responsibilities assigned to the different members can evolve over time. [text adapted from the worksheet pp. 32-33]

Who? The core team who has been working on the pilot

How long? 30'-2h

Worksheet: pp. 32-39

Our recommendations: It could be interesting to revisit the composition of the lead team. This could mean integrating (an) expert(s) in Design and participatory processes, with complementary expertise to the one already in the team. This could be for instance a member of the Helsinki Lab specialized in participatory processes or someone involved in Ruuti, to benefit from their experience working with youth aged 13-17 on participatory processes. This could also entail bringing on board someone in the Education department with complementary knowledge about the challenge to be tackled. Whether or not the team composition changes, it could be worth reassessing who takes which role and responsibilities.

⇨ The lead team should move to the creation of the impact group (essential route).

Scenario: Dreaming up New Cities

Sidewalk Toronto was an ambitious smart city initiative in Canada headed by Sidewalk Labs (owned by Alphabet) and Waterfront Toronto, that was ultimately dropped after

two and a half years and after a commitment of USD 50 million. It aimed to revitalize what used to be an industrial port into a smart micro-city that would provide eco-friendly, low-emission housing alongside other promises.

There are many reasons for its untimely end, including the disruptions from the COVID-19 pandemic, but also what appeared to be a lack of trust, transparency and agreement with the community at hand. One of the particularly contentious issues was that it aimed to tackle environmental problems through the fine-grain and minute monitoring of daily life. This included occupancy sensors in every home to inform smarter temperature and energy usage; smart street lamps to adjust for optimal lighting usage, and smart cameras to analyze traffic patterns. Through the process, privacy concerns were dismissed, channels of communication with the public broke down, where eventually four people resigned from the project's advisory board due to various disagreements. Amongst the project's harsher critics, the endeavour was called "dystopian" (Roger McNamee), "an experiment in surveillance capitalism" (BlackBerry founder Jim Balsillie), and an example of where 'community' became used as a branding technique" (Spencer Wicks)

Wicks (2020) in a deep analysis of the project argued that there was a lack of responsiveness from planners to the issue raised by participants; where development goals, methods, and success metrics were not adapted to initial feedback from participants which appeared to only have symbolic value. The public appeared to be excluded; from the decision to partner with Waterfront Toronto to the priorities of the project.

How do we ensure smart projects are truly collaborative? How do we define who our communities are, how they will be heard, and how we co-produce with them?

2.2 Exploration phase

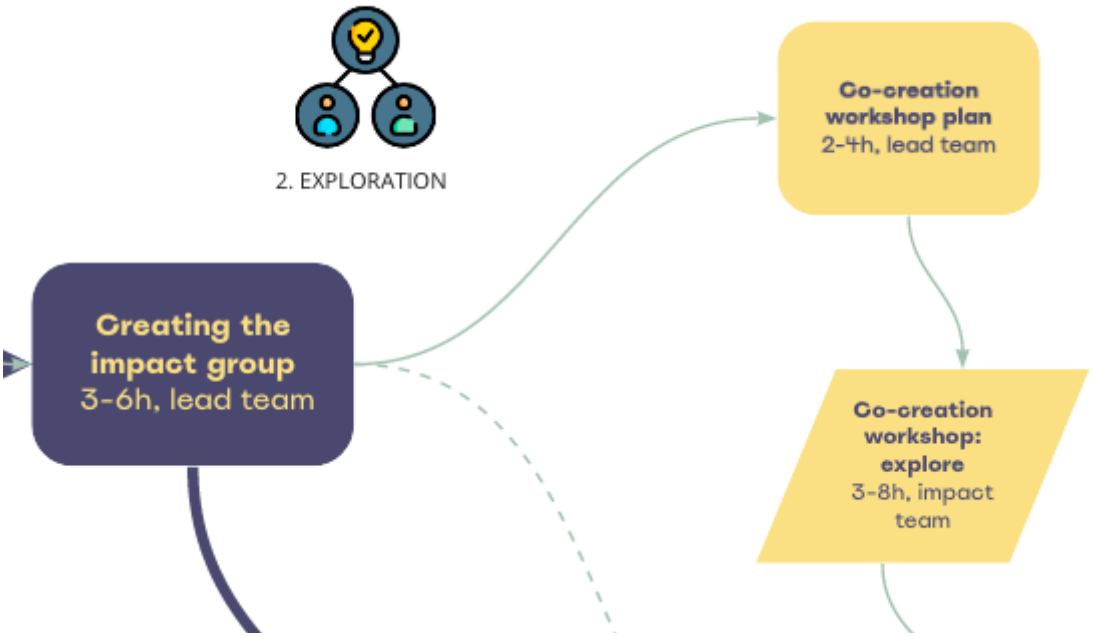


Fig. 3. The Exploration phase overview.

What is it about?

The Exploration phase seeks to provide the knowledge needed to define a shared vision for the city challenge, here eg. bringing an AI innovation to the education sector. City challenges are often complex and affect different actors in different ways. To address them, it is important to initiate collaborative processes with the actors affected or involved with the challenge in one way or another.

In this phase, the Lead team begins by identifying the relevant actors in the city challenge to later be able to convene and include them in the Impact group. The impact group has a crucial role. Its members have first-hand experience of the problem and, therefore, should play a central role in co-creating solutions that can bring about the desired change for the city. Depending on time and funding, the process of creating the impact group may be complemented by the planning and the execution of a co-creation workshop to explore the challenge and to create a shared vision of the future in collaboration with the impact group.

The ultimate aim of the Exploration phase is for the Lead team to incorporate multiple interests, priorities, and perspectives into the process to create a mutual diagnosis of the baseline situation and arrive at a shared vision of the desired future scenario.

[text adapted from the worksheet pp. 44-45]

How should we proceed?

Creating the impact group Essential

Why? The Impact group is comprised of the stakeholders with a vested interest in the challenge, who will work with the Lead team to solve the problem and devise actions. To create the Impact group, the lead team should first identify which actors are relevant to the challenge and analyze how to best integrate them based on their potential commitment and resources, which could range from participation in the co-creation process to keeping them informed or being observers for instance. Based on this prioritization of actors, the Lead team can determine who should be in the Impact group and how to best integrate them depending on their needs and characteristics. [text adapted from the worksheet pp. 48,49, 58 & 59]

Who? The Lead team

How long? 3-6h

Worksheet pp. 48-63

Our recommendations :

- Some of the key stakeholders regarding such an educational challenge would count among others: the students and teachers who will be using the tool, the school administrators where the tool is deployed, parents of the students, youth psychologists, the companies where vocational school students do their work placement. They may also be civic actors such as NGOs and local education activists, the city and the government educational division and potentially social services, service providers for the tool as well as experts in the realm of education and AI.
- Consider working with collectives, NGOs, and Unions here. There are various difficulties with identifying, recruiting, and incentivizing participants in these processes - and there might be particular barriers for minorities. If there are collectives who specialize in minority rights and other issues, it might be worth consulting them as they would have better links to local communities as well as a better understanding of historical contexts, barriers, and problems (an NGO and caseworkers who have worked with refugees and asylum seekers might have a better idea of the difficulties of datafication in an educational setting for example; likewise a Disability Collective might have better ideas on how AI might enforce neurotypical worldviews). To do so with young people, possible ways could be by working with schools, working with student unions, or with the Helsinki Youth Council for instance. To reach out to civil society organizations potentially interested, putting out a call could be a way of reaching out.

- To follow the rules of inclusive design and design justice, it's important to prioritize the voices of those most affected. If the challenge tackled is that of drop-out, groups most at risk are: students with advanced age at enrolment, female students, students with learning difficulties, students with low GPA in basic education certificate, students with parents with lower socioeconomic status (see Vehkasalo, 2020). Due to the Covid pandemic, there was a noticeable increase in dropouts among older students (ages 18 to 24) at Helsinki's vocational schools.
- It's also essential to think of compensation for members of the impact group for their time and expertise. Sometimes, organizations would be able to use their resources in this collaboration; though actors such as teachers, activists, students should be compensated with salary or school credits for students.
- Besides the suggested worksheet, other tools exist to help identify the actors such as ecosystem mapping.

The lead team can either move towards planning the main workshop (essential route) or planning an exploration workshop (recommended).

Co-creation workshop plan for the exploration phase

Recommended

Why? To design and structure the workshop precisely and coherently we must consider the participants' profiles and needs. The workshop helps to formalize the impact group by integrating, sharing, and recognizing the different perspectives and interests of the actors involved. It also helps to generate new relationships between them and start formalizing a shared vision for the future. [text adapted from the worksheet pp. 64-65]

Who? The lead team

How long? 2 - 4h

Worksheet pp. 64 - 71

Our recommendations: This exercise might be done mostly by a workshop facilitator or two if possible. The lead team should provide what type of resources are possible for the workshop (compensation for participants, place, time) and can review after the completed workshop plan.

The lead team should move towards the exploration workshop (recommended).

Co-creation workshop: explore Recommended

Why? To gain an in-depth understanding of the challenge, validate and reformulate it. To do so the workshop explores the causes and effects of the challenge as well as the opportunities, obstacles, and levers for change. It also helps create a shared vision for the future, that is agreed in a final target scenario for the challenge. [text adapted from the worksheet pp. 72, 73, 78, 79, 84 & 85]

Who? The lead team and impact group

How long? 4-8h

Worksheet pp. 72-89

Our recommendations: This exercise is being done in a shorter form in the main co-creation workshop that we introduce in Appendix B. We recommend doing this exercise first and then skipping the repetition in the main workshop if there are resources available. To map opportunities and obstacles, tools such as a SWOT or a PESTEL analysis can prove helpful too.

The lead team can either move towards planning the main workshop (essential route) in which case they should only focus on the part after the co-education phase or mapping the existing solutions (possible).

Scenario: Disability

In a report, AI Now warns that much of AI technologies targeted at disabled people "implicitly promises to make them more like non-disabled people". They gave the example of an AI-enabled app called Ava which promises to allow Deaf people to take part in spoken conversations, converting the spoken word into text in real-time. The premise here is that deafness is a hindrance to communication; however, they cite disability activist and scholar Eli Clare who wrote: "many deaf people claim themselves not as disabled but as a linguistic minority. They locate the trouble they experience not in their inability to hear but in the non-deaf world's unwillingness to learn and use sign language." Similar patterns of assuming needs and using neurotypical expectations can be seen across technologies made for disabled people – for instance, AI-enabled technologies that promise to coach autistic people to make eye contact and display emotions.

