societies

MDPI

Article

Training in Co-Creation as a Methodological Approach to Improve AI Fairness

Ian Slesinger [1,*], Evren Yalaz [1], Stavroula Rizou [2], Marta Gibin [3], Emmanouil Krasanakis [2] and Symeon Papadopoulos [2]

[1] Trilateral Research, London SW1X 7QA, UK; evren.yalaz@trilateralresearch.com
[2] Centre for Research & Technology Hellas (CERTH), 57001 Thermi, Greece; rizstavroula@iti.gr (S.R.); maniospas@iti.gr (E.K.); papadop@iti.gr (S.P.)
[3] Department of Sociology and Economic Law, University of Bologna, 40126 Bologna, Italy; marta.gibin2@unibo.it
* Correspondence: ian.slesinger@trilateralresearch.com

Abstract: Participatory design (PD) and co-creation (Co-C) approaches to building Artificial Intelligence (AI) systems have become increasingly popular exercises for ensuring greater social inclusion and fairness in technological transformation by accounting for the experiences of vulnerable or disadvantaged social groups; however, such design work is challenging in practice, partly because of the inaccessible domain of technical expertise inherent to AI design. This paper evaluates a methodological approach to make addressing AI bias more accessible by incorporating a training component on AI bias in a Co-C process with vulnerable and marginalized participant groups. This was applied by socio-technical researchers involved in creating an AI bias mitigation developer toolkit. This paper's analysis emphasizes that critical reflection on how to use training in Co-C appropriately and how such training should be designed and implemented is necessary to ensure training allows for a genuinely more inclusive approach to AI systems design when those most at risk of being adversely affected by AI technologies are often not the intended end-users of said technologies. This is acutely relevant as Co-C exercises are increasingly used to demonstrate regulatory compliance and ethical practice by powerful institutions and actors developing AI systems, particularly in the ethical and regulatory environment coalescing around the European Union's recent AI Act.

Keywords: participatory design; co-creation; AI bias; AI fairness; social inclusion; AI ethics; AI regulation; training

check for updates

Citation: Slesinger, I.; Yalaz, E.; Rizou, S.; Gibin, M.; Krasanakis, E.; Papadopoulos, S. Training in Co-Creation as a Methodological Approach to Improve AI Fairness. *Societies* **2024**, *14*, 259. https://doi.org/10.3390/soc14120259

Academic Editors: Theodora Saridou and Charalampos Dimoulas

Received: 26 September 2024
Revised: 27 November 2024
Accepted: 28 November 2024
Published: 3 December 2024

1. Introduction

The rapid innovation and permeation of AI technology require a range of socio-technical approaches to ensure its safe, responsible, and socially equitable deployment. One key aspect of this is the design and deployment of fair AI that avoids reproducing existing social biases and forms of discrimination against vulnerable or marginalized social groups. The objective of creating fair AI also aligns with emerging legal and regulatory requirements such as the EU AI Act [1] and codified ethical imperatives such as those encapsulated in UNESCO's Recommendation on the Ethics of Artificial Intelligence [2], the European Commission's (EC) High-level Expert Group's (HLEG) 2019 Ethical Guidelines for Trustworthy Artificial Intelligence [3], and the OECD's Principles for Responsible Stewardship of Trustworthy AI [4].

However, fair AI design presents particular challenges due to the profound societal, political, and economic implications of its ever-increasing proliferation and the rapid pace, complexity, and scale at which AI technology is developing. Participatory design (PD) and co-creation (Co-C) approach, if performed well, can give voice to stakeholders' perspectives—particularly less privileged ones, in the design processes of AI to enable fairer AI solutions. On the other hand, how to include stakeholders, especially the ones who are not knowledgeable about AI and how it works remains a challenging task. This paper focuses on one approach to tackle these challenges by incorporating training on defining AI and how its use can impact participants from identified vulnerable groups in the co-creation process as a way to engage them with the socio-technical complexity inherent to the problem of AI fairness.

Work to develop and implement AI fairness starts at the design phase of AI applications. This requires adapting existing PD and Co-C approaches that are frequently used in Human–Computer Interaction (HCI) and technology design. PD can be seen as a precursor to Co-C, whereby PD obtains input from end users to test and modify a prototype design that has been formulated by the design team, whereas Co-C is a process that works to incorporate users' requirements iteratively through cyclical consultation throughout the design process since its early phase [5]. This paper primarily draws from Co-C, although the term PD is also used throughout, particularly in relation to discussions of the existing literature, and there is significant overlap between the concerns and findings for both PD and Co-C.

This paper investigates the benefits and challenges of incorporating training components in the AI co-creation process through Co-C workshops conducted with vulnerable and marginalized stakeholder groups as part of the design process for an AI bias identification and mitigation tool of the Horizon Europe MAMMOth project. Bias in AI systems can be defined as "the inclination or prejudice of a decision made by an AI system which is for or against one person or group" [6], often leading to unfair treatment and discrimination [3,7], and has been recognized as a major ethical concern for AI systems [8–10]. Conversely, fairness in the context of AI systems is established by preventing bias, discrimination, and stigmatization [3].

The primary causes of AI biases are socio-technical in nature and often result from the inputting and processing of data that reproduce existing societal biases and inequalities to train AI models, which then reproduce and amplify these inequalities when deployed [6,11]. It is, therefore, important to recognize that AI bias can involve a feedback loop between data used to train and test AI systems, their design, application, results, and their possible social impacts [7]. This necessitates a comprehensive approach to mitigation, addressing the issue through the combined interdisciplinary efforts of sociology, computer science, and law.

MAMMOth tackles AI bias through an approach that is multi-modal—meaning it accounts for bias in a range of digital file types and formats, e.g., text, datasets, images, and audio files—and multi-attribute—which means it accounts for intersecting forms of bias that AI data processing can exacerbate based on multiple characteristics of vulnerability or marginalization, e.g., gender, age, ethnic origin, and sexual orientation. In the Co-C process for MAMMOth, training on AI definitions, terminology, and the project's socio-technical implications was carried out to empower non-technical stakeholders to articulate their insights on AI fairness clearly in evidence-gathering activities. In the Co-C workshops, data collection was carried out based on participants' evaluation of three use cases in which AI bias can have a significant impact—financial decisions, e.g., loan applications; biometric identity verification, e.g., automated passport control gates based on facial recognition; and bias in academic citation searches.

The next section will present a literature review that examines several aspects of PD with vulnerable groups and AI relevant to this paper. This is followed by a discussion of the legal and ethical aspects relevant to AI bias and the role of Co-C, and incorporating training in it. The remainder of the article will provide an empirical description of the role of training in the MAMMOth Co-C process and a reflective discussion of the benefits of incorporating training in AI, as well as its limitations and lessons learned.

2. Literature Review

2.1. Participatory Design with Marginalised Groups

PD proposes that the people who will be influenced by a new design or technology must be a part of its development [12]. However, the determination of who these stakeholders are has been debated and has shifted over time. While more traditional uses of PD emphasize "user-centered" design, social justice perspectives in PD move away from perceiving people solely in terms of their relationship to the use of technology and come up with the term "human-centered" design [13]. Human-centered design brings up a more complex understanding of users and stakeholders by accounting for potentially vulnerable, marginalized, hard-to-reach groups that can easily be ignored in a more generic user-centered perspective.

Vulnerability is a contested concept that is dynamic, multi-dimensional, and context-dependent [14,15]. A certain level of vulnerability exists in any human life [16]. While some people might be more vulnerable than others, who is vulnerable can change across different contexts. Attempts to define vulnerability are often coupled with narrowing it down to vulnerable groups. Vulnerable groups can include those who are socially excluded and experience disadvantages "due to not having access to material resources, being unable to exercise their voice, or being discriminated against on the basis of their age, sex, disability, race, ethnicity, or economic and migration status" [17].

The inclusion of vulnerable groups in PD and Co-C raises various questions. Why is the inclusion of vulnerable groups in PD/Co-C crucial? How is vulnerability defined? Which groups are included or left out? How are the participants recruited considering their hard-to-reach condition? Which research methods in PD projects are sensitive to the needs of vulnerable groups? How and to what extent do participants influence the decision-making procedure? In PD/Co-C, the goal is not only to empirically understand the activity, e.g., understanding people's perspectives to build a system for them but to collaboratively develop and redefine new technologies where participants' "tacit knowledge" and researchers' "more abstract and analytical knowledge" are bridged [18]. In this way, PD and Co-C can achieve practical and political improvements in participants' lives. PD has a long history of democratic practices, empowerment of marginalized groups, and giving voice to those people who may be invisible and unheard in existing power structures [19,20]. Self-reflective and deliberative planning of inclusion of vulnerable groups in PD is necessary for not reinforcing the benefits of already empowered groups in the design process [21].

