Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits

Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits

Malak Sadek
m.sadek21@imperial.ac.uk
Imperial College London
London, UK

Marios Constantinides
marios.constantinides@nokia-bell-labs.com
Nokia Bell Labs
Cambridge, UK

Daniele Quercia
daniele.quercia@nokia-bell-labs.com
Nokia Bell Labs
Cambridge, UK

Céline Mougenot
c.mougenot@imperial.ac.uk
Imperial College London
London, UK

ABSTRACT

Value Sensitive Design (VSD) is a framework for integrating human values throughout the technology design process. In parallel, Responsible AI (RAI) advocates for the development of systems aligning with ethical values, such as fairness and transparency. In this study, we posit that a VSD approach is not only compatible, but also advantageous to the development of RAI toolkits. To empirically assess this hypothesis, we conducted four workshops involving 17 early-career AI researchers. Our aim was to establish links between VSD and RAI values while examining how existing toolkits incorporate VSD principles in their design. Our findings show that collaborative and educational design features within these toolkits, including illustrative examples and open-ended cues, facilitate an understanding of human and ethical values, and empower researchers to incorporate values into AI systems. Drawing on these insights, we formulated six design guidelines for integrating VSD values into the development of RAI toolkits.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; *HCI design and evaluation methods*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

ethical AI, responsible AI, value sensitive design, toolkits

ACM Reference Format:
Malak Sadek, Marios Constantinides, Daniele Quercia, and Céline Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642810

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642810

1 INTRODUCTION

The increase of risks associated with Artificial Intelligence (AI) systems [11, 25] have led to a surge in the development of toolkits aimed at facilitating the practical design of Responsible AI (RAI) [88]. RAI advocates for the responsible design, development, and use of AI systems, aligning with values like fairness and transparency [79]. A framework with similar objectives is Value Sensitive Design (VSD), recognised for creating more human-centred AI systems [82]. This study hypothesises that VSD can effectively guide the creation of RAI toolkits, given its consideration of human values in the technology design process. However, the extent of alignment between VSD and RAI values and how existing RAI toolkits incorporate VSD remain unclear. Two critical aspects for supporting VSD involve: *a)* enabling stakeholder inclusion and collaboration in order to elicit and understand their values [49, 72], and *b)* facilitating the education of toolkit users in order to promote self-reflexivity and responsible decision-making [31, 60, 82].

This paper focuses on the the practical application of VSD in RAI toolkits, in two Research Questions (RQs):

- **RQ$_1$:** How closely do Value Sensitive Design (VSD) values align with Responsible AI (RAI) values?
- **RQ$_2$:** How do existing RAI toolkits incorporate VSD, and support collaboration and learning?

To address these questions, we conducted an empirical investigation through workshops with 17 participants (AI researchers) (§3). The workshops focused on the expression of VSD values within RAI toolkits, and the impact of toolkit design features on participants' perceptions of stakeholder collaboration and learning.

The contributions of this study are threefold: (i) **mapping VSD values** onto commonly used RAI values integrated into the toolkits, revealing a high degree of alignment between the two sets of values evidenced by consensus among workshop participants (§4.1), (ii) **identifying key links** between design features in RAI toolkits and their impact on promoting VSD. This included: navigation methods supporting iteration, open-ended cuing supporting collaboration versus solo work, examples and case studies providing learning opportunities, and value incorporation reducing cognitive load (§4.1), and (iii) **formulating six practical design recommendations** for enhancing value sensitivity in RAI toolkits (§5). These recommendations, focusing on concrete design features such as supporting actionability and shared knowledge, complement recent broader suggestions for the focuses and approaches of RAI toolkits [23].

2 BACKGROUND AND RELATED WORK

2.1 Value Sensitive Design and Alternative Frameworks

VSD is a theoretical design framework which advocates for the elicitation and inclusion of stakeholders' values in technology [6, 26]. The framework outlines three types of investigations that allow this to happen: conceptual investigations focusing on identifying relevant stakeholders and understanding their context, empirical investigations aimed at understanding stakeholders' needs and values, and technical investigations to reflect on how the technology being created can enable or violate these values [5, 26]. The framework of VSD is supportive of examining the role of values in emerging technologies [3] and in practice [29], and embedding values in design collaborations [90]. While there is strong support for eliciting contextual values directly from stakeholders for a given project [49], VSD also offers a list of "human values often implicated in system design" [26] [p.17] to consider. Figure 1 shows these values and their definitions exactly as stated by [26] in a Miro board.

While several other ethical and value-based frameworks exist, such as Utilitarian or Egalitarian Ethics [15], approaches where many people's values are aggregated, consolidated, or alternated have been recommended by experts in the context of AI systems [15, 30]. When coupled with (i) the consideration of technology generally [83], and algorithms specifically, as "value-laden artefacts" [52], (ii) the tendency to focus solely on economic values in AI systems and the need to consider broader human values [80], and (iii) the ability of VSD to promote self-reflexivity in AI practitioners [82] and practically bootstrap onto existing design processes for AI systems [81, 82], it becomes clear that VSD is an especially suitable framework to consider during AI system development. While design processes [68, 93], guideline [7] and methodologies [84] for value-sensitive AI are emerging, the explicit focus on VSD when designing RAI toolkits remains limited.

2.2 Responsible AI Toolkits

The space of theoretical interventions for responsible AI, such as guidelines and recommendations, while crucial, is quickly becoming overwhelmingly saturated [40]. Recent criticisms have expressed concern at the growing gap between theoretical interventions and the practical implementations of AI systems [14, 17, 21, 40, 50]. Theoretical frameworks are being viewed as too abstract [33], difficult to practically interpret [85], ineffective at resolving conflicts [65], offering little guidance [87], immeasurable in terms of their impact [32], hindering accountability [55, 61], and unimpactful on practitioners [54].