What assumptions are built into your systems?

2.3 Action Plan phase

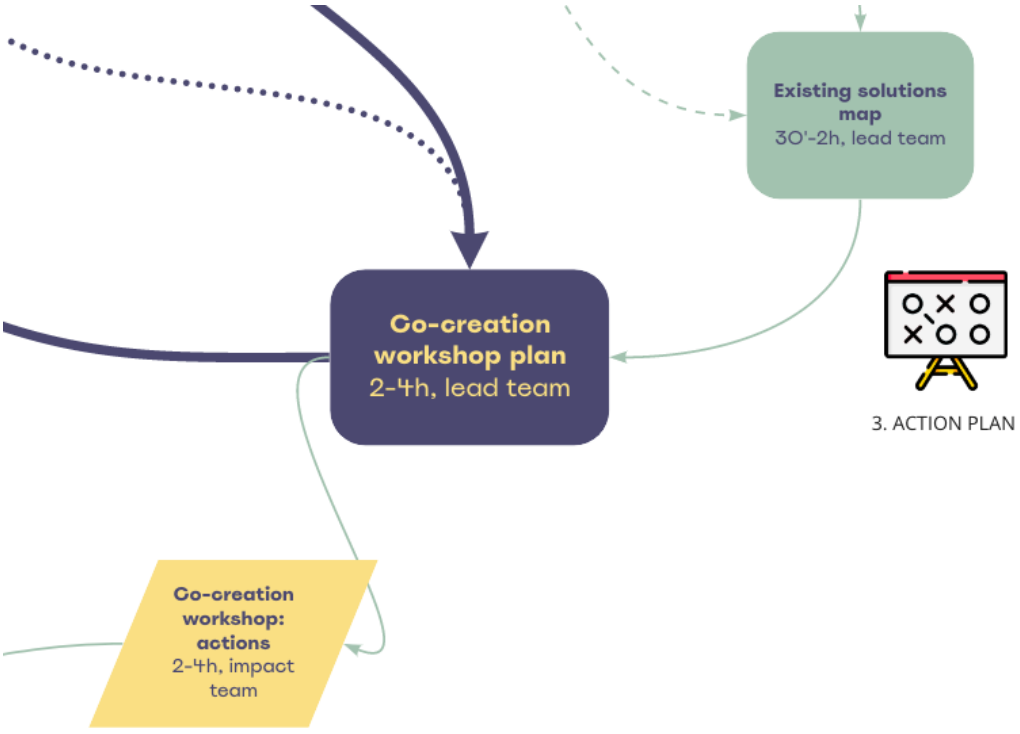


Fig. 4. The action plan phase overview.

What is it about?

In the Action Plan phase, all the information and knowledge generated during the previous phases should be used to set the stage for the collaborative work and to jointly propose viable lines of action consistent with the shared vision of the future for the target area of education in the city of Helsinki.

In case this has not been done by this stage, the Lead team must map out existing solutions to the challenge that it is trying to tackle, to make sure not to reinvent the wheel.

In the recommended path, the Lead team will plan on what will be the main co-creation workshop to convene the Impact group to integrate their different visions, to come up with new solutions, and to eventually create a mechanism for collective evaluation of the proposed solution.

The goal of the Action Plan phase is to set the stage for the main co-creation workshop, which represents the main element of the Implementation phase.

[text adapted from the worksheet pp. 98-99]

How should we proceed?

Mapping existing solutions Possible

Why? To identify key elements from similar challenges in other contexts that could be replicated or adapted, asking questions such as: What has been tried before? Which solutions were successful and why? Which initiatives failed and why? What are the main lessons learned? Who could give us more information? [text adapted from the worksheet pp. 100-101]

Who? the lead team

How long? 30'-2h

Worksheet: pp. 100-105

Our recommendations: Make a list of people or organizations associated with these solutions to contact them and learn from their experience.

The lead team can move towards planning the main workshop (essential route).

Co-creation workshop plan Essential

Why? To design and structure the workshop precisely and coherently we must consider the participants' profiles and needs. The workshop helps to formalize the impact group by integrating, sharing, and recognizing the different perspectives and interests of the actors involved. It also helps to generate new relationships between them and start formalizing a shared vision for the future. It will also help us devise new actions to work on the AI tools for education for Helsinki. [text adapted from the worksheet pp. 64-65]

Who? The lead team

How long? 2 - 4h

Worksheet pp. 64 - 71

Our recommendations:

- In the essential route, this will lead to the co-creation workshop introduced in the Implementation stage and described in detail in Appendix B. If there are resources, we recommend following the recommended path too, and perform one workshop at the Action Plan stage and one at the Implementation stage.
- Document your process so far to be able to explain the choices you have made about the model of the tool, the proxies and assumptions, the data used to train it, and the lesson from the pilot.

- This exercise might be done mostly by a workshop facilitator or two if possible. The lead team should provide what type of resources are possible for the workshop (compensation for participants, place, time) and can review after the completed workshop plan.

The lead team can move towards conducting the main workshop (essential route) or co-creation workshop: actions (recommended route).

Co-creation workshop: actions Recommended

Why? To identify new actions, establish criteria to prioritize the most significant and viable ones by assessing the feasibility of their implementation and capacity for impact and analyze the resources and capacities required to implement them. It can help generate a theory of change of the actions' desired impact in the short and long term which will then help define future monitoring and evaluation indicators. This can be particularly helpful when constructing from scratch an AI tool. [text adapted from the worksheet pp. 108, 109, 114, 115, 120 & 121]

Who? The lead team and the Impact group

How long? 2-4h

Worksheet: pp. 108-125

Our recommendations: When prioritizing actions, the group should pay special attention to which impact groups are put forward and what harms are associated with them, and how to take into account the legal framework, to avoid violation of the law. So to prioritize impact groups and related harms, a balancing approach is needed for consistency with the relevant legal framework that should be operationalized in the action phase.

The lead team can move towards conducting the main workshop (essential route) or co-creation workshop: implementation (recommended route).

Scenario: Data Request

The system has begun its roll out and there are considerable safeguards by now to protect personal data. E.g., the smart city authorities have decided that different sectors will not share data with each other and that only certain data is collected, such as attendance, basic personal information, and data on social interaction students have with their peers.

However, halfway through the school term – a crime has been committed. The police and intelligence agency have started requesting data on students who are potential suspects and witnesses. What are the guidelines for this request in protecting data, especially for minors? What level of evidence is needed for a data request and how much say do individuals and institutions have on this? What level of data granularity is needed for the AI system used here (is more data collected than necessary)? How will you get feedback on this process and are there complaints procedures in place in case someone has been treated unjustly? How can this contribute back to the iterative feedback loop to improve the system and its safeguard where the legal landscape of this will likely constantly shift?

Civilian intelligence legislation and Military intelligence legislation (June 2019); enables "civilian and military intelligence agencies to acquire information on alleged threats to national security through communications surveillance, without any requirement for a link to a specific criminal offence."

Civilian intelligence protects Finland's national security - Ministry of the Interior (intermin.fi)
Everything you need to know about human rights in Finland| Amnesty International | Amnesty International
Civilian intelligence protects Finland's national security - Ministry of the Interior (intermin.fi)

2.4 Implementation phase

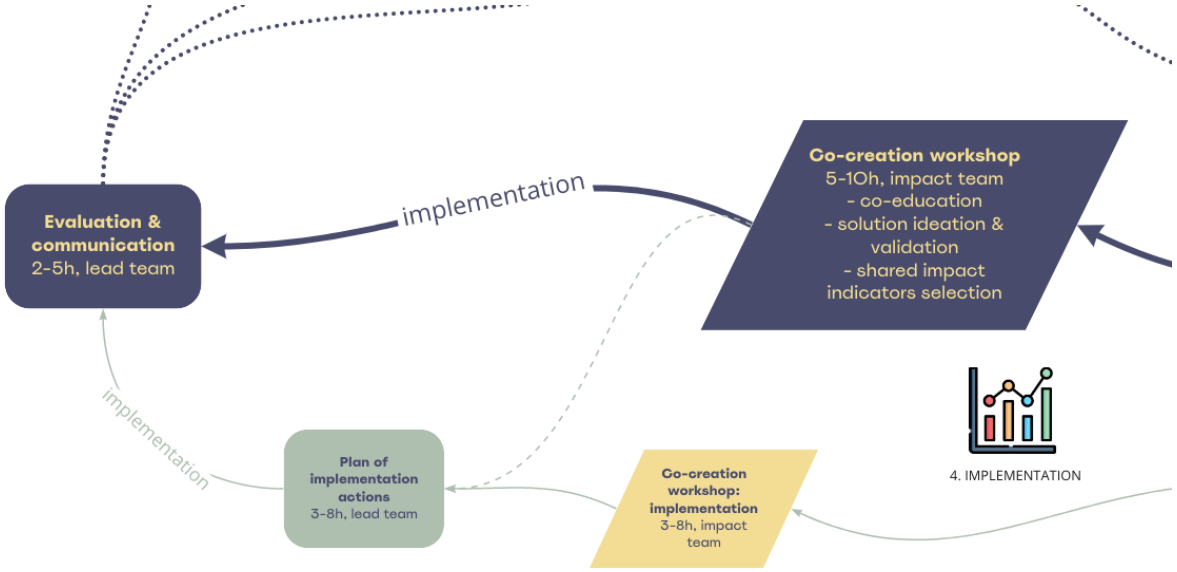


Fig. 5. The implementation phase overview.

What is it about?

The Implementation phase consists in collaboratively ideating and validating a solution and in coming up with ways to evaluate its impact that reflect the needs and vision of the Impact team. The goal is the implementation of the proposed solution, followed by its evaluation and its communication to the wider public.

This phase features the main co-creation workshop. The co-creation workshop aims to involve the Impact team in an exercise of co-education to each member's positional needs and knowledge, in an exercise of solution ideation and validation, and eventually, in the creation of shared impact indicators that reflect the success criteria of the envisioned solution. This latter step is essential for observing and measuring the progress towards the solution during its implementation and evaluating its final impact in both the short and medium term.