PD/Co-C projects might involve a wide range of vulnerable groups. A systematic literature review of PD studies with less privileged participants identified eight categories of participants who were perceived as less privileged individuals by the reviewed papers [17]: children; refugees, immigrants, and asylum seekers; older adults; people with cognitive or physical impairment; ethnic or aboriginal people; individuals in critical neighborhoods; women; unemployed people. These categories are not homogenous and might bring multiple vulnerabilities together and require an intersectional analysis, such as studies that work with children with disabilities [22] or children of color [23].

Reaching out to vulnerable groups and enabling their participation in PD/Co-C requires careful consideration. There might be various barriers to including vulnerable groups in PD and Co-C projects [24]. Such populations might be hard to reach because of their geographical or social locations (e.g., rural populations living in remote places) because they are hidden, stigmatized, or live in legal limbo (e.g., LGBT people or undocumented migrants), or because they lack necessary resources and skills (e.g., older people, children, those with poor literacy). There might be several solutions to overcome these participation barriers. The first strategy would be identifying the key vulnerabilities and devising the recruitment strategy accordingly rather than engaging in convenience sampling. Snowball sampling and its variations can be effective in recruiting traditionally underserved and vulnerable groups [25]. Building rapport and trust and following ethical engagement principles are other important factors facilitating the inclusion of vulnerable groups in PD/Co-C.

The methods of involving vulnerable groups in PD/Co-C need to be tailored according to research objectives and participants' needs [23]. Hands-on activities such as workshops are the most frequently used methods in PD projects with vulnerable groups [17]. While workshops with children often include playful activities, workshops with older adults might include pre-workshop support with technology. In addition to workshops, individual interviews and observations are other methods of involving vulnerable participants in PD. These methods can be accompanied by different toolkits (such as photo-elicitation, daily journals, and different workshop materials) to enrich the participatory process.

Last but not least, vulnerable populations can participate in PD/Co-C projects at different stages of design development and in different capacities. Different stages of involvement include the stage of problem framing (understanding the needs of the participants), ideation (creative envisioning activities), validation (testing activities), and development (developing prototypes) [17]. Depending on the conceptualization of PD, participants can have different roles, such as "informants", "learners", "research partners", and "design partners" [17].

2.2. PD for AI Fairness

PD and Co-C approaches to incorporating fairness in AI applications must negotiate the multiple notions of fairness held by diverse stakeholders [26–29]. In this way, machine learning algorithms can be used in deliberative PD processes as a form of S.L. Star's "boundary object" through which participants can negotiate shared beliefs and values with other stakeholders, as well as "the complexity of their differences within the problem space" [30]. A specific area of focus for PD/Co-C for AI is an explicitly value-led approach that aligns responsible AI design with social good. Wang et al. draw attention to the need for UX practitioners to develop new methods and strategies to incorporate responsible AI, including fairness, into technology design, including consideration of how UX practitioners conceptualize AI and elicit mental models of AI from co-design participants [31]. In work with AI system developers and practitioners, Madaio et al. indicate the value of co-designing AI fairness checklists within organizations as a way of formalizing ad-hoc processes and engaged advocacy from members in a way that encourages internal deliberation and affords permission for individuals to raise potential concerns [32]. Another approach is to invest AI practitioners in accountability for AI bias mitigation, including through awareness raising and bias mitigation training [33].

Several approaches encourage bringing AI into public spaces to broaden AI literacy and co-creation [34]. These methods include Real-World Laboratories (also called Living Labs) that empower citizens to make decisions about technological adoption [35] and TrustScapes that encourage end users to visually represent their ideas on data protection, algorithmic bias, and online safety to inform AI design and policy recommendations [36]. Another is PD fictions which encourage participants to experiment with and elucidate the multiple future pathways through which AI can influence the world to ensure AI design decisions that are sensitive to human values [37].

Zicari et al. suggest involving interdisciplinary experts in PD processes who can exert agency in designing, implementing, and regulating AI systems—such as healthcare professionals, technology designers, ethicists, legal experts, and social scientists—can arbitrate AI fairness in a "neutral" way to achieve fairer outcomes [38]. The expertise these interdisciplinary actors bring can help identify technical issues affecting fairness that would likely be opaque to laypersons, as well as identify and negotiate conflicts of interest between both direct and indirect stakeholder groups; however, as studies in the sociology of science have shown, the political neutrality of scientific knowledge is contestable. Each discipline is, in fact, an outcome and a reproduction of a certain way of producing knowledge and seeing the world, which evolves over time and is linked to the social context and power dynamics in which it develops [39].

Critical perspectives to PD/Co-C for social good are more circumspect about defining what social good entails and whose values and principles should be prioritized, particularly for minoritized or marginalized groups who are outliers in a utilitarian conception of social good. This includes challenging a perceived absence in general of participatory design and deliberative engagement with affected stakeholders of unfair machine learning algorithms, who are not necessarily the system end users with whom UX designers typically work in PD processes [40]. Other perspectives critique the assumption that when diverse stakeholder engagement in PD does take place, it automatically leads to greater fairness without reflexive engagement in the power relations implicated in the design and conduct of the PD or Co-C process [41,42]. Such perspectives also express concern about PD becoming a weak 'box-ticking' exercise rather than a genuine socially transformative process that empowers users and enables greater social equality [43,44]. Delgado et al. provide a framework for critical evaluation of the purpose, value, and efficacy of PD and similar approaches that consider the form that a particular design process takes, including how decisions are made and articulated, whom it includes and excludes, the scope and remit of the PD process, the extent to which it engages and empowers stakeholders [45]. Other methodologies for redressing this power imbalance that prioritize the voices of those individuals or groups disproportionately likely to face harm from poorly designed AI applications include a capabilities approach [46] and an iterative *agile participatory design* approach to designing technology objectives and user requirements with marginalized user groups [47].

Several recent empirical studies using PD and Co-C to address AI fairness in situated contexts with minorities, marginalized groups, and civil society actors provide new understandings and a range of methodological strategies for mitigating bias and harm from unfair algorithms, e.g., [48–50]. Suresh et al. highlight the need to design *with* communities of diverse and potentially marginalized stakeholders (including civil society activists) to ensure that systems "work not only for the mainstream, majority use case, but also for those on the margins" [48]. Katell et al. demonstrate how such work can benefit from incorporating non-technical interventions as part of a socio-technical approach to mitigating unfairness in Co-C [49]. In addition, approaches and resources that prioritize localized and context-specific issues pertinent to a particular application are often highly effective [26,49], although these can sometimes come into tension with the desire to achieve broadly applicable solutions through design.

2.3. Training in PD

Several studies add 'skills development' as the first step of PD/Co-C processes and conceptualize participants as 'learners.' The underlying assumption in adding skill development is that "people need not only the opportunity to participate, but capabilities, sense of agency, and literacy on the topics they will work on" [17]. For example, the inclusion of older adults in PD/Co-C processes requires developing technical skills as a first step [50]; however, there appears to be a gap in the literature regarding the focal approach of this paper on the critically reflective incorporation of expert-defined training on complex social scientific and technical concerns—in the present case, AI fairness—to share knowledge with end users in PD/Co-C processes. This paper both addresses this gap and suggests that including training in PD involving complex technology empowers end users and allied stakeholders to articulate their positional perspectives, particularly those from marginalized groups, and enables them to provide a more informed and detailed contribution to the design process to produce societally fairer outcomes that address their interests and needs.

3. How Training in Co-Creation Relates to EU Legal and Ethical Approaches to AI Bias

AI systems creators (including engineers, developers, and auditors) address the issue of achieving systems' compliance with evolving regulatory requirements from a technical perspective [51]. This first involves the identification of the applicable legislation based on its context and then its practical implementation [52], which both affect the technology design. Yet, the impact of training in the Co-C process extends beyond the technical development of an AI system, as it also has implications for how compliance regimes evolve in relation to the dynamic state of AI legislation. This is notable in the EU environment, namely with reference to the Ethics Guidelines for Trustworthy AI [3] and the AI Act [1], particularly concerning AI bias. This requires thoughtful engagement from training designers and facilitators to ensure that the training meaningfully contributes to fairer and safer AI design rather than becoming an instrumental vehicle for demonstrating compliance in a hollow way.