As a result, there has been a shift towards practical tools and processes to guide the implementation of AI systems. These come in several forms such as software [44] and design methods [84], activities [18], and toolkits [59]. The aim of practical interventions is to translate theoretical concepts and frameworks into a tangible, digestible form that practitioners can utilise within their workflows. Despite the basis of most of these practical interventions, and especially toolkits, on theoretical frameworks, very little work has been done to assess the extent to which they effectively operationalise the core concepts of those frameworks.

2.2.1 Forms and Mediums. RAI toolkits originate from both scholarly and industry-based sources. MIT's AI Blindspot cards [12] and the Digital Impact toolkit by Stanford Digital Civil Society Lab [43] are two examples of academic contributions, while Microsoft's Judgement Call cards [56] and Nokia AI Design toolkit [16] are examples of industrial contributions. These examples illustrate the variety of approaches used when designing RAI toolkits for both content and delivery medium. *In terms of content*, while both MIT's AI Blindspot cards and Microsoft's Judgement Call are decks of cards, their content serves different purposes. Microsoft's cards aim to foster empathy in practitioners through gamefication, whereas MIT's cards aim to educate practitioners on how to identify and address potential blindspots while building AI systems by providing examples, recommendations, and stakeholders to engage with for each blindspot. *In terms of delivery mediums*, Stanford's toolkit comes in the form of a collection of worksheets, templates and resources while Nokia AI Design toolkit comes in the form of an interactive website dedicated to a single tool to aid practitioners in ensuring they have made all the necessary considerations to build responsible AI.

By just considering these four toolkits, it already becomes apparent that RAI toolkits come in a variety of shapes and sizes, which have been recently comprehensively reviewed by Wong et al. [88]. In terms of their presentation or display medium, these range from physical mediums, such as decks of cards [12] and canvases [48], to digital mediums, such as code packages [69] or interactive websites [16]. While physical mediums, and decks of cards especially, are heavily used across design fields [1, 66, 70], digital mediums provide the added benefit of interactivity and adaptability. Currently, there has been no exploration of the effects of RAI toolkits' medium or form on its users and the outcomes they produce.

2.2.2 Design Decisions. Nevertheless, not all RAI toolkits are equivalent or interchangeable. While still largely unexplored, the design decisions made when creating a RAI toolkit can impact its effect on those using it and the outcomes it helps to produce. For example, when working with non-technical stakeholders, a toolkit's use of metaphors to explain AI capabilities (e.g., anthropomorphise conversational AI [47]) has been found to be much more effective than simply listing capabilities [62]. Another recently explored design decision is whether or not a toolkit de-couples or "decontextualises" [88] its view of ethics from that of a specific domain or context. Such a design decision can significantly impact the use a toolkit by allowing practitioners to ignore contexts or abstract away inconvenient details, which, in turn, can encourage destructive behaviours such as shifting responsibility to other stakeholders [60].

Wong et al.'s review of 27 toolkits that focus on AI ethics also highlights broader trends within the design decisions of these interventions [88]. In terms of narrative, toolkits tend to focus on either societal harms of AI or organizational risks. They also sometimes focus on 'opportunities' as potential positive outcomes. Toolkits are either based on what is seen as responsible; on laws and regulations; or on some form of human rights, values or principles [88]. When aimed at developers and technical stakeholders, ethics is framed as a series of specifications or requirements, when aimed at business owners and executives, ethics is framed as business strategy and risk assessment. In terms of limitations, many toolkits focus on the

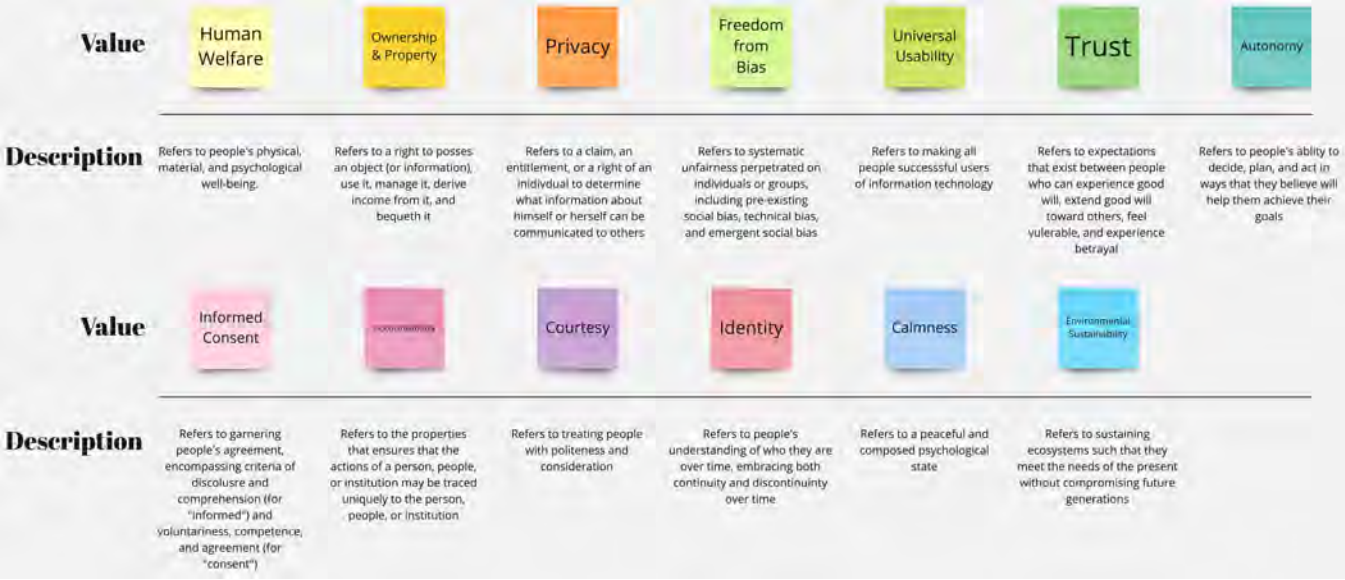


Figure 1: A list of the values stated by VSD as being "often implicated in system design" [26] and their descriptions in a Miro board.

technical aspects of ethics and make it difficult for non-technical stakeholders to get involved by offering little support for the "translational work" needed to bridge between disciplinary knowledge [22, 88]. They also advocate for stakeholder participation, but offer little guidance in terms of identifying and engaging stakeholders.