In this stage, the Lead team needs to identify the elements required to implement the proposed solution and plan some key aspects, such as the specific activities, the assigned budget, responsibilities, and the roadmap to be followed to achieve the expected results. Those can be done by following the possible path or utilizing inside-organization resources. Once these elements have been defined in more detail, a common narrative should be articulated to communicate the purpose and added value of the solution and raise awareness among the Impact group and all those involved and/or affected by it. In the essential path, the narrative exercise is included in the main workshop.

It should be noted that as well as evaluating the final results of the solution, it will be important to also carry out intermediate evaluations during the implementation process to decide whether you need to reorientate the focus of the actions implemented or even repeat any activities from the previous phases. This will be further elaborated on in the Closing the Loop section.

[text adapted from the worksheet pp. 134-135]

How should we proceed?

Co-creation workshop Essential

Why? To complete all the essential actions with the impact group in the scenario, where resources are allowing for only one participatory workshop in the loop.

Who? The lead team and the Impact group

How long? 5 - 10h

Details: Appendix B: the main co-creation workshop.

Our recommendations: In case the recommended path is followed, we suggest adding Part 3: How do we make the tool to the Co-creation workshop: implementation, as this is the AI-specific part of the workshop.

The lead team can move towards Evaluation & communication (essential route) or Plan of implementation actions (possible route).

Co-creation workshop: implementation Recommended

Why? The purpose of this workshop is to define the final actions that have to do with building/updating the prototype. It will help test and validate the assumptions of the model, incorporate improvements and generate a common representation for the Impact group of what the tool will be like as well as generate a common narrative. It will also help define how to monitor the implementation of the tool to receive feedback on its progress and analyse critically the results. It will help define shared indicators that reflect the success criteria. This will allow you to guide the process on which relevant data to gather and ensure all actors are on the same page when it comes to evaluating the results and impact. [text adapted from the worksheet pp. 136, 137, 148 & 149]

Who? The Lead team and the Impact group

How long? 4-8h

Worksheet: pp. 136 - 155

Our recommendations: We suggest adding Part 3: How do we make the tool to the Co-creation workshop: implementation, as this is the AI-specific part of the workshop.

The lead team can move towards Implementation or Plan of implementation actions (possible route).

Plan of implementation Possible

Why? To define the elements to be implemented, identify the requirements in terms of resources, capacities, and alliances needed for the implementation. The different associated tasks will then be scheduled and prioritized by defining responsibilities, the timeframe, and budget and then establishing a roadmap. [text adapted from the worksheet pp. 156-157]

Who? The lead team and the Impact group

How long? 3-8h

Worksheet: pp. 156 - 169

Our recommendations: The worksheet proposes the actions that can be taken together with the impact group. If the resources permit, we recommend inviting the impact group for planning together. We understand, however, that the City of Helsinki might have its resources and methodologies of listing the responsibilities and creating the roadmap, and those tasks, in that case, might not be relevant.

The lead team should move towards Implementation.

Implemententation Essential

Why? To implement the actions planned in previous steps.

Who? The lead team and all people involved in developing the tool.

Our recommendations: We recommend working iteratively, that is, to make small changes and updates, test, evaluate the outcome and close the loop from the chosen point. The list of the actions to be completed should be prioritized by the outputs from co-creation workshops with the impact group.

The lead team should move towards Evaluation and communication.

Evaluation and communication Essential

Why? Analyzing data from the shared indicators will help to perform a critical analysis of the final results. It is also a moment to share learnings on the effectiveness of the implementation of the tool and its medium-term impact. This will help improve the tool for further iterations. It is a communal and transparent step that can help reinforce the sense of commitment and co-responsibility of the different stakeholders. [text adapted from the worksheet pp. 172-173]

Who? The lead team and the impact group

Worksheet: pp 172 - 177

Our recommendations :

- In the worksheet, the evaluation and communication is a separate 1-4 hour-long activity for the lead and impact group (pp. 172-177). The activity also refers to the shared indicators chosen with the impact group.
- In our playbook, we included more detailed legal and technical frameworks that you can use for human oversight in the Closing the loop and Mitigation Technologies and Processes sections.
- It is good to have a control group to assess the effectiveness of the tool.
- It's essential to keep stakeholders (therein users) informed about the evaluation results. Think of a platform for sharing those (can AI register serve for it?).
- Think also of having some repetitive evaluation meetings with the impact group. At first, there should be more often (eg. every month), later they can get more rare (eg. every half a year). During those meetings, you should present the evaluation results and how those were obtained. You can also decide then with a group what should be the next step.

⇨ The lead team should move towards the oversight actions, described in the section Closing the loop; and then follow with the next iteration from a chosen point in the path.

Scenario: Automated A-Levels

In 2020, Foxglove together with students forced the UK government to make a U-turn on a controversial A-level algorithm system. Amid the COVID19 pandemic, the government had put in place a system to predict final year A-level student grades instead of in-person examinations. This predicted grade would determine decisions including whether or not a student was accepted to their University offers. The system required teachers to estimate how well they thought their students would perform, then they adjusted these grades using an algorithm that weighted scores based on the historic performance of the particular high school the students went to. This was based on the idea that this

weighting would compensate for the tendency of teachers to inflate the expected performance of their students. Of course, this also meant that students with high grades from less–advantaged schools were most likely to have their scores downgraded, while students from richer schools were more likely to have their scores raised. Throughout this process, there was no clear consulting processes nor were there straightforward complaints mechanisms. The government eventually scrapped this algorithm.

What could have been done in the observation, exploration, action plan, and implementation stages to have prevented this controversy?

3. Closing the loop

This is the part of the process dedicated to ensuring accountability. In this playbook, we have decided to focus on accountability as "Human Oversight". By borrowing the division from the Ethics Guidelines for Trustworthy AI, we analyze the concept of Human Oversight on three different levels; Human In The Loop (HITL), Human On The Loop (HOTL), and Human In Command (HIC). While these three are not mutually exclusive, they may entail different interventions at different phases of the process to uphold their requirements.

Even though we have split the lifecycle of an ML project into 4 phases, we do not believe this to be a linear development but instead an iterative process where human oversight happens by developers continuously looping through the 4 phases. We refer to this process as "Closing the Loop". As part of closing the loop, retraining machine learning models is a common industry practice, although there is currently no standard set of guidelines detailing when and why this should occur. Instead of this being merely a technical decision, however, **model retraining should also be an ethical decision, focusing on the holistic behaviours of the model**.

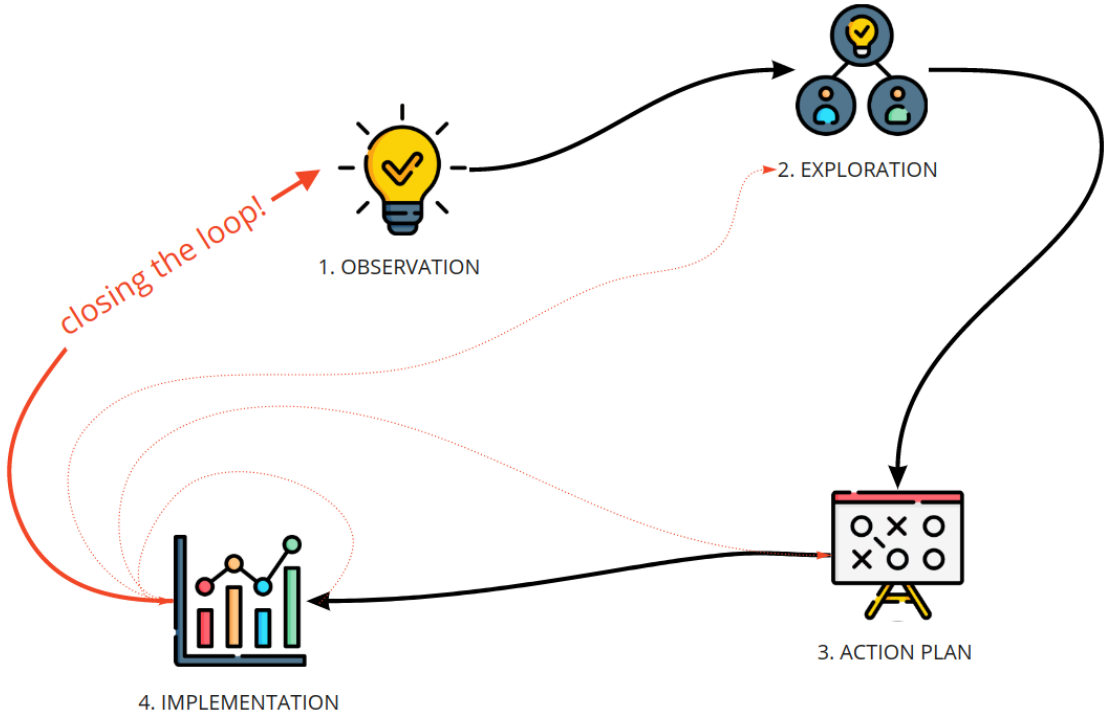


Fig. 6. The simplified process overview with the closing the loop focus.

Several subjects can be involved in the human oversight process, which we envisage as a multilayered and collaborative process among different stakeholders to the project, in consistency with the pyramid below.