3.1. Ethical Guidelines for Trustworthy AI

Several non-legally binding codes from inter-governmental organizations set out ethical principles for the safe and socially responsible implementation of AI technologies, including trustworthiness and fairness. Both the OECD Principles for Responsible Stewardship of Trustworthy AI and UNESCO Recommendation on the Ethics of Artificial Intelligence call for AI Actors to protect fairness and reduce social, cultural, economic, and gender inequalities throughout a given AI system's lifecycle by ensuring the inclusion of underrepresented and marginalized populations in the development and implementation of AI systems [2,4]. The methodological approach of including training in Co-C argued for in this paper instantiates the ethical principles of fairness, participatory inclusion, social justice, and knowledge-building codified in the aforementioned documents.

In the EU context, the third chapter of the Ethics Guidelines for Trustworthy AI, which was the main non-legally binding instrument regarding AI at the EU level prior to the introduction of the AI Act, sets out an assessment list for the compliance of AI systems with its seven key requirements: (a) human agency and oversight; (b) technical robustness and safety; (c) privacy and data governance; (d) transparency; (e) diversity, non-discrimination and fairness; (f) societal and environmental well-being; (g) accountability. The Ethics Guidelines for Trustworthy AI has identified the fostering of AI literacy, including training and, in general, the involvement of all the stakeholders in AI systems as mechanisms for realizing accountability and improving the trustworthiness of an AI system. [3]. Notably, the provision of informed participation and education about AI systems includes all relevant stakeholders, including the groups affected by AI-based decisions, which, in the case of discrimination, often refer to marginalized communities.

In addition, the Assessment List for Trustworthy AI (ALTAI) [53], which was derived through multi-stakeholder Co-C processes, examines the needed proactive steps to reduce and prevent hazards associated with the seven above-mentioned ethics guidelines [54] and introduces training as a mechanism to ensure human oversight, transparency, non-discrimination, societal and environmental well-being, and accountability. Regarding non-discrimination, it lays down the need for consultation with stakeholders who may directly or indirectly be affected by the AI system. To achieve substantive benefits rather than merely fulfilling procedural requirements, the involvement of the widest range of stakeholders, as outlined in the list, should be incorporated throughout all stages of the system.

3.2. AI Act Provisions Associated with AI Bias and Training

The AI Act [1] is one of the first and most comprehensive legally binding instruments for AI regarding discrimination in AI systems. The AI Act formalizes the requirement for the identification and mitigation of unfair discrimination in high-risk AI systems [Article 10 par. 2 (f), (g)]. It requires AI system providers to identify the risk level of each AI system, taking into consideration the importance of preventing the (re)production of biases that can lead to discrimination, as well as the potential protected ground(s) of discrimination under EU law. In parallel, the requirement for conducting a Fundamental Rights Impact Assessment (FRIA) under Article 27 for high-risk systems aims at mitigating discrimination in general, among other fundamental rights.

The involvement of stakeholders and the endorsement of training in Co-C can be applied as mechanisms for implementing the regulation. First, at the level of the deployer's obligations, the AI Act introduces a risk management approach for high-risk AI systems by organizations that includes a set of measures (including education and training) as factors that can mitigate the high-risk systems' risks (Article 9 par. 5, subpar. 3). At the level of supporting the enforcement and implementation of the regulation, the AI Office is expected to include involvement with relevant stakeholders (as defined in Article 4 of the Commission decision (Brussels, 24 January 2024, C (2024) 390 final)) within its criteria for compliance in its guidance and compliance monitoring activities [55]. The following section will describe the methodology applied to implement engagement with vulnerable and marginalized stakeholders through training in the MAMMOth Co-C process.

4. The Use of Training in Co-C to Support the Design of an AI Bias Mitigation Toolkit

Co-C with vulnerable groups was applied during the creation of a toolkit that helps AI system creators assess the bias of their datasets and systems. The toolkit offers a user interface to run methods for tackling three critical challenges: bias contained within the ML training data of AI components, AI-induced bias that could take the form of spurious correlations (e.g., using postal codes as a proxy of race to determine economic status) and bias occurring during the interpretation of system outcomes. In the toolkit, AI system creators can run existing and new methods to quantify bias for multi-attribute and multi-modal data, including preliminary assessment during design and continuous re-evaluation during maintenance. Toolkit users can also obtain qualitative suggestions of how found biases may be mitigated.

To extract the requirements for the toolkit's creation, a multi-faceted and comprehensive design approach was followed that involved the engagement of diverse stakeholders to a significant degree. This involved desk research into existing solutions, consultations with domain experts and the MAMMOth consortium, and a set of Co-C activities with vulnerable and marginalized groups (the focus of this paper) to ensure a more equitable representation of their interests. To enhance the quality and depth of the process outcomes, the consortium adapted training materials for public engagement on AI bias and discrimination developed by the University of Groningen (RUG) for use in the Co-C workshops. Inputs from the consultation with vulnerable users influenced the selection of the following requirements: (1) user requirements, where underrepresented groups affected by evaluated AI systems are also considered as indirect users; (2) research requirements for evaluating the project use cases and effectiveness of the toolkit; (3) functional requirements that capture actions that the toolkit needs to perform to comply with the state-of-the-art and the state-of-practice; (4) non-functional requirements, such as usability and data security.

4.1. Incorporating Training in the Co-C Process

Initial steps in the co-creation process included conceptualization of the desired objectives of the toolkit and the envisaged end users, which were defined during an internal workshop carried out amongst the MAMMOth project members. Surveys were also conducted amongst communities of vulnerable users recruited through local civil society organizations involved in the project, including women at risk of social exclusion, e.g., ethnic minority and domestic violence victims, migrants, and recipients of social protection services (see Table 1). These surveys set out to investigate the level of awareness of people from vulnerable communities on AI and their perspectives on AI fairness and to recruit participants to participate in the Co-C workshops with their informed consent.

A series of Co-C workshops were carried out alongside the other design activities in MAMMOth to advise how the requirements for the MAMMOth bias mitigation tool were formulated. The aim of these workshops was to clarify from a socio-technical perspective how stakeholders from identified marginalized groups considered and felt affected by, or that they would be affected by AI bias. The topics that were explored during the workshops included, but were not limited to, the following:

- Participants' opinions on the use of AI in the specific use cases of finance decisions, identity verification, and academic citations and collaborations;
- What they considered a fair outcome when using these systems in real life;

- Their concerns about the use of AI in the use cases and the expected benefits;
- Their previous experiences, if any, with the use of AI in specific use cases;
- What groups of people they thought might be negatively affected by the use of AI in the use cases.

Table 1. Organized Co-creation workshops.

	Use Case(s)	Moderator	Location	Number of Participants	Language	Marginalized Group(s) Represented
1	Finance, identity verification	Associació Fòrum Dona Activa (DAF)	Spain and online	17 on-site and 2 online	Catalan and Spanish (mixed)	Women at risk of social exclusion, including ethnic minority and domestic violence victims
2	Finance	IASIS	Greece	20	English	Young people
				4	English	Ethnic minorities and young people
				4	Greek	Unemployed receiving social support
3	Finance, identity verification	Diversity Development Group (Lithuania; DDG)	Online	5	English	Migrants, Non-EU citizens, and ethnic minorities
4	Academic citations and collaborations	University of Bologna; University of Groningen, Complexity Science Hub Vienna	Online	7	English	Early career researchers belonging to the LGBTQI+ community and/or ethnic and religious minorities

In total, six workshops with marginalized groups were conducted (see Table 1 for details) in Spain (hybrid online), Greece, and online in collaboration with the local civil society organizations representing the vulnerable groups listed above. The Spanish and one of the Greek workshops were facilitated in local languages, and two of the Greek workshops (mainly with non-Greek participants) and the online workshops were conducted in English. Ethical approval for this research was obtained in advance from the research ethics committees (REC) of the Centre for Research and Technology Hellas (CERTH) and the University of Bologna (UNIBO). Researchers were careful to be transparent about the research objectives and to obtain informed consent for participation from participants.