Wong et al. [88]'s work begins to provide a taxonomy for RAI toolkits based on certain design decisions, such as the narrative they support or the stakeholders they target. Our work is specifically interested in exploring which design decisions relate to the operationalisation of VSD's core principles and in what ways.

3 METHODOLOGY

Author Positionality Statement. To position ourselves as researchers and clarify our perspectives on the study [28, 35]: This research was conducted in a Western, European context. The research team includes two women and two men from North-Eastern Africa, and Southern and Western Europe, working in academia and industry. With individual backgrounds in Human-Computer Interaction, Design, Computer Science, and AI, the team shares a common interest in the Design of Human-centred AI and Responsible AI.

Below, we start by outlining the process used to select RAI toolkits for our user study (§3.1). This is followed by a discussion on how we mapped VSD values to those of RAI (§3.2), answering our (RQ_1) and then describe how we conducted workshops to investigate whether and how RAI toolkits support VSD in their design and facilitate collaboration and learning (§3.3), answering our (RQ_2).

3.1 Selecting and presenting RAI toolkits

To identify and select RAI toolkits for the workshops, we followed a similar methodology to that conducted by a recent review of RAI toolkits [88]. For our initial corpus, we reviewed a total of 63 toolkits; 27 toolkits from a recent RAI toolkits taxonomy [88], and an additional set of 36 toolkits from a large collection of practical tools for legal, ethical, and societal aspects of AI and data driven applications [75]. We removed two toolkits because they were duplicates in both sources (i.e., the AI Ethics Cards and Aequitas). Additionally, since the review of RAI toolkits was conducted in 2022 [88] and the online repository was undated, we included two additional toolkits that were released in 2023. Table 3 shows the full list of toolkits that were reviewed, and Figure 7 provides a breakdown of the selection process. The four-step process included:

- **Step 1 - Target Users**: We selected toolkits designed for AI technical practitioners (e.g. developers, data scientists), resulting in 38 candidate toolkits.
- **Step 2 - Focus on Regulation**: We excluded toolkits from regulatory institutes focusing on regulations, resulting in 21 candidate toolkits.
- **Step 3 - Indication of Use**: We excluded toolkits lacking evidence of recent use, resulting in 6 candidate toolkits. We did so by following Wong et al. [88] methodology using proxies such as toolkits' appearance in practitioner-made resource lists, search rankings, and, signs of community use.
- **Step 4 - Comparability**: We selected toolkits with comparable design features, collaboration and learning support (i.e. content division, graphics or illustrations, and provocative cues/questions) in order to control for any effects on the study results. The resulting toolkits were Nokia AI Design toolkit and the MIT's AI Blindspots toolkit. For brevity, we will refer to them as the Nokia AI Design toolkit and the MIT Blindspots toolkit respectively.

After selection, the toolkits were accessed and presented in the following manner for our study: **Nokia AI Design toolkit (Figure**

2): We could access the source code, allowing us to create a copy without creator mentions. User interaction involved sequential card navigation with answer boxes, a progress bar, and an option to save and export answers. **MIT Blindspots toolkit (Figure 3):** Unable to access the source code, we replicated the toolkit through a PowerPoint presentation. User interaction featured clickable thumbnails for detailed views, with QR codes on cards linking to additional information.

3.2 Mapping VSD values to RAI values

Three of the authors went through the list of VSD values and those of RAI. VSD defines a list of "human values often implicated in system design" [26] [p.17] to consider (Figure 1). Similarly, Responsible AI is about creating AI systems that are fair, transparent, and accountable, making a positive impact to society. To obtain the RAI values that are often used to design RAI toolkits, we relied on the NIST AI Risk Management Framework [63]. The framework identifies characteristics that contribute to AI systems that are fairness, explainable, accountable, privacy-preserving, secure and reliable, and sustainable. Alternatives include the Principled Artificial Intelligence from the Berkman Klein Center [24], which aligns with the NIST framework.

During this exercise, the authors found that it was difficult to conduct this mapping on the MIT Blindspots toolkit given the limited number of cards and due to the cards explicitly mentioning values such as fairness, explainability, accountability, safety, and so on; which defeated the purpose of the exercise. The Nokia AI Design toolkit proved much more effective due to the larger number and variety of cards, and the more implicit embedding of values within its cards and recommendations.

3.3 Conducting Workshops

The objective of our study was to investigate the extent to which VSD values align with RAI values, and whether and how RAI toolkits support VSD values in their design by promoting collaboration and learning. To do so, we conducted workshops with participants who engaged in value mapping and brainstorming while using selected RAI toolkits (Figure 4). The use of collaborative design workshops has been recommended when creating responsible AI [33] and is an effective approach for gathering interdisciplinary and in-depth insights [51] among several other benefits in the context of AI design [71]. We first describe the participants, followed up by the workshop activities, then by the data collection and data analysis process.