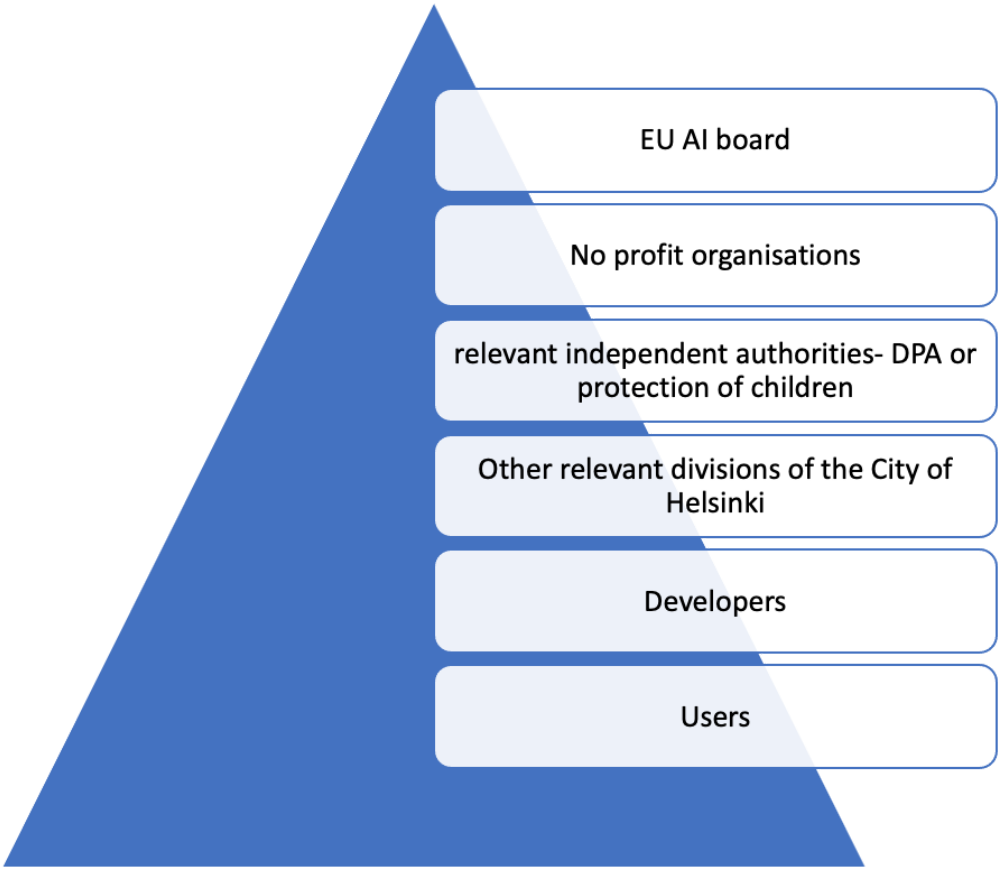


Fig. 7. The pyramid of stakeholders that might be involved in the oversight.

The process of ensuring adequate human oversight is a continual process in which there is no one size fits all. In this section we provide some starting points and prompts to think about this in the project, including legal recommendations and technical recommendations.

3.1 Why close the loop?

Closing the loop is especially critical because of the impact that deploying such a system has on the data and behaviours it is trying to model: the AI system builds a model of the world which it uses to make predictions, but its very existence and predictions intervene on the world, causing it to change. This makes the system's model go stale, and thus requires frequent updating. It would be prudent to assume that the very deployment of this system will result in a change in student behaviour, and this new training data needs to be fed into the predictive system (following the participatory 4 phase process detailed in our playbook) for it to function well. This is why the three levels of human oversight are very important to assure that impact groups retain full control of how the

model functions in practice and enable the developer to govern the model in case it works in an unintended way.

The three levels of human oversight are enshrined in the European Commission High-Level Expert Group on AI's Guidelines on a Trustworthy AI and are related to I) human in the loop: the capability for human intervention in every decision cycle of the system; II) human on the loop: the capability for human intervention during the design cycle of the system and monitoring the system's operation; III) human in command: the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation.

Parable: Snakes & Loops

In colonial India, the British Raj decided to control the local population of snakes by instituting a bounty to be paid out in return for dead snakes. The idea was simple and effective - pay people to kill snakes, and gradually the snake population will reduce. This worked, for a time. However, the policy was never updated to reflect the new reality of gradually reducing (but still high) snake populations. The set-and-forget approach was adopted, and once instituted and verified to be working, the policy was left untouched, without any updates to reflect the new state of the world.

This policy directly intervened in the world, and the lack of a close-the-loop approach led to disastrous consequences. As the snake population reduced in urban centres, it became easier to breed domesticated snakes, rather than to go out into the jungle and kill wild ones. Thus, entire communities were transformed into snake breeders, and growing and harvesting snakes became financially viable due to the new government policy.

When the government finally realized what was happening and withdrew the bounty, the snake breeders simply let all the snakes loose, which had the effect of increasing the overall snake population dramatically. Thus, not closing the loop led to the system behaving in a manner that was exactly the opposite of what was intended!

source: Pervasive incentive - The original cobra effect at Wikipedia

3.2 Human in the Loop Oversight

"The capability for human intervention in every decision cycle of the system"

Regularly test the assumptions of the model

To assure that the artificial intelligence-based model efficiently achieves the set goals, the assumptions of the AI model must be regularly tested to realign the model's

assumptions with the changed reality in which it intervenes. For human-in-the-loop oversight, the adequate enactment of set normative conditions requiring interaction between the developer and the user is essential. We suggest that this enactment should occur following a participatory approach maximizing the set interaction to the benefit of full protection of the relevant fundamental rights that are illustrated in the chart below.

Problems	**Principles**
Transparency	Legal Rules (12-15 GDPR; Technical documentation - Record-Keeping; Provision of information to users- Reg. AI) Ethical Principles- EC Trustworthy AI
Explainability	Legal Rules Art. 13-15 GDPR ("meaningful information) Technical- Explainable AI is a technical option
Contestability/ Objectability	Legal Rule Art. 21 GDPR/art. 22 GDPR Ethical principles Trustworthy AI
Human Supervision	Legal Rules- Art. 14 AI Regulation Ethical Principles-Trustworthy AI
Self-determination (Standardisation of Cultural Models)	*Fundamental Right to Dignity and Autonomy* *ECFR*
Freedom of Expression	*Fundamental Right ECFR*
Institutional Autonomy	*Constitutional Right*
Digital Divides - nondiscrimination	*Fundamental Right ECFR*
Privacy Policies- Terms and Conditions - Clear and understandable	
Data quality and management	*Legal Rules- Art. 9 Reg. AI* *Ethical principles- inclusivity-Diversity.*

Define the list of responsibilities-also in terms of legal obligations

The responsibilities for correct implementation of the project shall be strictly adherent to the envisaged risks and harms. They should be also adequately measured and defined upon the applicable legal framework providing for specific obligations. Itis thus extremely important to reconstruct the relevant framework, giving specific

consideration to the General Data Protection Regulation, the Cybersecurity Act as well as the EC Proposal on AI. Also, because the project is developed by a public administration, it is important to consider applicable open data regulations, for the enactment of requested open data policies.

In general, the project will have to pay due attention to the procedures and/or systems implemented for data collection, storage, protection, retention, and destruction, ensuring that they comply with national and EU legislation on the subject. In particular, the interaction between the General Data Protection Regulation (GDPR) and the other mentioned regulations should be taken into account. For these purposes, we think that guidelines and the protocols specifically designed for the management of the project should be developed as well as made readily public on institutional channels of the City of Helsinki.

Moreover, a protocol on the retention of data together with a procedure for disposal of original records which could identify data subjects should be enacted.

Given the high risk posed by envisaged processing activities that involve also the treatment of special categories of personal data under art. 9 GDPR, a Data Protection Impact Assessment (DPIA) should be conducted under Article 35 of the General Data Protection Regulation 2016/679, and a policy on incidental findings should be elaborated if needed as part of a more general ethical assessment possibly performed by an ad-hoc ethics committee.

General Data Protection Regulation (Reg. UE 679/2016)

Sensitive personal data management

GDPR considers both personal data and “special categories of personal data” that by their nature are particularly sensitive concerning fundamental rights and freedoms. Under art. 9 GDPR, these special categories of personal data (sensitive data according to the previous definition) encompass amongst others i) biometric data and, more generally, ii) “data concerning health”; iii) data related to racial or ethnic origin, political opinions, religious or philosophical beliefs. Education-related data often contains this type of sensitive information. It must be borne in mind that the processing of sensitive data is, as a default, prohibited (art. 9(1) GDPR) unless one of them (rather broad) exceptions apply (art. 9(2) GDPR) providing a suitable legal basis. A suitable legal basis for the processing of these data in the context of digital education performances could be found in art. 9(2)(d) GDPR related to processing endeavours in the context of “legitimate activities” or in art. 9(2)(g) GDPR related to reasons of “substantial public interest” as can be education. This appears the most adherent legal basis for education-related processing activities, but it comes along with the requirement for

data controllers to “provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject”. A first safeguard is surely given by anonymisation or pseudonymisation techniques, but additional safeguards, such as the enactment of data segregation, should be more generally defined within a code of conduct.

Transparency and Explainability
Arts 12-15 and 22 GDPR require the provision of meaningful information regarding the logic involved in the processing. As many commentators have suggested these rules require the explainability of enacted automatised processing models: for the digital education context these requirements should be implemented following the differentiated explainability model, which takes into account the different information needs of different impact groups. These provisions are important for assuring the human in the loop phase.

Contestability/Objectability
Following art. 21 GDPR the results of the processing activities have to be objectable and contestable, this relating to the possibility of the model to correct and rectify its results. Also, the implementation of this moment guarantees a human-in-the-loop oversight.

Cybersecurity Act (Reg. UE 881/2019)
The adoption of cybersecurity certifications in accordance with the Cybersecurity Act (art. 8) could be a relevant means for adherence to developed standards in the field of education technology.

3.3 Human on the Loop Oversight

“The capability for human intervention during the design cycle of the system and monitoring the system’s operation”

Model retraining is a common feature of industrial AI deployments. Model performance decays over time, and retraining helps keep AI systems healthy. In this sense, our recommendation of “closing the loop” is not new at all. However, our proposition of closing the loop not by returning to the start of the implementation phase, but by (optionally, depending on the resources available) returning to an earlier phase of development is unique. This modification of typical iterations in data science projects is important because of the following reasons:

- it builds trust in the community, rather than a set-and-forget strategy (which often results in regret), periodic engagement keeps systems healthy, relevant, and valid

- It allows for greater system adaptability, ensuring that the project can evolve and succeed with ever-changing demands and mutating contexts.
- It directly addresses the impacts that deploying the system itself has on the joint data distribution fed into the training data of the model

EC Proposal for an AI Act (COM/2021/206 final)
Although not yet in force, the AI ACT offers significant relevant tools for transparency and accountability purposes, related to the requirements under art. 9, requiring the generation of documentation of risk management systems; art. 10 related to the documentation of data governance and management systems; art. 11 regarding technical documentation of the enacted system; art. 12 concerning the enactment of automatic recording logs. These rules assure a human on the loop oversight.