The workshops all shared a similar structure. They started with a brief introduction to the MAMMOth project and the aim of the workshop, along with rules for participation. The first part of these workshops involved the delivery of a training module on AI fairness conducted by the workshop facilitators using training materials designed by science education and communication specialists from RUG. This methodological innovation in the Co-C process precedes evidence-gathering activities with training to equip participants with the explicit conceptual knowledge and terminology to articulate their perspectives on AI. This is both for the purposes of the training and for the broader purpose of empowering participants in their daily lives since they use or have their personal data processed by AI, although they may hitherto be less consciously aware of when and how this occurs. An analogy for this provided by one of the training module designers is explaining the rules and lingo of soccer to someone watching a match for the first time so they can more clearly understand what is happening in the match.

The training material was prepared previously as educational materials designed for public engagement with laypersons on issues concerning AI bias and subsequently adapted for the MAMMOth Co-C workshops. These materials were based on research gathered from focus groups, interviews, and informal conversations with very different audiences about how laypeople use AI and their attitudes and feelings toward it. The material focused on the non-technical aspects of AI and specifically on its socio-cultural implications. The technicalities were intentionally avoided so as not to overwhelm participants with aspects not considered useful for the discussion. The training phase followed an inquiry-based teaching and learning approach [56] aimed at first exploring laypersons' misconceptions about AI—mainly its association with robots and the idea that AI is a neutral and objective technology—providing daily life examples, adding new knowledge that also questions these misconceptions and then checking the level of understanding of participants and the effectiveness of the training. These steps were followed in an iterative manner. The training, due to its interactive nature with the participants, acted as a "workshop in the workshop". The rationale for incorporating training into the Co-C of the MAMMOth tool was to empower representatives from marginalized groups with knowledge of how AI bias in decision-making can impact them and create a common frame of reference by ensuring that every participant had a basic understanding of the topic.

Prior to conducting the workshops, the training developers from RUG conducted a session with the workshop facilitators to familiarize them with the training materials and share knowledge on best practices for delivering the training modules. The material was first prepared in English and then translated into the languages used in the workshops. The training material was intended as a tool that facilitators could then adapt to their purpose and audiences. In this regard, Donna Activa Forum (DAF), the civil society partner of the MAMMOth project working with under-served women, adapted the material to include examples related to gender biases—something the participants of the workshop could relate to. These training components explained in an accessible way how AI decision-making algorithms work and ways in which they can reproduce discrimination, including demonstrating real-life examples.

The second half of the workshops focused on discussing participants' perspectives based on their analysis of finance, identity verification, and academic citations and collaborations using example use-cases prepared by a sociologist from the University of Bologna (Marta Gibin, co-author) to obtain evidence for the design process. For finance, this included a use case based on a bank's use of AI algorithms to evaluate participants' home loan applications. The identity verification use case was based on examples of everyday life applications of this service (e.g., to open a bank account, to rent a car, to subscribe to the services of an Internet provider) and how it can reproduce biases. The academic citations and collaborations use case was introduced by showing possible biases reproduced by Google Scholar when searching for authors. For each use case, participants were invited to think about the potential groups that could be at an elevated risk of experiencing adverse effects as a result of the deployment of an AI system. This was performed in order to identify potential sources of bias. For example, with reference to the use case on identity verification, participants were asked to reflect on different categories of facial attributes that could lead to bias in ID verification systems (e.g., facial decorations such as piercings and tattoos; religious and cultural symbols such as hijabs and turbans; and so forth) and rank them in order of importance. In terms of fairness, participants were encouraged to express their views on the fairness of the use of AI systems in the given examples. If they felt that the use of an AI system was unfair, they were invited to suggest an alternative outcome that would be perceived as fair. This allowed us to explore their conceptions of fairness. Participants' evaluations of the use cases were gathered as research data and subsequently analyzed by a small team of project researchers, including the aforementioned sociologist and a software engineer, to inform the requirements of the MAMMOth toolkit. This process will be elaborated on in the following section.

4.2. Workshop Outcomes

The findings from the individual workshops were varied. In the workshop conducted by DAF, only three participants affirmed that they had previous knowledge about AI. In the workshops conducted by IASIS, most of the participants had little to average knowledge of the topic. When asked to provide a definition of AI, they mentioned robots with human behavior, computer systems, ChatGPT, something complex, algorithms, and something related to math and data. During the workshop, once shown some examples of everyday use of AI systems, they realized that they make use of some form of AI regularly in daily activities. The participants of the workshop organized by the Diversity Development Group (DDG) were, on average, familiar with AI. In particular, they unanimously reported daily interactions with AI through digital tools (Google Maps, Translate, Spotify, etc.), reflecting the pervasive role of AI in modern life.

In their reflections, all workshop facilitators acknowledged the key role of the training module in enabling participants to discuss AI bias in an informed way in workshop discussions. Feedback from facilitators included that the training module provided stimuli for participants to engage in the workshop and raised all participants within a group to a similar level of understanding about AI bias. Similarly, several participants mentioned that they found the training session to be useful. After the presentation, all participants agreed that they are using narrow-purposed AI daily without realizing it. Some participants were surprised to realize that AI financial decision-making applications exist and are being used in some banking systems. Some participants also became more aware of the distinction between e-banking procedures and credit score and loan request procedures as a result of the training.

Participants exhibited nuanced perspectives on the objectivity of AI; some argued that AI systems are inherently biased because of non-objective data, while others believed that objectivity could vary depending on training data. In the workshop conducted by DDG, an interesting consensus emerged in terms of who would benefit the most from AI-based credit scoring: white males. This illustrates prevalent concerns about biases in AI systems. The concept of fairness in AI decisions was interrogated, resulting in a variety of definitions, including unbiased, non-discriminatory, equitable, and transparent. An important consideration was that AI should exhibit contextual awareness, considering individual circumstances. It was also mentioned that there should be a combination of human critical thinking and AI, i.e., a human-in-the-loop, and the overall and final monitoring of impactful decisions should be made by humans; however, participants expressed cynicism regarding the inherent fairness of AI-driven processes and the motives of AI developers, such as in the case of AI-made decisions on bank loan applications where some participants suggested that banks would not priorities fairness over profitability. Furthermore, participants acknowledged that the fairness of AI decisions is contingent on factors such as design, implementation, data quality, the algorithms used, and ethical and technical considerations. As one participant noted, “Statistics is not about fairness, but about the majority and/or probability”.

4.3. User Requirements and Research Requirements

User requirements gathered a set of expectations for what the MAMMOth bias toolkit should be capable of and the parameters of how it should work from the standpoint of end users. The primary end users targeted through the project’s methodology were the data scientists whose systems ended up affecting under-represented groups, among others; however, it was also recognized that affected stakeholders, including members of such groups, are also indirectly affected, and their input was also considered. User requirements were shaped based on the amalgamated results of the Co-C workshops with a pilot survey with under-represented groups, and the expert user needs analysis, which included both data scientists and social scientists. This combined approach aimed at capturing both the suggestions received from experts and the concerns identified by the affected stakeholders that could arise when the toolkit is deployed. The construction of the list of Research requirements followed a similar process. Research requirements refer to additional input, mainly coming from the opinions and concerns of under-represented groups, that should guide the research on biases and mitigation conducted by technical partners of the project.

Each partner conducting a workshop provided the research team with an English language summary of the workshop contents with key points, which were then collated by the research team for comments and validation by consortium partners. The results were also presented at a plenary meeting with the entire consortium in order to be discussed collectively. To translate the social scientific data gathered from the workshop into technical terms, the sociologist prepared a summary of key points and worked with the software engineer to clarify any ambiguous points and reduce the number of requirements by clustering similar points. This was an ongoing and iterative process that negotiated the different terminologies and epistemic approaches used by technical researchers and social scientists working on the project to enable clear communication and the successful formulation of evidence-based user and research requirements. Activities were part of systems design that eventually led to the creation of user story mockups and eventually to a modular open-source architecture whose individual modules deploy algorithmic strategies for bias detection and mitigation. Figure 1 illustrates the MAMMOth methodology designed to collect, utilize, and leverage the requirements for an AI bias toolkit. This underscores the systematic efforts to ensure the effectiveness and reliability of the toolkit in reducing AI biases through the implementation of collaborative Co-C processes.