3.3.1 Participants. Participants were recruited on a voluntary basis. The inclusion criterion was that participants were "familiar with how AI systems work" and "how to build at least one type of AI systems". This was checked through a questionnaire where participants described in details how they had learned these skills, e.g. through formal education (e.g., university courses) or self-learning (e.g., online courses). Participants were recruited throughout the study until a saturation was perceived to be reached (i.e., our process of interpreting the data collected yeilded no new insights [9]), in line with the grounded theory approach [42] and studies with similar methods [73]. Recognizing the subjective nature of saturation, in this study, saturation was deemed reached when the majority of themes generated post-workshop were consistent with themes identified in previous workshops [9].

Table 1: Workshop participants' demographics.

Participant ID	Age	Gender	Role	AI Experience (years)
1	32	Male	MSc Student	0.5
2	31	Female	PhD Student	1
3	22	Male	MSc Student	1.5
4	23	Male	MSc Student	2
5	26	Male	MSc Student	2
6	22	Female	BSc Student	2
7	26	Male	PhD Student	2
8	28	Female	MSc Student	3
9	22	Male	MSc Student	3
10	22	Male	MSc Student	3
11	22	Male	MSc Student	3
12	25	Male	MSc Student	3
13	29	Female	PhD Student	3
14	30	Female	PhD Student	4
15	22	Female	MSc Student	4
16	31	Male	Post-Doc	6
17	27	Male	MSc Student	7

We recruited 17 participants (11 male, 6 female), whose ages ranged from 22 to 32 (M = 25.9, SD = 3.7). All were researchers with varying levels of experience with AI systems, ranging from 0.5 to 6 years (M = 2.9, SD = 1.6). Early-career researchers were screened to confirm either their current status as AI practitioners or their intention to pursue careers as AI practitioners. Table 1 summarises participants' demographics. The study was approved by the Science Engineering Technology Research Ethics Committee at Imperial College London under the SETREC reference 21IC7361. Participants signed consent forms prior to attending the workshop, and received £25 Amazon gift cards as compensation for their involvement.

3.3.2 Workshop Structure and Activities. The workshop was 2.5 hours long, and was conducted 4 times in Spring 2023, with different participants each time (groups of 4, 3, 4, and 6). It was conducted online using Microsoft Teams and Miro, and consisted of three activities and three surveys (following each activity). The overall procedure for the study is shown in Figure 4 and described below.

Activity 1 (Value Mapping) lasted 30 minutes. The goal of this activity was to empirically obtain a mapping between the Nokia AI Design toolkit and VSD values. The structure of this activity was derived from the methodological approaches of affinity mapping [34] and card sorting[64] as methods of directly assigning values by participants, as opposed to more subjective and implicit methods used in previous works [76]. First, participants were asked to reread the cards in the Nokia AI Design toolkit and then read a list of "universal values" outlined by VSD as being "often impacted upon by technology" [27]. The Miro board layout for this activity is shown in Figure 5. Participants were then asked to assign values to each card based on which values they felt the card was respecting

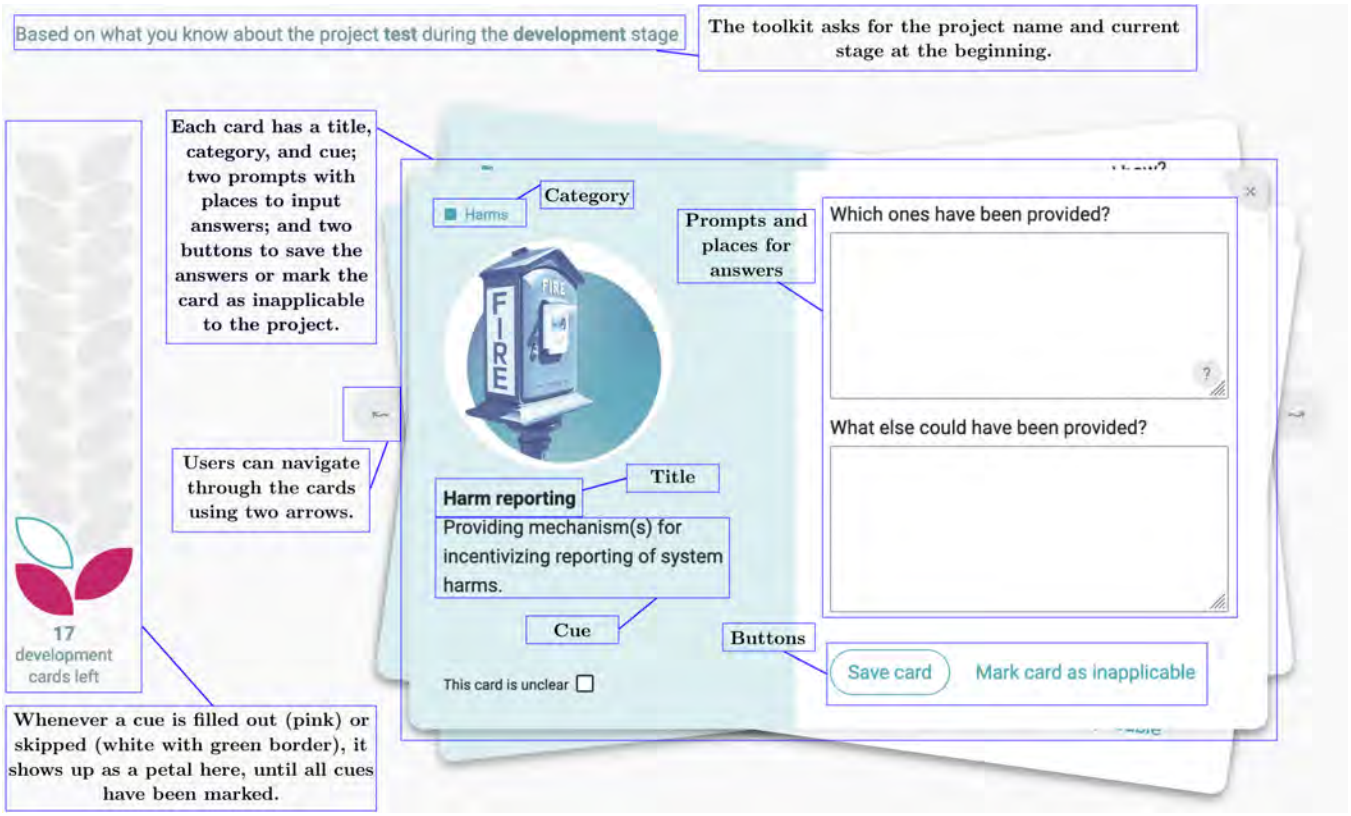


Figure 2: A screenshot of the Nokia AI Design toolkit with descriptions of each element in the interface in blue boxes.