Implement constant monitoring mechanisms for early detection and risk mitigation
The mentioned measures identified in the AI Act could well serve monitoring purposes if the results of collected documentation are made publicly accessible by open access policies.

3.4 Human in Command Oversight

"The capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation".

Codes of Conducts
Art. 40(2) recommends the enactment of codes of conduct to be drafted together with relevant bodies of expertise that can be entrusted to conduct mandatory monitoring of compliance. The enactment of such a code of conduct and the identification of relevant bodies of expertise outside the group of the City of Helsinki in charge of the development of the project is an extremely important moment for assuring the human in command oversight.

Open Data Directive (Dir. UE 1024/2019)
The legal framework of access regimes regarding publicly held data greatly relies on open access policies. Accordingly, the Open Data Directive (Dir. EU 2019/1024) places particular emphasis on the value of access and transferability of research data, in consistency with the paradigm of open science and innovation, aiming at fostering the interaction between research results and market innovation: art. 10 of the Directive requires publicly funded data to be openly available through open access policies. The

definition of open access policies regarding collected education-related data could facilitate oversight over enacted data collection and processing activities by identified impact groups, first of all, students and teachers. The enactment of these open data policies can assure human in command control.

Good practice: Metadicim

Metadicim or how to collectively govern a city's digital infrastructure and platform (see with Johanna Laukkanen for Helsinki's experience with the Metadecidim community).

Decidim - 'we decide' in Catalan - is a digital common's infrastructure, a framework, as well as an open-source digital platform for citizen participation made with free software. It is known for its open and horizontal governance model with a strong community behind it. Created by the city of Barcelona, it is now an association to guarantee the project's autonomy, transparency and democratic organizing. The community controls the domain and trademark, signs agreements for the platform's dissemination and development, and oversees the development and management of the platform. The city of Barcelona remains its main funder and co-participant in the product development governance.

Metadecidim is the democratic community that manages the Decidim project in all its dimensions, from software development to community engagement. The community is open to any person or institution that wants to participate in the debate, the proposal and the execution of relevant issues of the Decidim platform such as the (re)design of the features, the improvement projects and their uses and future possibilities.

Metadecidim - is organized into multiple spaces and channels. The meta.decidim.org website is organized along four pillars:

- Participatory Processes: They include a support forum where the members of the Community can ask questions, share tutorials and interact with other members to exchange knowledge. Other participatory processes include reporting bugs, getting familiar with the platform with tutorials and a demo and proposing new features.
- Assemblies: The Community meetings or SOMs are open, reflexive and collaborative spaces of participation to share experiences of the use of Decidim, prioritize development lines, to give support to new features, to solve doubts and empower citizens to appropriate the tool, to contribute to its development and to be co-participants in its construction. The SOM meetings enable to repeatedly engage the community with the platform and

foster an ongoing process of co-creation and innovation, allowing for transnational and cross-scale participation. Different working groups have been created to work along specific lines: governance, technology, processes, participant experience, communication, and a verification committee among others. Other assemblies include the yearly general assembly of the association and the LAB Metadecidim is an open and collaborative research space aimed at addressing key issues for the development of the Decidim platform and nline democracy.

- Consultations are used to ask questions regarding features, governance, etc. to all the people from the community, make a call to participate in the consultation, spark and order the debate for or against a response. When the consultation date arrives, the community can vote and publish the results of the votes.
- Conference: DecidimFEST is the annual meeting of the Decidim project and Metadecidim. Its objectives are: 1) public presentation of the project, of the last version of the software and the advances in the development of the platform, 2) celebration of work sessions, hackathons and specialized conferences, 3) promoting a space to share experiences with other cities and organizations to rethink in an open, transparent and collaborative way the future of Decidim.

4. Mitigation Technologies and Processes

4.1 Eval Stores

A model monitoring system, also termed evaluation store, should help "close-the-loop" as mentioned in the previous system. This is a tool that can either be built in-house, procured externally, or even created by adapting existing open-source software. Allowing with monitoring for system health though (like a typical model observability platform), emphasis should also be given to features that go beyond traditional monitoring metrics. These could involve:

- The ability to detect data and concept drifts: as real-world statistical distributions change, being able to detect them offers **clear early-warning signals about when model performance is about to degrade**. Extra vigilance at this stage can be much more helpful than trying to mitigate harmful model predictions after the fact
- The ability to explain model predictions: the eval store should come with added functionality that allows users to drill down and understand why models behave the way they do. If we predict a drop in fluency for a student, for example, the teacher assigned to review the case and decide whether an intervention is necessary would benefit greatly with an additional **understanding of the reasons behind the model prediction**, in addition to the prediction itself.
- Monitoring fairness violations: the eval store should monitor model performance not just in terms of traditional metrics such as accuracy or F1-scores, but should also **continuously display fairness metrics** (like CDD) and flag potential violations here.
- Flexible alerting: model observability platforms often come with built-in alerting abilities, which allow the flagging of anomalous behaviour. This should be customizable, with the Helsinki team being able to use its understanding of the local context and its experience in the problem domain, along with recommendations from ethics and education experts to **customize the quantities being alerted on by the eval store**.

4.2 Transparency

A good way to build transparent systems is to ensure the exploration and model validation phases (phases 2, and 3, as per our taxonomy) contain model cards that are accessible to a wide range of audiences. In addition to this industry standard, we propose the following additions (which might be thought of as features of an extended eval store):

- Model Explainability: we propose that developers use interpretable ML techniques wherever possible. This is a prerequisite for robust, safe, and fair systems (Rudin et al). This would allow for better human-in-the-loop oversight as well, better equipping teachers to act on model recommendations about interventions for students at risk of dropping out.
- Recourse: we propose that model predictions be non-binding and that counterfactual recourse be explored as a means to ensure both fairness and accountability in system predictions. This could be a simple recommendation that could be used both to aid human-in-the-loop decision-makers, as well as to guide the proposed interventions for students identified to be at high risk of dropping out. For example, the AI system would not just flag that a student's fluency is low, but would also simultaneously provide recommendations about how to improve it. (Wachter et al).

5. Limitations

5.1 The Limits of Participation

This playbook has been shaped extensively by the directives provided by the city of Helsinki and the Berkman Klein Center. It focuses almost exclusively on building a participatory framework and allowing for stakeholder oversight involved in a public sector AI undertaking. However, these do not form the complete picture when considering the ethical pitfalls of delivering a large-scale AI project. It is perhaps best to consider these per the Anna Karenina principle: these are necessary aspects for such an endeavour to succeed, but they are not sufficient. We do not explore alternative points of failure in this playbook, nor do we justify that participation and oversight are the most important pieces of the puzzle. This direction is informed not by our experiences or expertise, but rather due to the directives involved in the charter of forming this research clinic.

Participation and accountability are only a little piece of the ethical puzzle that an AI project applied to society ought to grapple with. The city of Helsinki should try to involve stakeholders in the model development and deployment process, but be constantly aware that this is no guarantee against broader ethical pitfalls. The only way to know whether the developed AI system functions well is to check if it is functioning well *after* it has been deployed, and to iterate on the process frequently until it is right. While doing that, it is important to assess its ethical impact more broadly in terms of privacy, autonomy, explainability, transparency, objectability, bias and discrimination, data quality, etc. Several ethical frameworks can be referred to practically uphold these principles. The main example is the Ethics Guidelines for Trustworthy AI and the assessment list provided therein.

Parable: Coca-Cola

In 1985, in response to shrinking market share, the Coca-Cola company decided to replace the original, secret Coke recipe with a new one. There had been growing competition from Pepsi-Cola, which had recently surpassed Coke in supermarket sales, and was outperforming coke in blind taste tests.

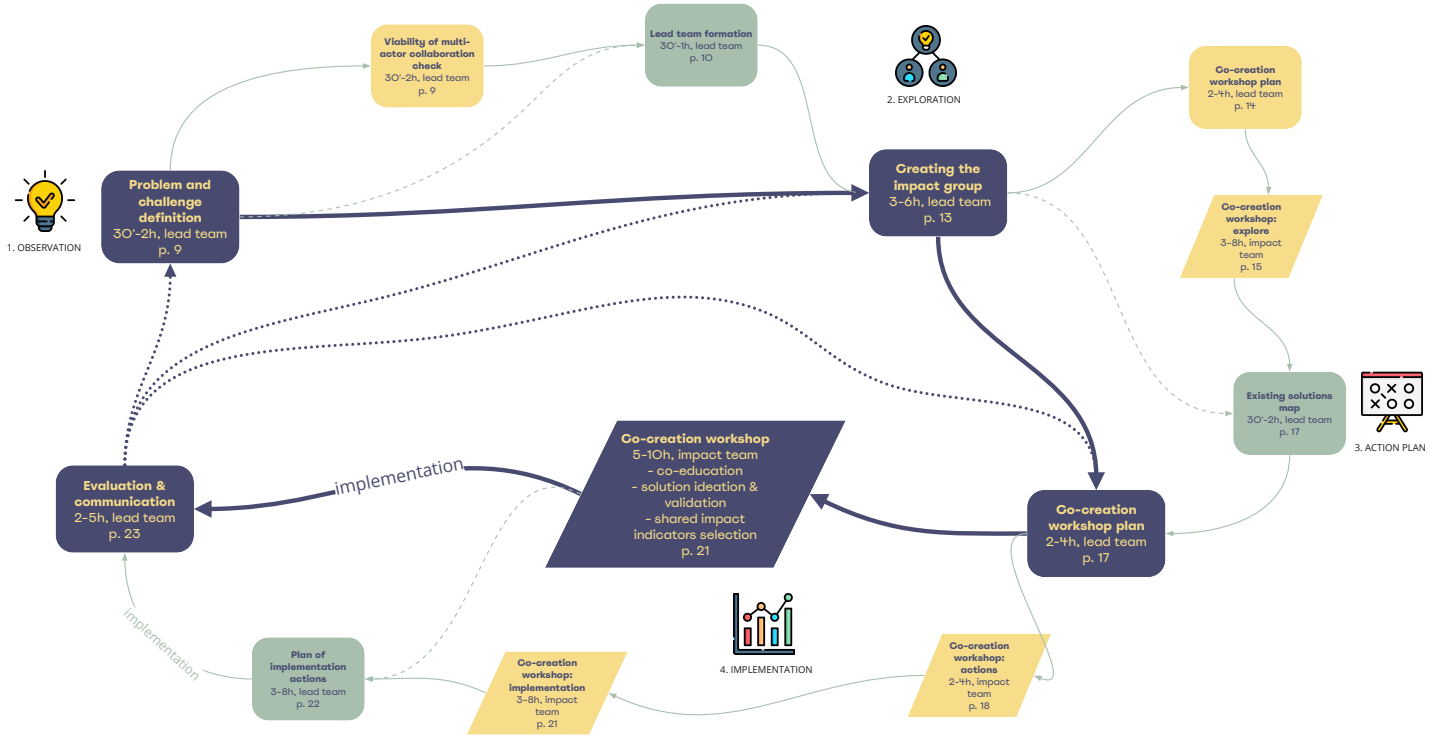
To sell soft drinks, blind taste tests can be thought of as the ultimate participatory process. If conducted well, and across a representative cross-section of the potential consumer market, they provide a clear view into the potential market success of a soft-drink product. Coke created a new product, finally changing its famous secret recipe, that was finally able to outperform Pepsi in blind taste tests. This was a robust participatory process too, with tens of thousands of sample taste tests. Coke was finally sure that its new product would not lose to Pepsi in blind taste tests.