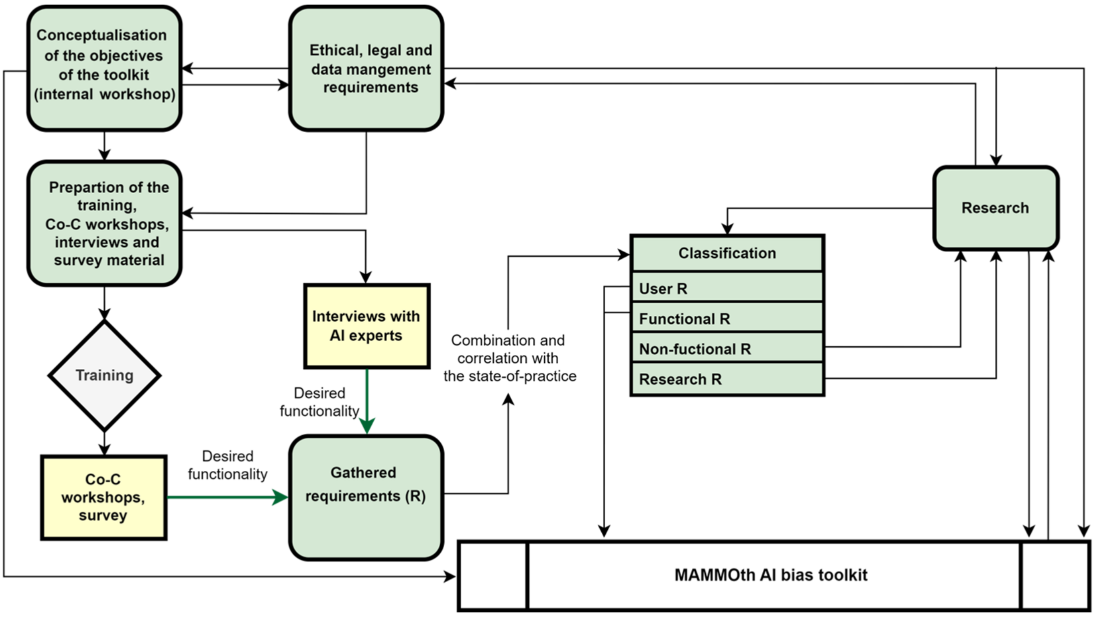
Conceptualisation of the objectives of the toolkit (internal workshop)
Ethical, legal and data mangement requirements
Prepartion of the training, Co-C workshops, interviews and survey material
Research
Classification
User R
Functional R
Non-fuctional R
Research R
Interviews with AI experts
Combination and correlation with the state-of-practice
Training
Desired functionality
Desired functionality
Co-C workshops, survey
Gathered requirements (R)
MAMMOth AI bias toolkit

Figure 1. MAMMOth approach to gather and utilize the requirements for the AI bias toolkit.

The next section will provide a discussion that draws from critical perspectives on PD and Co-C to reflect on the value, methodological considerations, and challenges of integrating training in stakeholder Co-C processes on AI bias, particularly in relation to working with marginalized and vulnerable populations.

5. Discussion

5.1. Reflections on Designing and Integrating Training in Co-C

Incorporating training into AI Co-C requires careful planning and implementation, including clarifying expectations and assumptions about participants' exposure to, engagement with, and prior knowledge about AI. RUG designed the training with the assumption that participants would have little prior knowledge of AI technology and that any knowledge they did have would be limited to utopian/dystopian cultural representations of AI. This was corroborated by at least one facilitator's experience that most participants had an overall low level of knowledge about AI, although several participants did have some introductory level of knowledge, and at least one had some experience using a generative AI application. With this in mind, RUG took a deliberate strategy of ensuring the training was designed and presented in a way that participants required zero prior technical knowledge to understand the material and that the amount of technical content included was kept as minimal as possible. The rationale for this approach is that a socio-technical approach would be relatable to participants, whereas an emphasis on technical knowledge would likely alienate participants and cause them to 'switch off' and thus have the opposite effect on the training's goal of encouraging engagement and inclusion.

Upon reflecting on the training, designers did question whether they underestimated participants' capacity for engaging with more technical content and what different outcomes or understandings this might enable. At the same time, it was clear that some workshop target communities, e.g., a group of older unemployed women in their 40s and 50s, struggled more with understanding the content than other groups. This leads to two recommendations: (1) additional research should be conducted on the relationship between greater technical knowledge and enhanced socio-technical understanding of systemic factors contributing to AI bias by training participants; (2) provided the increased inclusion of technical content can contribute to greater understanding and empowerment of contributing factors to AI bias, the extent of the inclusion of technical content should be reflected upon and calibrated in relation to both the participants needs and capacities and the objectives of the Co-C process.

Facilitators also independently adapted the workshops in order to be relevant to the experiences and needs of participants, particularly in the specific contexts of the civil society organizations conducting the Co-C workshops. This included translating the workshop materials into local languages where necessary and adding real-life examples and use cases relevant to the participants. DAF added an example based on generative AI queries on the responsibilities of fathers and mothers that reproduced outdated and stereotypical gender roles. Another example they added to the training demonstrated how Google Translate reproduced stereotypical gender masculine and feminine endings in translating text from Catalan to Spanish, whereby masculine and feminine versions of the job role were both provided for short answers, but in longer answers, stereotypical gender language was used (e.g., 'judge' had male ending, 'nurse' had female ending). The fact that these examples were real-life as opposed to hypothetical was also important, as it added authenticity to the significance and pervasiveness of AI technology bias, as well as its often oblique and implicit nature.

There was also critical engagement from the facilitators on the representation of AI in the training. One trainer was concerned that the training materials framed AI technology in an exclusively negative way that ignores the potential benefits or instrumental value of AI technology, including its potential to enhance opportunities and mitigate discrimination if applied conscientiously towards these ends. They found this to be a particularly acute issue when working with the older unemployed women participant group, as it felt outside of this group's sphere of interest and understanding about potential benefits and harms beyond automation and AI 'taking jobs' to explore more nuanced aspects of the technology's societal implications, including its potential benefits if used in a fair and ethical way; however, the facilitators from other organizations disagreed that the training had an overly negative orientation, but instead focused on raising awareness of the pitfalls of AI technology in its current trajectory of innovation and on promoting empowerment. They agreed that it would be interesting to explore in more detail how AI bias can be mitigated in order to promote a fairer AI.

5.2. The Methodological Value of Incorporating Training in Co-C

As discussed in Section 4.2, the purpose of incorporating training in the Co-C workshops was to empower participants from marginalized communities with knowledge of what AI is, how it affects their lives, and how AI bias in decision-making can cause them harm. As commonly agreed by the workshop facilitators, the training module provided a smooth kick-start for the co-creation process by ensuring that all participants had a shared understanding of the topic. In this way, the training module democratized the Co-C process by enabling the participation of the most disadvantaged participants in the discussion.

Participants, even though they came from different backgrounds and had different levels of AI knowledge, could participate in the discussion almost on an equal footing. As one of the workshop facilitators commented, conducting a Co-C workshop without a training module would not be possible. The training also helped raise a general awareness of AI bias and how it affects participants' everyday lives. A jargon-free, non-technical and example-based training module enabled participants to understand how AI has penetrated almost every area of their lives and how AI bias could affect them.

Figure 2 provides a visual summary of how the training contributed to the Co-C process and this, in turn, to the requirements. The training formed the basis of the Co-C process and contributed to it in three ways. First of all, it provided clarity about AI and how it works and the level of understanding of the training material was iteratively checked by the trainers through questioning and interaction with the participants. The training also increased the level of awareness of the participants on the use of AI in everyday life and challenged misconceptions related to AI; after being provided with real-life examples, the participants were able to recognize when they use AI in their lives. The training also contributed to the activation of critical thinking by providing examples of possible forms of biases and discrimination reproduced by AI systems, questioning overly optimistic narratives regarding the use of AI systems.

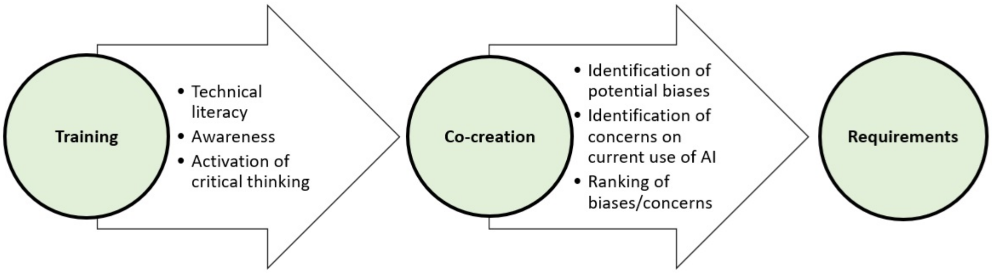
Training
• Technical literacy
• Awareness
• Activation of critical thinking
Co-creation
• Identification of potential biases
• Identification of concerns on current use of AI
• Ranking of biases/concerns
Requirements

Figure 2. Contribution of the training and the Co-C process to the requirements.