or advocating for. They were told that they could assign multiple values to each card and assign a value to multiple cards. This was done individually for 15 minutes and then as a group for 15 minutes where participants aggregated all the values they had assigned to each card and then discussed and changed values until a consensus was reached for each card. Participants were then given another 15 minutes to go through the values assigned by the whole group, discuss any discrepancies, and reach a consensus together.

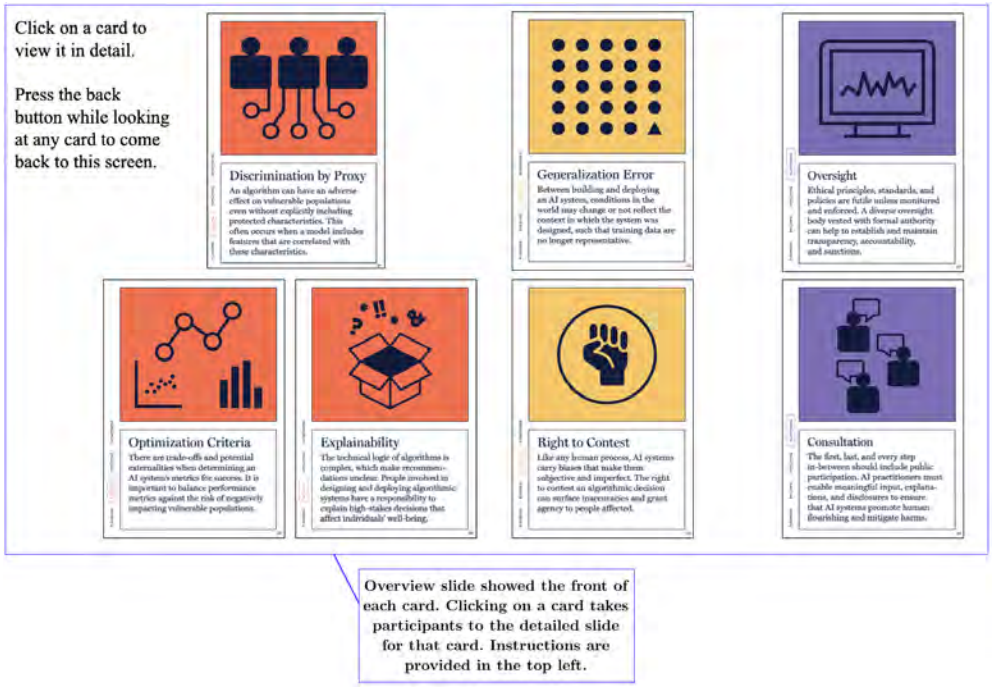
Activities 2 and 3 (Brainstorming) lasted 30 minutes each. The goal was to learn about how participants used and envisioned themselves using the toolkits, and to analyse the differences between the two toolkits in terms of their effects on participants and the ideas they helped them produce. Participants were given access to the toolkits through links embedded in the Miro board. To increase the comparability between the toolkits, both were presented in an interactive form, and we only included cards that related to the same phases (i.e. "designing", "deploying" and "using" in the Nokia AI Design toolkit; "building", "deploying", or "monitoring" in the MIT Blindspots toolkit) . As such, the Nokia AI Design toolkit had 20 cards and the MIT Blindspots toolkit had 7 cards.

Participants were either assigned the Nokia AI Design toolkit for Brainstorming 1 then MIT Blindspots toolkit for Brainstorming 2 (N = 8) , or the MIT Blindspots toolkit then the Nokia AI Design toolkit (N = 9). Both these activities were conducted by participants individually. Participants were asked to brainstorm as many activities, steps, or considerations needed to ensure that a fictional AI system is 'responsible' before deployment. They were given the following fictional scenario to work with:

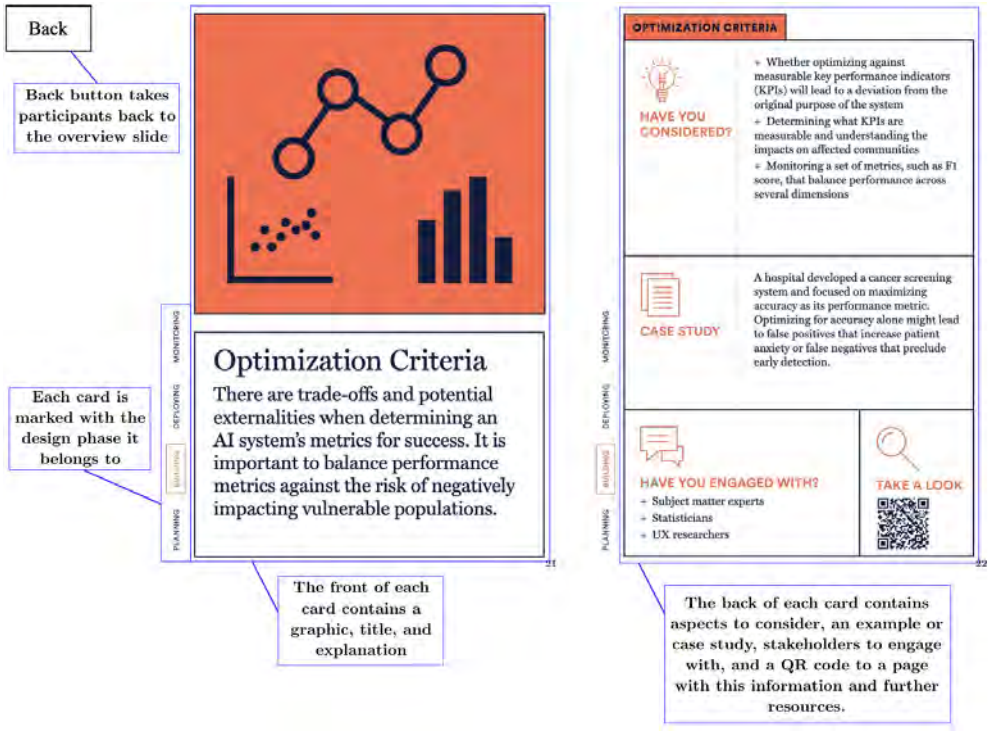
> *"You are on a team building an AI-powered chatbot for your company that will help people self-manage their health. The initial planning and design phases are complete and you are now building and training the AI model. You need to make a list of activities, steps, or considerations that your team will need to make moving forward to ensure the chatbot is responsible and ethical. These should be focused on the building, deployment and monitoring phases."*

Healthcare-related use-cases have been used in previous studies when exploring aspects relating to responsible AI [45]. This speculative healthcare context was chosen as a context that many participants are likely to be familiar with, and as suitable context to explore human values [77]. It is not our intent to focus this work solely on AI for healthcare or frame our contributions as such. Furthermore, the addition of the AI-powered chatbot was made to provide a relatable, relevant and interesting AI technology given the recent advent of large language models such as the ChatGPT.

3.3.3 Data Collection & Analysis. The study employed a qualitative approach using data collected throughout the workshops: value



(a) Screenshot of the MIT Blindspots toolkit's overview slide showing all the cards available.



(b) Screenshot of one of the MIT Blindspots toolkit's card slides showing one card in detail.

Figure 3: Screenshots of the two types of slides in the MIT Blindspots toolkit: the overview slide (top), and an example of a detailed slide (bottom).

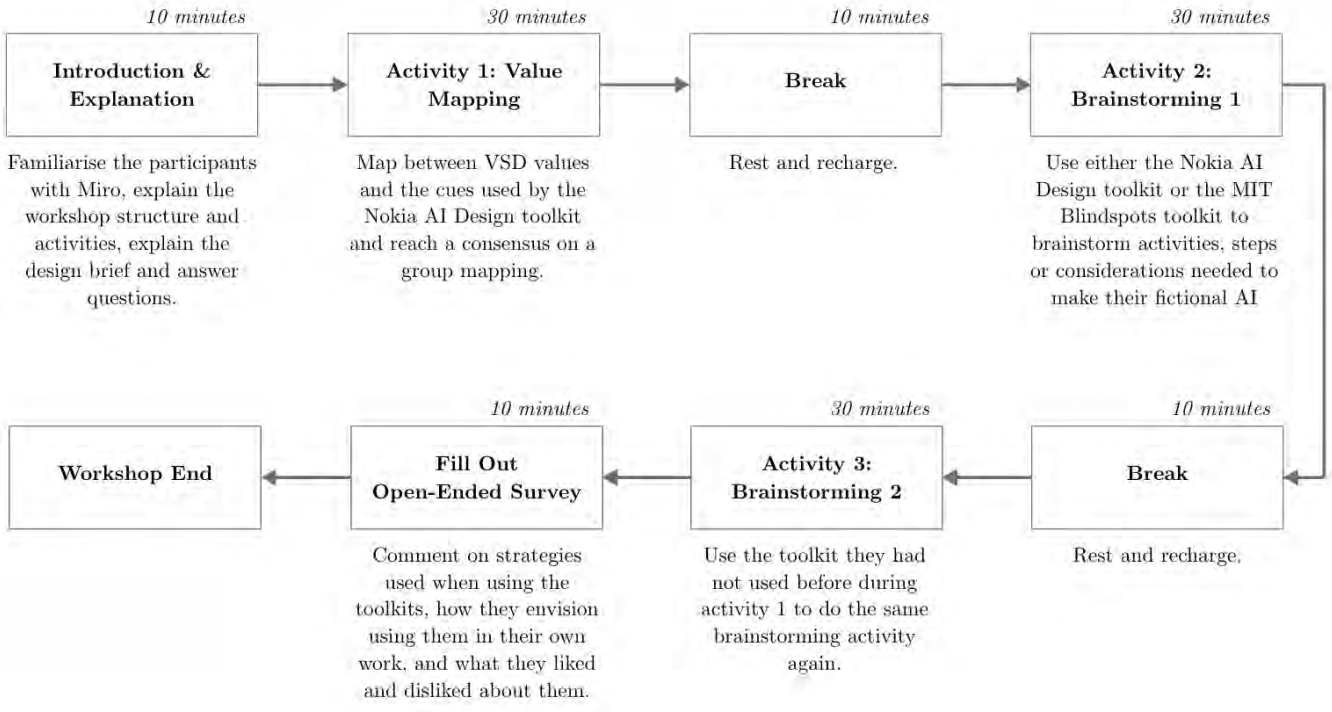


Figure 4: A flow diagram outlining the workshops' activities and their goals and duration.

Figure 5: The Miro board for Activity 1 after participants had assigned values to the Nokia AI Design toolkit and reached a consensus.