> Upon launch of the product, not only did new Coke fail to impress consumers, but the backlash was so harsh that Coca-Cola had to bring back original Coke, and eventually retire the new recipe entirely. This resulted in massive financial losses, despite all the best practices being followed, and massive user testing before product rollout.
>
> *source: New Coke at Wikipedia*

Even though useful, such ethical frameworks might fall into the danger of becoming convenient check-boxes that certify the ethics of a project without ensuring it in practice. When dealing with AI Ethics principles, of which accountability, diversity, and non-discrimination are part, it is crucial to appreciate their contribution without uncritically relying on it. There are a host of concerns associated with relying on principles from AI Ethics or Fairness, Accountability and Transparency (FAccT) to solve societal problems and the Helsinki team ought to be aware of them. More information about it can be found in Critiquing and Rethinking Accountability, Fairness and Transparency (CRAFT), a new program based in the ACM FAccT conference and works such as EngageMedia: A critical view of AI ethics: Looking at the substance of ethical guidelines | Association for Progressive Communications (apc.org) and Ethical AI principles won't solve a human rights crisis | Amnesty International.

Finally, it is important to acknowledge that the brevity of the Policy Clinic detracts from the comprehensiveness of this document. This piece of advice has been gathered in the short span of a month. Additionally, the advice could not be informed by interviews with stakeholders nor by a deeper knowledge of the contextual constraints of this project because of the limited time available. Rather than disqualifying the work, this consideration warns against relying on this document as the only piece of advice. We invite the Helsinki team to continue gathering information from as many diverse sources as possible and we wish them luck.

Appendix A: Process overview



Legend

TYPE OF EVENT

Group exercise in Lead Team

co-creation workshop with Impact Team

PRIORITY

essential actions

recommneded actions

possible actions

PATH TYPE

Connecting essential actions

Available paths for closing the loop

Connecting recommended actions

Connecting possible actions

Appendix B: Co-creation workshop plan

In this section, we introduce you to our idea for the co-creation workshop with the impact group. For the essential path, but also to contextualize the aspect of participation, we decided to merge the Coboi Lab tools with our ideas and knowledge-based approaches. The co-creation workshop is specifically designed for a product that is already past the initial ideation and design stage (so no ideation for new solutions is included)[2]. The main goals of this workshop are:

- establishing the relationships among the members of the impact group,
- co-educating among participants on the problems that are to be solved and AI solution,
- reframing the product for the next implementation (therein data, model, processes...),
- selecting the shared indicators for evaluation.

This is the main participatory part of the process, in which you include the impact group selected in the Exploration phase. The workshop is a place for all stakeholders to share their opinions, ideas, concerns and needs. By doing so, you lower the chances of creating something harmful for users (and non-users), start building relationships between actors that will help with the regular evaluation and monitoring of the tool and increase the chances for success for all.

In case of following the recommended path, you might notice that some of the parts of the following workshop were already covered or will be in other workshops. Therefore, we suggest you skip most of the parts, but the Part 3: How do we make the tool?. This exercise, we encourage to add to the Co-creation workshop: actions or Co-creation workshop: implementation. The Part 3 exercise will help you to make those workshops more focused on the AI solutions and will help to focus the conversations on important problems that should be discussed among the impact group.

The workshop process

The full workshop should take between 5 and 10 hours to take (excluding breaks). Important to mark, we are not offering here a ready-to-go solution, but rather a framework and recommendations that should be then completed in the Co-creation workshop plan activity. The specific parts are dependent on the iteration progress, size of the impact group, and time & place workshop resources. We recommend relying on two facilitators, one could be a specialist on participation and AI already present in the team, another one could come from the Helsinki Lab or someone involved in the Ruuti participative process.

[2] In case the process is to be applied for the challenge that yet has no solution, we suggest conducting the big loop (yellow path).

The process of the workshop is following:

- welcoming, introduction and icebreaker [30 min]
- part 1: co-education [1-2h]
- part 2: specifying common goals for the product [30 min-1h]
- part 3: reframing the product [1:30-3h]
- part 4: establishing shared indicators [1-2h]
- summary and communication exercise [30min-1:30 h].

Each of those stages is described in the below subsections.

The workshop requires some preparation from both the lead team and participants. The former should communicate the initial information about the problem they wish to solve and the solution idea, without going into too much detail. The participants can be requested to fill in the form from part 1: co-education. By doing so, the co-education can be of better depth and the workshop can take less time. However, one should remember that it is not always possible for participants to prepare something beforehand.

Welcoming and ice-breaking

Time: 30 min

In the essential path, this co-creation workshop is the first opportunity for the impact group to meet. Therefore, it is essential to support networking from the very beginning. Online, networking can be enhanced by asking people to start introducing themselves in the chat. Offline, you can ask them to pin the name tags and introduce themselves to a person nearby.

Once the time comes to start the workshop, the facilitators should share the agenda for the day, set the rules for the workshop, and introduce the topic. Then, we suggest that the whole group spends 10 to 20 minutes on an icebreaker, which would be prepared beforehand by facilitators. The goal of setting the rules and the icebreaker is to establish a safe, energetic, and creative atmosphere, where everyone feels welcome to share their experiences and opinions. Here are some inspirations for setting the rules and here for icebreakers.

Part 1: What can we learn from each other?

Time: 1-2 h
Worksheet: pp. 74-77

In this part, each participant will share their perspective on the issues related to the problem to be solved and to the AI HOKS solutions based on their knowledge and

experience. We suggest using the framework from the worksheet to help participants focus and produce results in a similar framework. If the workshop is longer, participants can spend the first 30 minutes filling in the framework during the workshop. Otherwise, they should prepare it before coming.

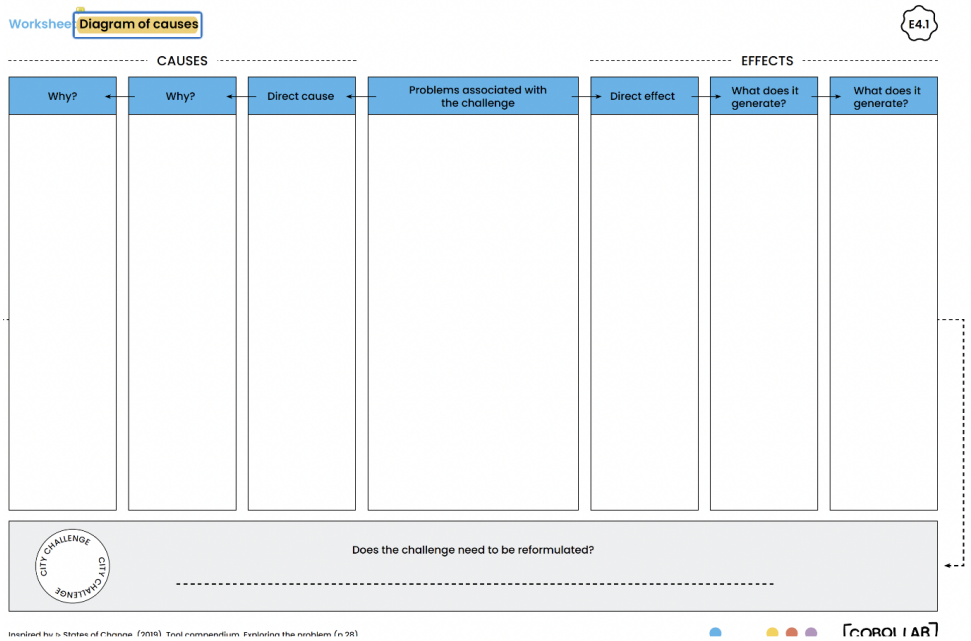


Fig. 7. Suggested framework to be used for co-education.

One of the most important rules of participatory design is that everyone has their own expertise to bring to the challenge (otherwise, they would not be there). Co-education is the last moment to make sure that everyone in the workshop realizes so, especially for the *experiential experts*, those whose expertise relies on their own experience of the challenge to solve. By making sure that it is clear from the very beginning, we help establish respect among participants and support by sharing important knowledge and experiences with each other.

The objective of this step would be to make sure that everyone gets a holistic vision of the challenge. For example, technical experts would learn more about the problem they are trying to solve, future users would learn more about technical possibilities and risks.

In this specific case, we imagine that eg. students can share their perspective on issues like dropouts, study plans, or relations with teachers; teachers could share their experiences of supporting students and the state of relations with them; psychologists can tell more about dropouts; AI specialists about some of the technical trade-offs, biases, etc. Moreover, the lead team or product developer can share what is possible and not with the AI solution. However, they should not go into the specific details of the solution itself.

Part 2: What are our common goals for the product?

Time: 30’ - 1h

In this part of the workshop, we encourage you to spend around half an hour agreeing on the common goals and priorities for the product. Firstly, we suggest using eg. sticky notes where each of the participants would write for 5 to 15 minutes what goals they would like this product to achieve. Depending on the group size, this can be done individually at first or immediately in a group. Example goals could be: AI HOKS should help students in getting better study results; AI HOKS should support teachers in administrative tasks; AI HOKS should minimize student dropouts etc.