The results of the Co-C process provided three types of feedback to the MAMMOth requirements. During the workshops, the participants identified the potential sources of bias related to the use of AI in the MAMMOth use cases. For example, when discussing the use case of academic citations and collaborations, several potential biases were highlighted, such as self-citations, multiple citation counts due to the same document being published on different platforms, popular authors who are consistently cited, and the lack of representation of the diversity and multi-disciplinarity of a research community. The workshops also highlighted the participants' concerns about the use of AI. For example, in the finance use case, participants expressed concerns about the potential dehumanization of the relationship between banks and customers following the use of AI. Moreover, similar to the ID verification use case, participants were asked to rank potential forms of bias/concerns in order of importance, helping the researchers to identify what was most important to them.

In addition, it is important to acknowledge the diverse forms of value that the training can bring to different stakeholders in the Co-C process, including not only the participants but also the facilitators and their organizations, researchers, and technology designers. A reflective critique from the MAMMOth Co-C practitioners suggests that harnessing value for all of these stakeholders might require integrating training in Co-C as an ongoing and iterative process rather than a one-shot activity. Workshop participants expressed in their feedback on the sessions a desire to learn more about the societal implications of AI after participating in the session. In addition, one of the facilitators decided to change her psychology master's degree dissertation topic to AI anxiety after the learning process she underwent for leading her organization's sessions. Both she and the other trainers stated that they desired greater involvement in an ongoing process and would like to have been kept informed of the training outcomes and the impacts generated by the evidence they gathered.

5.3. Challenges and Complexities of Integrating Training in Co-C

Incorporating training into Co-C processes raises a number of methodological and epistemological questions. The most crucial is whether incorporating training to impart specific sets of knowledge or empower social outcomes based on notions of empowerment or social justice influences the responses, findings, and outcomes of the Co-C process. This also has implications for debates around the boundaries between research and activism [57–59] and whether the deliberate merging of the two is desirable or problematic [60–62]. In the case of the Co-C process for MAMMOth, civil society partners with defined activist missions were responsible for delivering the training and the Co-C workshops that fed into the production of user and research requirements. Here, the justification for merging research with activism is a deliberate and direct strategy to attain the MAMMOth project's goal of mitigating discrimination and adverse social outcomes for vulnerable groups that result from the perpetuation of biases in AI data training and processing.

The critical approaches to PD in AI technology development discussed in this paper's literature review challenge the extent to which these processes, as they are often conducted, are, in fact, a 'light touch' exercise to claim inclusive representation that has little transformative value in producing more fair and inclusive innovation. Similarly, PD and Co-D can be a crucial factor in supporting the implementation of the applicable legal requirements, which aim to address AI bias, and in demonstrating legal compliance as well, as discussed in Section 3. Through the presentation of the EU legal and ethical framework for AI, the essential importance of compliance through inclusivity, stakeholders' involvement, training, and, in general, the promotion of AI literacy is emphasized, particularly in the case of AI bias. This approach can verify that AI system creators do not merely fulfill formal requirements but also heed the insights of social scientists regarding true fairness and social transformation. This raises questions about how incorporating training can become a catalyst for more profound outcomes for a socially just and ethical digital transformation.

Alongside this, critical perspectives in the literature also indicate a lack of critical reflexivity on training in PD with vulnerable groups [41]. Incorporating training in itself does not automatically lead to more fair or just outcomes, but rather, the training must be incorporated in a reflective and considered way in terms of both the selected content, mode of delivery and context in which it is delivered, as well as how it integrates with the focus and purpose of the Co-C/PD exercise in which it takes place. In the design process for MAMMOth, attention was paid by the training designers not only to support the development of the MAMMOth tool but also to empower participants to understand AI better and express a voice or opinion on how AI affects their lives. This was furthered in how particular civil society organizations conducting the training added real-world examples to the training that were targeted to engage the specific groups with which their organizations worked.

Another related aspect that requires reflective consideration in planning and executing training in the Co-C process is negotiating the balance between imparting the necessary technical depth to obtain meaningful input from participants, and ensuring clarity and understanding of the training content so it is accessible and actionable for participants. In the initial training design, a deliberate decision was taken to minimize detailed technical information in order to avoid overwhelming or confusing participants. Alongside this, real-life examples were selected and emphasized as a strategy to make the training content engaging and understandable to non-specialists without relying on complex technical concepts.

Reflective analysis also requires engaging with both the limitations of incorporating training in AI Co-C with vulnerable groups and other stakeholder categories, as well as the way it was implemented in the MAMMOth Co-C design. Due to the comparatively small sample size of the MAMMOth CO-C process, it was not possible for researchers to collect statistically significant or reliable quantitative data to support their analysis. In addition, in the MAMMOth Co-C, the workshops were conducted on a one-off basis with each stakeholder group. As a result, the potential for further dialogue and iterative development was missed. Both workshop participants and other stakeholders (facilitators and researchers) expressed a desire to learn more about the topic of AI bias and to remain engaged in the design process to see the impact of their input on the research and technology design outcomes from the project. This demonstrates that to achieve better integration, researchers and facilitators should integrate training as an ongoing process-within-a-process in technology Co-C. This entails planning for multiple workshops or sessions throughout the design process, starting from an early stage, and at each stage modifying the design and adapting the Co-C approach in response to participants' input in a way that demonstrates to participants that their voices are being listened to and are contributing to meaningful social change.

On a broader scale, it is reasonable to question the extent to which training within Co-C activities for AI technology development actually disrupts inherent power dynamics or whether training risks adding an additional layer of ostensible credibility to 'ethics-washing' practices [63] or 'thin' compliance tick-box exercise that fails to influence meaningful change in identifying and mitigating vectors of discrimination and other forms of social inequality. This can be addressed in the process of how the training material is prepared, whereby it is produced through participatory methods in which potential groups to be trained have an opportunity to shape the material. In the case of MAMMOth, to address this challenge, training material was prepared along with a consultation process with a wide range of stakeholders, which included a number of interviews and focus groups. This counteracts the pitfall of material being produced in a top-down way where mainly technology developers' interests are reflected.

Using training in Co-C can also support the implementation of EU ethical and legal requirements by the AI systems creators, particularly concerning the requirements outlined in the recently enacted EU AI Act. The provision of education and training, especially to those groups affected by AI bias, can supply system creators with real-life scenarios and facilitate the effectiveness of stakeholders' involvement in the implementation of the AI Act [55]; however, there is a risk of training in co-design, along with other co-creation best practices, to be misused for 'ethics-washing' or leading to regulatory compliance exercises that fail to meaningfully address the discriminatory effects they are meant to mitigate. As such, the AI literacy of subjects should be considered an essential factor in contributing to informed consent for their participation in AI co-design since the conditions under which informed consent is obtained in many cases require a certain level of knowledge about the functionality of AI technology and its technical context. Thus, providing guidance and training that addresses AI bias and discrimination could contribute to the AI literacy of affected stakeholders. Furthermore, effective legal compliance that provides substantive protection against AI bias requires the clear identification and careful examination of protected grounds directly related to the context of the technology under investigation, for instance, when conducting the Fundamental Rights Impact Assessment (FRIA) under Article 27 of the AI Act.

6. Conclusions

This paper draws from both applied and critical approaches to PD and Co-C to argue for the inclusion of training on AI bias in co-creation processes with non-expert stakeholder groups whose input would otherwise be excluded from AI design and integration decisions, particularly those from vulnerable or marginalized groups. Incorporating training can enhance the Co-C process in several ways: it supports participants with the conceptual frameworks and lexicon that can allow them better to articulate their opinions, observations, questions, and concerns; it injects both shared understandings and specific examples that problematize abstract concepts and controversies in a concrete way to stimulate active discussion and engagement in Co-C activities; it empowers end-users to be more aware of how AI bias can impact them and to recognize (often obfuscated) sites where AI is being used and where there is a risk of discrimination occurring. Training in Co-C can also provide a mechanism in conjunction with other technology design and implementation best practices for demonstrating compliance with emerging regulatory and ethical regimes, such as the EU's AI Act.