Table 2: Mappings between the Nokia AI Design toolkit's cues, their responsible AI pillars, and the VSD values assigned to them by consensus.

Nokia AI Design Toolkit	RAI Value	VSD Value
Identifying intended uses in consultation with relevant parties	Accountability	No consensus
Having an ethics committee or similar body approve intended uses	Accountability, Safety	Human Welfare
Providing mechanism(s) for incentivising reporting of system harms	Accountability	Accountability
Developing mitigation strategies to tackle harms or risks for each intended use	Fairness	Freedom from Bias
Documenting system components (including AI models) to allow for reproducibility and scrutiny	Accountability, Fairness	Accountability
Performing code review	Accountability, Fairness, Reliability	Accountability
Reporting evaluation metrics and checking them in different groups	Fairness, Transparency	Trust
Having interpretable outputs to allow for auditing purposes and informed decisions	Transparency	Trust
Documenting security of the system components in consultation with experts	Security	Privacy, Trust
Providing environmental assessments of the system	Sustainability	Environmental Sustainability
Developing feedback mechanism(s) that help update the system	Accountability	Accountability
Handling data without breaching any agreements or legal requirements	Accountability	Accountability
Comparing training and testing datasets in terms of data quality, representativeness, and fit with intended uses	Fairness, Transparency	Freedom from Bias
Identifying measurement errors in input data, including assumptions behind it	Transparency	Trust
Protecting sensitive variables in training/testing datasets	Accountability, Privacy	Privacy
Continuously testing the system's performance and reverting to the previous version if the tests are not successful	Reliability, Safety, Security	Trust
Including human oversight in the system	Accountability	Accountability
Ensuring team diversity	Accountability, Fairness	Freedom from Bias
Training team members on ethical values and regulations	Accountability	No consensus

mappings, transcripts and outcomes of brainstorming activities, and participants' responses to the open-ended survey.

Thematic analysis [10] was used to identify and cluster themes the researchers' identified within activity outcomes, workshop transcripts and open-ended survey questions. Initially, top-down coding relied on researchers' workshop observations (e.g "mentioning examples") and based on conceptual categories (e.g. "negative aspects mentioned regarding the Nokia AI Design toolkit/the MIT Blindspots toolkit"), while subsequent bottom-up coding constructed sub-themes based on researchers' understanding of the data [8]. The analysis, conducted in Miro using sticky notes from the workshop, involved participants' survey answers, quotes, and researchers' observations from the workshops and transcripts. These sticky notes were then clustered by the researchers into the top-down themes mentioned earlier. Afterwards, individual researchers organised themes into sub-themes using a bottom-up approach. Finally, discussions took place until a consensus was reached.

The resulting themes (see 4.2.2 in Section 4 for details) are as follows: "navigation", "considering stakeholder perspectives", "collaboration versus solo work", "open-ended cuing", "user experience and content", "lack of adaptive responses", "providing examples and case studies", "practical support needed".

4 RESULTS

4.1 RQ_1: How closely do Value Sensitive Design (VSD) values align with Responsible AI (RAI) values integrated into RAI toolkits?

VSD values align, to a great extent, with RAI values. When asked to map between the Nokia AI Design toolkit and VSD values, participants considered a number of stakeholders' perspectives and considered the values from the point of view of testers, developers, ethics boards, users. Overall, a consensus was reached between the experts' mapping and the workshop participants' mappings on all but three cards for: 'identifying intended users in consultation with relevant parties', 'training team members on ethical values and considerations', and 'having an ethics committee or similar body approve of intended uses' where the researchers assigned all three the value of 'accountability' and workshop participants assigned them the values of 'informed consent', 'universal usability', and 'human welfare' respectively. This has lead to an overall consensus of 85% (17/20 cards) across researchers and workshop participants for the following mappings:

Overall, the value of accountability was most implicitly represented by the cards provided within the Nokia AI Design toolkit (6/20), followed by trust (5/20). Out of the 13 VSD values provided by [26], 6 are represented in the toolkit's cards, although all the VSD values were assigned to various cards by at least one person during the workshops. The fact that the three cards where a consensus was not reached were assigned the value of 'accountability' by researchers indicates that the conceptual definition for that value held by the researchers might have differed from participants, which is supportive of previous work highlighting different groups having different value definitions and priorities [39].

It was interesting to note that almost all workshop participants struggled with the values of 'calmness' and 'courtesy' as they felt "unfamiliar" with them and were "not how they would refer to AI ethics aspects". One participant mentioned that an alternative value to those could be "competence or effectiveness" in the sense of "acting with due diligence, care and vigilance and making sure quality was good enough". Four participants also felt that these values were "secondary byproducts" as opposed to "primary concerns" for them. They appreciated that the cards would "factor in" or consider these values for them in the actions and recommendations they offered so that they would not have to think about them actively themselves.

Summary. A consensus was reached between the experts' mapping and the workshop participants' mappings on all but three cards where the researchers had assigned the value of 'accountability'. The cards represented the values of 'accountability' and 'trust' most commonly. Practitioners struggled with unfamiliar values and felt that some values had a secondary importance.

4.2 RQ_2: How do existing RAI toolkits incorporate VSD, and support collaboration and learning?

We begin by discussing the outcomes of the brainstorming session, followed by the design choices of the two toolkits and their support for collaboration and learning.

Outcomes of Brainstorming Session. Participants who used the Nokia AI Design toolkit first generated a total of 84 ideas across the 4 workshops, and a total of 35 ideas when they then used the MIT Blindspots toolkit. Conversely, participants who used the MIT Blindspots toolkit first generated 42 ideas across the workshops and 69 ideas when they then used the Nokia AI Design toolkit In both cases, participants generated a higher number of ideas using the Nokia AI Design toolkit, even when starting with the MIT Blindspots toolkit, despite our expectation that participants' second activity might generate fewer ideas than their first. Participants using the Nokia AI Design toolkit also generated a greater breadth of ideas and had a greater range of considerations under each idea or theme. Figure 6 shows the coding trees for the themes generated while brainstorming using each toolkit in both orders to highlight a disparity across toolkits.