Next, the goals should be ordered by priority. First, we encourage a small discussion on the viability and importance of the listed goals. Then, simple voting can be conducted to understand the group priorities. For example, each participant can be given three votes to distribute on the goals.

If the time permits, the lead team can decide to use a specific framework for goal setting, which will take around 1 hour, from this webpage.

This exercise is meant to help participants to establish the same vision on the main goals of the challenge after the co-education part.

Part 3: How do we make the tool?

Time: 1:30 - 3h

Part three is the culmination point for the workshop. The goal is to reframe the initial prototype. More specifically, participants would discuss and select the data, processes, models and metrics, power distribution, etc. for the next product implementation. We suggest this is done by reviewing with them different harm and technical trade-offs scenarios with prompts. In this playbook, we presented you with different parables and scenarios, which we grouped in Appendix C. We add below also a technical note on the trade-offs. We invite you to use those texts as an inspiration for the facilitators or workshop participants to conduct and focus the discussions. The results of the discussion, eg. discussion notes, created materials, or recordings, should be later thoroughly analyzed by the facilitators.

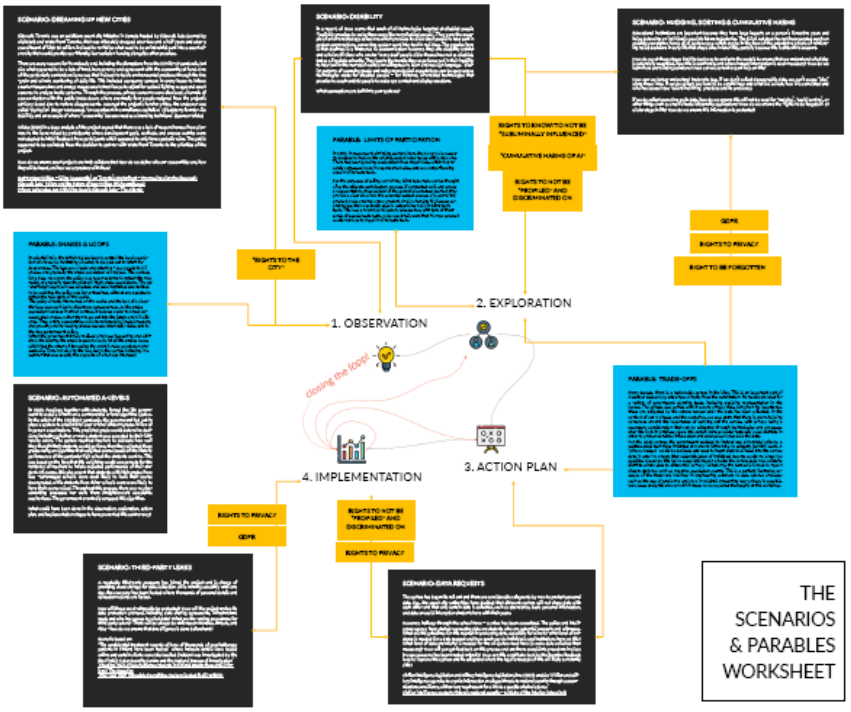


Fig. 8. See Appendix C: The Scenarios & Parables Worksheet.

Tradeoff Scenarios

In an ideal world, we would want a solution that can resolve the problems of eg. student dropouts, successfully allowing early interventions with measurable reductions in student dropout rates. Simultaneously, we want a solution that is known not to have any direct or indirect discrimination (in US terms: disparate treatment or disparate impact), measured against vulnerable groups in Helsinki's specific context. The system must also be privacy-preserving, successfully masking sensitive information and protected attributes not just from system admins but from the predictive algorithms themselves. In addition, there may be other desiderata that get revealed in the course of conducting the workshop.

Unfortunately, a general TANSTAAFL principle tempers our expectations in this regard. One should expect to make trade-offs between these desiderata, and the co-creation workshop is the perfect mechanism for taking the pulse of the concerned individuals when making these choices. There is already literature proving the existence of such tradeoffs in specific scenarios, and more such evidence should be expected as the science of machine learning matures. In the meanwhile, it would be a good starting point to conduct exercises in the workshop where such tradeoffs are hypothetically assumed to exist, and acceptable solutions are sought from the participants. These can take the form of simple, open-ended questions, for example: *given a choice between a predictive system that does not access race as an input, but with outputs that have been shown to be different for different races, and another predictive system that explicitly considers race as input but is*

able to provide balanced outputs that are not different for different races, which system should the city of Helsinki employ to predict student outcomes? They can also be more specific and direct: *what levels of noise are acceptable for the differential privacy layer between the cities data storage layer and model training/inference layers*? or more open: *Given a tension between accuracy and fairness, what are the considerations that should guide the prioritisation of one over the other? And why?.* We invite technical experts to discuss and prepare such questions with facilitators before the workshop.

Part 4: How do we make sure it works?

Time: 1 - 2 h
Worksheet: pp. 148 - 155

Part four is where the accountability process starts. That is also a moment where the power is given to the impact group, therein future users, to choose how, by whom, and what should be audited from their perspective. For the time-restricted option, we suggest focusing only on the *what* part using the framework from the worksheet. There, the impact group would agree on the shared indicators that would reflect the success criteria for reaching the goals set in part 2 of the workshop.

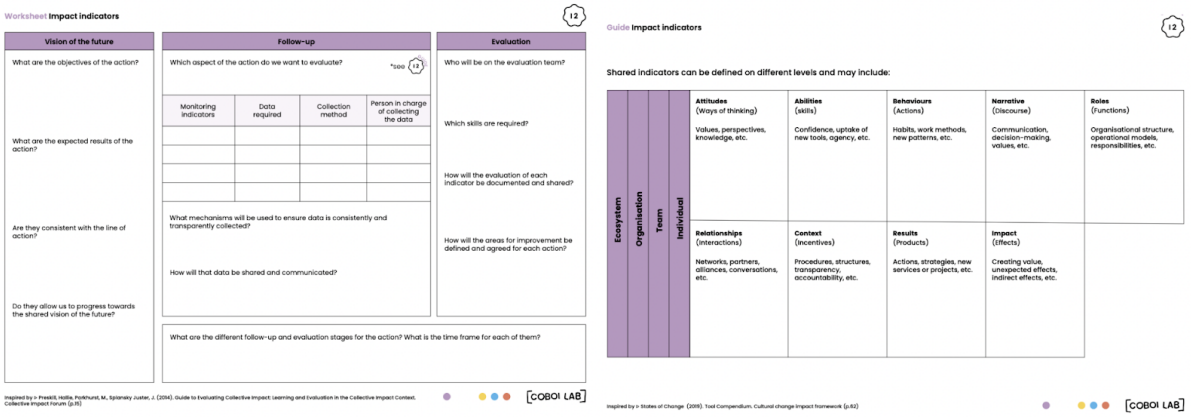


Fig. 9. Impact indicators worksheets.

In case there is time left, we encourage you to also discuss with participants on the "who" and "how" parts: so who should be involved in evaluating the product according to the chosen metrics, and how (when, where) it should happen. Otherwise, we suggest answers to those questions in the section *Closing the loop*.

What did we learn and how do we communicate about it?

Time: 30' - 1:30 h
Framework: pp. 144 - 147

Time for the workshop to be finished. As a nice exercise of summarizing what has happened, we suggest the exercise from the worksheet. The framework includes spaces

such as the problem to be solved, who benefits from solving it, what makes it unique, and in what format it would be communicated. This is a nice final action that shows participants that their opinions matter and that they should be included in any publication of this process.

Afterwards, we suggest that there is some space left for participants to openly discuss the workshop and share their initial thoughts after it. We also encourage you to share with the participants a feedback form. The meetings with the impact group might be repeated in the next iterations, therefore it is important to work on the quality of those.

Parable: Trade-offs

Every decade, there is a nationwide census in the USA. This is an important part of American democracy and draws directly from the constitution. Its results are used for a variety of government planning tasks, including deciding representation in the House. The census also carries with it severe privacy risks, but given its importance, these are mitigated by the census bureau after the data has been collected. In the context of our 4 phases and the workshop, we may state that there is overwhelming consensus around the importance of carrying out the census, with privacy being a secondary consideration that can be mitigated through technologies and processes after the fact. In previous years, the census bureau would randomly swap districts to protect sensitive information and anonymise/randomise the data.

For the 2020 census, the government decided to instead use differential privacy, a mathematical technique invented at Harvard University, to mitigate just this issue. A “privacy budget” would be decided, and used to insert statistical noise into the census data to ensure that reidentification of individual records would no longer be possible. However, this led to some unforeseen effects: the census data now contains districts where (due to differential privacy limitations), the bureau is forced to report bizarre statistics such as negative population counts. This is a perfectly illustrative example of the trade-offs involved in engineering solutions to data science problems such as the use of predictive analytics in Helsinki. Preserving user privacy is possible, but comes at a significant cost that needs to be explored thoroughly at the workshop.

Source: Harvard Researchers Identify Accuracy Concerns in Census Bureau’s New Privacy System at The Harvard Crimson

Appendix C: The scenarios & parables worksheet

SCENARIO: DREAMING UP NEW CITIES

Sidewalk Toronto was an ambitious smart city initiative in Canada headed by Sidewalk Labs (owned by Alphabet) and Waterfront Toronto, that was ultimately dropped after two and a half years and after a commitment of USD 50 million. It aimed to revitalize what used to be an industrial port into a smart micro-city that would provide eco-friendly, low-emission housing alongside other promises.