Whilst the focus of this article was on training in Co-C for addressing AI bias and discrimination, training in Co-C could almost certainly be applied to benefit Co-C processes in other complex areas of technology design with high-stakes social, political and economic implications; however, for such training to work in a way that both empowers stakeholders and supports the objectives of the Co-C process, including establishing regulatory compliance, it is necessary to design and conduct training in a critically engaged and reflexive way. This requires training designers and facilitators to engage with the particular contexts of specific stakeholder groups by differentiating the language, content, medium, and pedagogical approach of delivery to be relevant and accessible to that group. It also requires careful consideration and choice-making on how to approach topics of technical complexity in an accessible way, and calibrating the extent to which the training content engages with technical concepts and details (or avoids them altogether). This sometimes requires finding techniques to communicate how a technology works and its socio-technical implications without resorting to technical knowledge. Experiential findings from the MAMMOth Co-C process also indicate the importance of emphasizing real-world examples that are relevant to the context, experience and potential impacts for the participant group collaborating in the Co-C process.

Author Contributions: Conceptualization: I.S., E.Y., S.R., M.G., E.K. and S.P.; methodology, M.G., E.K. and S.P.; software: E.K., S.P. and M.G.; investigation: I.S., E.Y., S.R., M.G. and E.K.; writing—original draft preparation: I.S., E.Y., S.R., M.G. and E.K.; writing—review and editing: I.S., S.P., S.R., M.G. and E.K.; visualization, S.R., M.G. and I.S.; supervision: S.P.; project administration: S.R. and S.P.; funding acquisition: S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by the European Union under the Horizon Europe MAMMOth project, Grant Agreement ID: 101070285. UK participant in Horizon Europe Project MAMMOth is supported by UKRI grant number 10041914 (Trilateral Research LTD).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics Committees of the Centre for Research & Technology, Hellas (CERTH) (protocol number 014787, approved on 10 March 2023) and the University of Bologna (protocol number 0058267, approved on 3 March 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: MAMMOth Deliverable D1.1 User Requirements and Architecture is openly available here: https://mammoth-ai.eu/wp-content/uploads/2023/11/d1.1-v1.0.pdf. MAMMOth training materials produced by the University of Groningen are openly available here: https://mammoth-ai.eu/training-material/. The Mammoth Fairness Definition Guide and Toolkit are openly available on the Github repository here: https://github.com/mammoth-eu.

Acknowledgments: The authors would like to acknowledge the contribution of Dagmar Heeg from the University of Groningen, who produced the AI bias training materials that were adapted for the MAMMOth co-creation workshops, and who shared her insights on producing and adapting the training materials. The authors also wish to thank Anna Pujol-Ferran and Virginia Carreras from Donna Activa Forum and Elli Nikolakopoulou from IASIS, who facilitated some of the co-creation workshops, for their engaged reflections on the MAMMOth co-creation process.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA Relevance. 2024. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 16 July 2024).
2. UNESCO Recommendation on the Ethics of Artificial Intelligence. 2021. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed on 20 November 2024).

3. Directorate-General for Communications Networks, Content and Technology (European Commission). *High-Level Expert Group on Artificial Intelligence Ethics Guidelines for Trustworthy AI*; Publications Office of the European Union: Luxembourg, 2019; ISBN 978-92-76-11998-2. Available online: https://data.europa.eu/doi/10.2759/346720 (accessed on 16 July 2024).
4. OECD. Recommendation of the Council on Artificial Intelligence. 2024. Available online: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed on 18 November 2024).
5. Sanders, E.B.-N.; Stappers, P.J. Co-Creation and the New Landscapes of Design. *CoDesign* **2008**, *4*, 5–18. [CrossRef]
6. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in Data-driven Artificial Intelligence Systems—An Introductory Survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1356. [CrossRef]
7. Information Commissioner's Office (ICO) Guidance on AI and Data Protection. Available online: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/ (accessed on 16 July 2024).
8. Access Now. *Human Rights in the Age of Artificial Intelligence*; Access Now: New York, NY, USA, 2018; Available online: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf (accessed on 31 July 2024).
9. Latonero, M. Governing Artificial Intelligence. Available online: https://datasociety.net/library/governing-artificial-intelligence/ (accessed on 31 July 2024).
10. Muller, C. *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*; Ad Hoc Committee on Artificial Intelligence (CAHAI), Council of Europe: Strasbourg, France, 2020; Available online: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da (accessed on 31 July 2024).
11. *Bias—The TAILOR Handbook of Trustworthy AI*. Available online: http://tailor.isti.cnr.it/handbookTAI/Diversity_Non-Discrimination_and_Fairness/bias.html (accessed on 30 August 2024).
12. Bødker, S.; Dindler, C.; Iversen, O.S.; Smith, R.C. What Is Participatory Design? In *Participatory Design*; Bødker, S., Dindler, C., Iversen, O.S., Smith, R.C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 5–13. ISBN 978-3-031-02235-7.
13. Rose, E.J. Design as Advocacy: Using a Human-Centered Approach to Investigate the Needs of Vulnerable Populations. *J. Tech. Writ. Commun.* **2016**, *46*, 427–445. [CrossRef]
14. Jo, A. *Participatory Research: Working with Vulnerable Groups in Research and Practice*; Policy Press: Bristol, UK, 2016; ISBN 978-1-4473-2556-7.
15. Limantė, A.; Tereškinas, A. Definition of Vulnerable Groups. In *Legal Protection of Vulnerable Groups in Lithuania, Latvia, Estonia and Poland: Trends and Perspectives*; Limantė, A., Pūraitė-Andrikienė, D., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 3–27. ISBN 978-3-031-06998-7.
16. Ippolito, F.; Sánchez, S.I. Introduction. In *Protecting Vulnerable Groups: The European Human Rights Framework*; Ippolito, F., Sánchez, S.I., Eds.; Bloomsbury Publishing: London, UK, 2015; pp. 1–22. ISBN 978-1-78225-613-7.
17. Laura Ramírez Galleguillos, M.; Coşkun, A. How Do I Matter? A Review of the Participatory Design Practice with Less Privileged Participants. In Proceedings of the 16th Participatory Design Conference 2020—Participation(s) Otherwise, Manizales, Colombia, 15–20 June 2020; Association for Computing Machinery: New York, NY, USA, 2020; Volume 1, pp. 137–147.
18. Spinuzzi, C. The Methodology of Participatory Design. *Tech. Commun.* **2005**, *52*, 163–174.
19. Luck, R. What Is It That Makes Participation in Design Participatory Design? *Des. Stud.* **2018**, *59*, 1–8. [CrossRef]
20. Charlotte Smith, R.; Winschiers-Theophilus, H.; Loi, D.; Paula Kambunga, A.; Muudeni Samuel, M.; De Paula, R. Decolonising Participatory Design Practices: Towards Participations Otherwise. In Proceedings of the 16th Participatory Design Conference 2020—Participation(s) Otherwise, Manizales, Colombia, 15–20 June 2020; ACM: Manizales, Colombia, 2020; Volume 2, pp. 206–208.
21. Bowler, L.; Wang, K.; Lopatovska, I.; Rosin, M. The Meaning of "Participation" in Co-Design with Children and Youth: Relationships, Roles, and Interactions. *Proc. Assoc. Inf. Sci. Technol.* **2021**, *58*, 13–24. [CrossRef]
22. Hussain, S. Empowering Marginalised Children in Developing Countries through Participatory Design Processes. *CoDesign* **2010**, *6*, 99–117. [CrossRef]
23. Buddemeyer, A.; Nwogu, J.; Solyst, J.; Walker, E.; Nkrumah, T.; Ogan, A.; Hatley, L.; Stewart, A. Unwritten Magic: Participatory Design of AI Dialogue to Empower Marginalized Voices. In Proceedings of the 2022 ACM Conference on Information Technology for Social Good, Limassol, Cyprus, 7–9 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 366–372.
24. Ellard-Gray, A.; Jeffrey, N.K.; Choubak, M.; Crann, S.E. Finding the Hidden Participant: Solutions for Recruiting Hidden, Hard-to-Reach, and Vulnerable Populations. *Int. J. Qual. Methods* **2015**, *14*, 1609406915621420. [CrossRef]
25. Sadler, G.R.; Lee, H.-C.; Lim, R.S.-H.; Fullerton, J. Research Article: Recruitment of Hard-to-Reach Population Subgroups via Adaptations of the Snowball Sampling Strategy. *Nurs. Health Sci.* **2010**, *12*, 369–374. [CrossRef] [PubMed]
26. Lee, M.K.; Kim, J.T.; Lizarondo, L. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; ACM: Denver, CO, USA, 2017; pp. 3365–3376.
27. Park, H.; Ahn, D.; Hosanagar, K.; Lee, J. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–22.

28. Koene, A.; Perez, E.; Ceppi, S.; Rovatsos, M.; Webb, H.; Patel, M.; Jirotka, M.; Lane, G. Algorithmic Fairness in Online Information Mediating Systems. In Proceedings of the 2017 ACM on Web Science Conference, Troy, NY, USA, 25–28 June 2017; ACM: Troy, NY, USA, 2017; pp. 391–392.
29. Starke, C.; Baleis, J.; Keller, B.; Marcinkowski, F. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Big Data Soc.* **2022**, *9*, 20539517221115189. [CrossRef]
30. Zhang, A.; Walker, O.; Nguyen, K.; Dai, J.; Chen, A.; Lee, M.K. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proc. ACM Hum. Comput. Interact.* **2023**, *7*, 1–32. [CrossRef]
31. Wang, Q.; Madaio, M.; Kane, S.; Kapania, S.; Terry, M.; Wilcox, L. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; ACM: Hamburg, Germany, 2023; pp. 1–16.
32. Madaio, M.A.; Stark, L.; Wortman Vaughan, J.; Wallach, H. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; ACM: Honolulu, HI, USA, 2020; pp. 1–14.
33. Lancaster, C.M.; Schulenberg, K.; Flathmann, C.; McNeese, N.J.; Freeman, G. "It's Everybody's Role to Speak Up... But Not Everyone Will": Understanding AI Professionals' Perceptions of Accountability for AI Bias Mitigation. *ACM J. Responsible Comput.* **2023**, *1*, 1–30. [CrossRef]
34. Long, D.; Jacob, M.; Magerko, B. Designing Co-Creative AI for Public Spaces. In Proceedings of the 2019 Conference on Creativity and Cognition, San Diego, CA, USA, 23–26 June 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 271–284.
35. Gerling, K.; Vinel, A.; Müller, K.; Nierling, L.; Stiefelhagen, R.; Karmann, C.; Lang, D.; Asfour, T. Technology-Centric Real-World Labs: Challenges and Opportunities for a New Mode of Participatory Research From the Perspective of Computer Science. Mensch und Computer 2023—Workshopband, 2023.
36. Ito-Jaeger, S.; Lane, G.; Dowthwaite, L.; Webb, H.; Patel, M.; Rawsthorne, M.; Portillo, V.; Jirotka, M.; Perez Vallejos, E. TrustScapes: A Visualisation Tool to Capture Stakeholders' Concerns and Recommendations About Data Protection, Algorithmic Bias, and Online Safety. *Int. J. Qual. Methods* **2023**, *22*, 1–10. [CrossRef]
37. Liao, Q.V.; Muller, M. Enabling Value Sensitive AI Systems through Participatory Design Fictions. 2019. Available online: https://arxiv.org/abs/1912.07381 (accessed on 6 March 2024).
38. Zicari, R.V.; Ahmed, S.; Amann, J.; Braun, S.A.; Brodersen, J.; Bruneault, F.; Brusseau, J.; Campano, E.; Coffee, M.; Dengel, A.; et al. Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Front. Hum. Dyn.* **2021**, *3*, 1–20. [CrossRef]
39. Cetina, K.K. *Epistemic Cultures: How the Sciences Make Knowledge*; Harvard University Press: Cambridge, MA, USA, 1999; ISBN 978-0-674-03968-1.
40. Weinberg, L. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *J. Artif. Intell. Res.* **2022**, *74*, 75–109. [CrossRef]
41. Donia, J.; Shaw, J.A. Co-Design and Ethical Artificial Intelligence for Health: An Agenda for Critical Research and Practice. *Big Data Soc.* **2021**, *8*, 20539517211065248. [CrossRef]
42. Donia, J.; Shaw, J. Co-Design and Ethical Artificial Intelligence for Health: Myths and Misconceptions. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; p. 77.
43. Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M.C.; Gabriel, I.; Mohamed, S. Power to the People? Opportunities and Challenges for Participatory AI. In Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization, Arlington, VA, USA, 6–9 October 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–8.
44. Delgado, F.; Yang, S.; Madaio, M.; Yang, Q. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Boston, MA, USA, 30 October–1 November 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–23.
45. Delgado, F.; Yang, S.; Madaio, M.; Yang, Q. Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir". In Proceedings of the 35th Conference on Neural Information Processing Systems, Sydney, Australia, 1 November 2021.
46. Bondi, E.; Xu, L.; Acosta-Navas, D.; Killian, J.A. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 425–436.
47. Hossain, S.; Ahmed, S.I. Towards a New Participatory Approach for Designing Artificial Intelligence and Data-Driven Technologies. 2021. Available online: https://arxiv.org/abs/2104.04072 (accessed on 27 November 2024).
48. Suresh, H.; Movva, R.; Dogan, A.L.; Bhargava, R.; Cruxen, I.; Cuba, A.M.; Taurino, G.; So, W.; D'Ignazio, C. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; ACM: Seoul, Republic of Korea, 2022; pp. 667–678.
49. Katell, M.; Young, M.; Dailey, D.; Herman, B.; Guetler, V.; Tam, A.; Bintz, C.; Raz, D.; Krafft, P.M. Toward Situated Interventions for Algorithmic Equity: Lessons from the Field. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; ACM: Barcelona, Spain, 2020; pp. 45–55.

50. Hornung, D.; Müller, C.; Shklovski, I.; Jakobi, T.; Wulf, V. Navigating Relationships and Boundaries: Concerns around ICT-Uptake for Elderly People. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 7057–7069.
51. Otto, P.N.; Anton, A.I. Addressing Legal Requirements in Requirements Engineering. In Proceedings of the 15th IEEE International Requirements Engineering Conference (RE 2007), New Delhi, India, 15–19 October 2007; pp. 5–14.
52. Kerrigan, S.; Law, K.H. Logic-Based Regulation Compliance-Assistance. In Proceedings of the 9th International Conference on Artificial Intelligence and Law, Edinburgh, UK, 24–28 June 2003; Association for Computing Machinery: New York, NY, USA, 2003; pp. 126–135.
53. Directorate-General for Communications Networks, Content and Technology (European Commission). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*; Publications Office of the European Union: Luxembourg, 2020; ISBN 978-92-76-20008-6.
54. Krasanakis, E.; Gibin, M.; Rizou, S. AI Fairness Definition Guide 2024. Available online: https://github.com/mammoth-eu/FairnessDefinitionGuide/blob/master/AI%20Fairness%20Definition%20Guide.pdf (accessed on 16 July 2024).
55. Novelli, C.; Hacker, P.; Morley, J.; Trondal, J.; Floridi, L. A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. *Eur. J. Risk Regul.* **2024**. [CrossRef]
56. Kahn, P.; O'Rourke, K. Understanding Enquiry-Based Learning. In *Handbook of Enquiry & Problem Based Learning*; CELT: Galway, Republic of Ireland, 2005.
57. Young, A.M.; Battaglia, A.; Cloud, D.L. (UN)Disciplining the Scholar Activist: Policing the Boundaries of Political Engagement. *Q. J. Speech* **2010**, *96*, 427–435. [CrossRef]
58. Kende, A. Separating Social Science Research on Activism from Social Science as Activism. *J. Soc. Issues* **2016**, *72*, 399–412. [CrossRef]
59. Choudry, A. Reflections on Academia, Activism, and the Politics of Knowledge and Learning. *Int. J. Hum. Rights* **2020**, *24*, 28–45. [CrossRef]
60. Lehtiniemi, T.; Ruckenstein, M. The Social Imaginaries of Data Activism. *Big Data Soc.* **2019**, *6*, 2053951718821146. [CrossRef]
61. Gilson, L. Activism versus Criticism? The Case for a Distinctive Role for Social Critics. *Am. Polit. Sci. Rev.* **2024**, *118*, 862–875. [CrossRef]
62. Gillan, K.; Pickerill, J. *Research Ethics and Social Movements: Scholarship, Activism and Knowledge Production*; Routledge: London, UK, 2016; ISBN 978-1-317-58602-9.
63. Wagner, B. Ethics as an Escape from Regulation: From "Ethics-Washing" to Ethics-Shopping? In *Being Profiled*; Bayamlioğlu, E., Baraliuc, I., Janssens, L., Hildebrandt, M., Eds.; COGITAS ERGO SUM: 10 Years of Profiling the European Citizen; Amsterdam University Press: Amsterdam, The Netherlands, 2018; pp. 84–89.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.