There are many reasons for its untimely end, including the disruptions from the COVID-19 pandemic, but also what appeared to be a lack of trust, transparency and agreement with the community at hand. One of the particularly contentious issue was that it aimed to tackle environmental problems through the fine-grain and minute monitoring of daily life. This included occupancy sensors in every home to inform smarter temperature and energy usage; smart street lamps to adjust for optimal lighting usage; and smart cameras to analyze traffic patterns. Through the process, privacy concerns were dismissed, channels of communication with the public broke down, where eventually four people resigned from the project's advisory board due to various disagreements. Amongst the project's harsher critics, the endeavor was called "dystopian" (Roger McNamee), "an experiment in surveillance capitalism" (BlackBerry founder Jim Balsillie), and an example of where "'community' became used as a branding technique" (Spencer Wicks)

Wicks (2020) in a deep analysis of the project argued that there was a lack of responsiveness from planners to the issue raised by participants; where development goals, methods, and success metrics were not adapted to initial feedback from participants which appeared to only have symbolic value. The public appeared to be excluded; from the decision to partner with Waterfront Toronto to the priorities of the project.

How do we ensure smart projects are truly collaborative? How do we define who our communities are, how they will be heard, and how we co-produce with them?

Not-So-Smart Cities – Using 'community' on Toronto's Waterfront – Progressing planning (lse.ac.uk)
Sidewalk Labs' Failure and the Future of Smart Cities (triplepundit.com)
Privacy Advocates Are Criticizing Google Sidewalk Labs - The Atlantic

SCENARIO: DISABILITY

In a report, AI Now warns that much of AI technologies targeted at disabled people "implicitly promises to make them more like non-disabled people". They gave the example of an AI-enabled app called Ava which promises to allow Deaf people to take part in spoken conversations, converting spoken word into text in real time. The premise here is that deafness is a hindrance to communication however, they cite disability activist and scholar, Eli Clare who wrote: "many deaf people claim themselves not as disabled, but as a linguistic minority. They locate the trouble they experience not in their inability to hear but in the non-deaf world's unwillingness to learn and use sign language." Similar patterns of assuming needs and using neurotypical expectations can be seen across technologies made for disabled people – for instance, AI-enabled technologies that promise to coach autistic people to make eye contact and display emotions.

What assumptions are built into your systems?

SCENARIO: NUDGING, SORTING & CUMULATIVE HARMS

Educational institutions are important because they have large impacts on a person's formative years and helps determine an individual's possible future trajectories. The EU AI Act draft has not incorporated much on possible cumulative harms of AI systems (e.g. which can be in the form of the cumulative effects of reinforcing racial; decisions in early life that shape ones in later life), partially because this is difficult to measure.

How do we at these stages find the best way to evaluate the models to ensure that we understand what bias is potentially magnified, how this is monitored, and where human intervention is most necessary? How do we learn from pre-existing solutions? How can we consult expert help on this?

How can we better understand trade-offs (e.g. if we don't collect demographic data, we can't assess "bias" along these lines. If we do collect, we have to be careful how and what we collect, how it is correlated and who has access? See "colorblind hiring" practices and its problems).

If we do collect more fine grain data, how do we ensure this will not be used for "nudging", "social sorting", or other things such as credit checks/citizenship applications? How do we ensure the "rights to be forgotten" at a later stage in life? How do we ensure this information is protected?

PARABLE: LIMITS OF PARTICIPATION

In 1985, in response to shrinking market share, the Coca-Cola company decided to replace the original, secret Coke recipe with a new one. There had been growing competition from Pepsi-Cola, which had recently surpassed Coke in supermarket sales, and was outperforming coke in blind taste tests.

For the purposes of selling soft-drinks, blind taste tests can be thought of as the ultimate participatory process. If conducted well, and across a representative cross section of the potential consumer market, they provide a clear view into the potential market success of a soft-drink product. Coke created a new product, finally changing it's famous secret recipe, that was finally able to outperform Pepsi in blind taste tests. This was a robust participatory process too, with tens of thousands of sample taste tests. Coke was finally sure that it's new product would not lose to Pepsi in blind taste tests.

RIGHTS TO KNOW/TO NOT BE "SUBLIMINALLY INFLUENCED"

"CUMULATIVE HARMS OF AI"

RIGHTS TO NOT BE "PROFILED" AND DISCRIMINATED ON

GDPR

RIGHTS TO PRIVACY

RIGHT TO BE FORGOTTEN

"RIGHTS TO THE CITY"

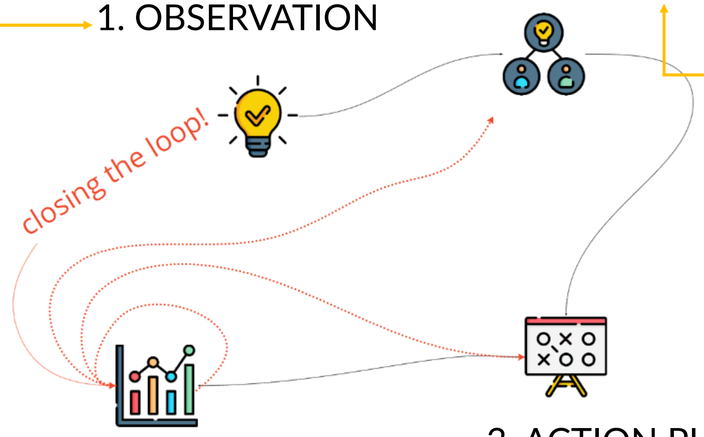
PARABLE: SNAKES & LOOPS

In colonial India, the British Raj decided to control the local population of snakes by instituting a bounty to be paid out in return for dead snakes. The idea was simple and effective - pay people to kill snakes, and gradually the snake population will reduce. This worked, for a time. However, the policy was never updated to reflect the new reality of gradually reducing (but still high) snake populations. The set-and-forget approach was adopted, and once instituted and verified to be working, the policy was left untouched, without any updates to reflect the new state of the world.
This policy directly intervened on the world, and the lack of a close-the-loop approach led to disastrous consequences. As the snake population reduced in urban centres, it became easier to breed domesticated snakes, rather than to go out into the jungle and kill wild ones. Thus, entire communities were transformed to snake breeders, and growing and harvesting snakes became financially viable due to the new government policy.
When the government finally realised what was happening and withdrew the bounty, the snake breeders simply let all the snakes loose, which had the effect of increasing the overall snake population dramatically. Thus, not closing the loop led to the system behaving in a manner that was exactly the opposite of what was intended!

SCENARIO: AUTOMATED A-LEVELS

In 2020, Foxglove together with students, forced the UK government to make a U-turn on a controversial A-level algorithm system. In the midst of the COVID19 pandemic, the government had put in place a system to predict final year A-level student grades in lieu of in person examinations. This predicted grade would determine decisions including whether or not a student was accepted to their University offers. The system required teachers to estimate how well they thought their students would perform, then they adjusted these grades using an algorithm that weighted scores based on the historic performance of the particular high school the students went to. This was based on the idea that this weighting would compensate for the tendency of teachers to inflate expected performance of their student. Of course, this also meant that students with high grades from less –advantaged schools were most likely to have their scores downgraded, while students from richer schools were more likely to have their scores raised. Throughout this process, there was no clear consulting processes nor were there straightforward complaints mechanisms. The government eventually scrapped this algorithm.

What could have been done in the observation, exploration, action plan, and implementation stages to have prevented this controversy?



PARABLE: TRADE-OFFS

Every decade, there is a nationwide census in the USA. This is an important part of American democracy, and draws directly from the constitution. Its results are used for a variety of government planning tasks, including deciding representation in the House. The census also carries with it severe privacy risks, but given its importance, these are mitigated by the census bureau after the data has been collected. In the context of our 4 phases and the workshop, we may state that there is overwhelming consensus around the importance of carrying out the census, with privacy being a secondary consideration that can be mitigated through technologies and processes after the fact. In previous years, the census bureau would randomly swap districts in order to protect sensitive information and anonymise/randomise the data.
For the 2020 census, the government decided to instead use differential privacy, a mathematical technique invented at Harvard University, to mitigate just this issue. A "privacy budget" would be decided, and used to insert statistical noise into the census data in order to ensure that reidentification of individual records would no longer be possible. However, this led to some unforeseen effects: the census data now contains districts where (due to differential privacy limitations), the bureau is forced to report bizarre statistics such as negative population counts. This is a perfect illustrative example of the trade-offs involved in engineering solutions to data science problems such as the use of predictive analytics in Helsinki. Preserving user privacy is possible, but comes at significant cost which needs to be explored thoroughly at the workshop.

RIGHTS TO PRIVACY

GDPR

RIGHTS TO NOT BE "PROFILED" AND DISCRIMINATED ON

RIGHTS TO PRIVACY

SCENARIO: THIRD-PARTY LEAKS

A reputable third-party company has joined the project and is charge of providing cloud storage for data collection. All is running smoothly until one day, the company has been hacked where thousands of personal details and education records are leaked.

How will those most vulnerable be protected? How will the project review its data protection protocol, including data sharing agreements, infrastructure used, and who has access to what data? What are the vetting procedures for third party actors? There will be always be unforeseen troubles, threats, and risks - how do we ensure that due diligence is done beforehand?

Scenario based on:
"The confidential treatment records of tens of thousands of psychotherapy patients in Finland have been hacked" where intimate details were leaked online and certain clients were blackmailed (incident was investigated by the the Finnish Cyber Security Centre and the National Bureau of Investigation"
'Shocking' hack of psychotherapy records in Finland affects thousands | Finland | The Guardian
They Told Their Therapists Everything. Hackers Leaked It All | WIRED

SCENARIO: DATA REQUESTS

The system has begun its roll out and there are considerable safeguards by now to protect personal data. E.g., the smart city authorities have decided that different sectors will not share data with each other and that only certain data is collected, such as attendance, basic personal information, and data on social interaction students have with their peers.

However, halfway through the school term – a crime has been committed. The police and intelligence agency have started requesting data on students who are potential suspects and witnesses. What are the guidelines for this request in protecting data, especially for minors? What level of evidence is needed for a data request and how much say do individuals and institutions have on this? What level of data granularity is needed for the AI system used here (is more data collected than necessary)? How will you get feedback on this process and are there complaints procedure in place in case someone has been treated unjustly? How can this contribute back to the iterative feedback loop to improve the system and its safeguard where the legal landscape of this will likely constantly shift?

Civilian intelligence legislation and Military intelligence legislation (June 2019); enables "civilian and military intelligence agencies to acquire information on alleged threats to national security through communications surveillance, without any requirement for a link to a specific criminal offence."
Civilian intelligence protects Finland's national security - Ministry of the Interior (intermin.fi)

THE SCENARIOS & PARABLES WORKSHEET