Design Choices. In terms of navigation, participants contrasted the navigation strategies that both toolkits afforded. On one hand, the MIT Blindspots toolkit offered more back-and-forth navigation. Three participants preferred being able to return to the 'overview' screen and select the desired card. On the other hand, the Nokia AI Design toolkit offered more sequential navigation. Four participants preferred this more sequential nature as it forced them to consider each card *"one by one"* and write down *"what ideas or actions it made [them] consider"* and not skip ones they assumed to be irrelevant.

Four of the participants who started with the Nokia AI Design toolkit filled out their answers directly into the tool itself and sent the generated PDF to the researchers as opposed to using the Miro board, and enjoyed using the interface directly. Participants also strongly appreciated the ability to save their responses as a PDF afterwards. One participant mentioned that the *"PDF consolidated review and the option to upload an old result for comparison/review was really good"* and another participant complained that in the MIT Blindspots toolkit there was *"no way to evaluate or summarise [their] ideas/thoughts as [they] go through the toolkit"*.

Participants also commented on the Nokia AI Design toolkit's flexibility, allowing them to use it throughout the design process and therefore making them feel more efficient and productive during the brainstorming activity. Because of its ability to diverge during brainstorming sessions, four participants felt that they would want to use the Nokia AI Design toolkit during early planning phases of a project as a *"starting point"* to *"devising a plan on how to design an AI"*, *"make [them] think about what [they] would need to think about to ensure this tool is ethical"*, and *"adjust the system design to be more responsible and ethical"*, as well as to *"raise [their] concerns to [their] colleagues and team and communicate [their] views."* Five participants felt that they could also use the toolkit towards the end of a project for *"auditing", "testing", and "evaluation"*, with two participants stating that it could be used repeatedly throughout. Participants described their brainstorming sessions as *"more productive", "more aware", "more critical", "more efficient", and "more comprehensive"* having used the Nokia AI Design toolkit. Four participants referred to the toolkit giving them ideas that *"did not come to mind"*, and *"new solutions"* that *"cover blindspots"* they originally had.

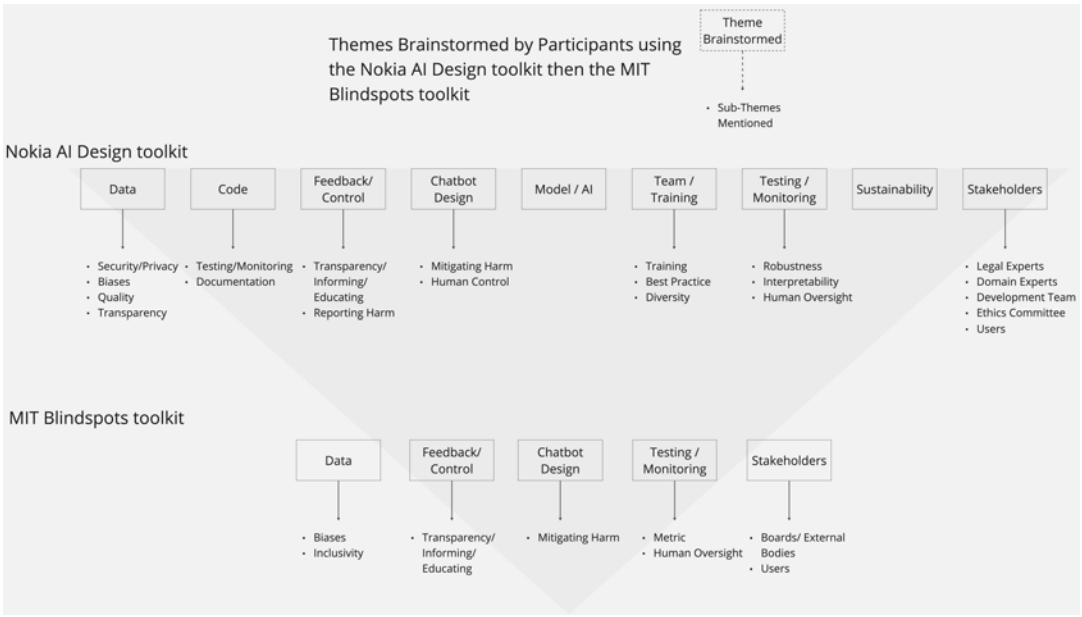
Participants also commented about the Nokia AI Design toolkit's lack of adaptive responses. One participant felt they could fill in anything in the boxes provided and the tool would say *"well done"* or *"40% done/considered"*, but that would not be true and would be misleading to think. Another participant felt that the Nokia AI Design toolkit *"did not evaluate at all what [they] wrote"* and was *"unusable"*, and another commented that the tool *"does not give an actual indicator of how good the system is already"*. They were worried that the tool relied too much on how well and how reliably people explain their systems. One participant suggested that the tool should *"provide a more custom response (e.g., analyse the Github repo and answers) instead of just repeating user input"*.

Collaboration and Learning. In terms of collaboration, participants found that the Nokia AI Design toolkit fostered more open-ended brainstorming and thus allowed for discussion and collaboration, especially within teams. They described their brainstorming with the toolkit as *"organic"* and *"unbiased"* given the lack of direction and the open/general nature of the cards and that the toolkit *"supported open-ended ideation to let out what you feel"* and *"provided many ideas from which it was easy to formulate more specific activities and considerations"*. Participants used the cards more as starting points or *"springboards"* that allowed them to "sprout ideas" and provide *"inspiration"*.

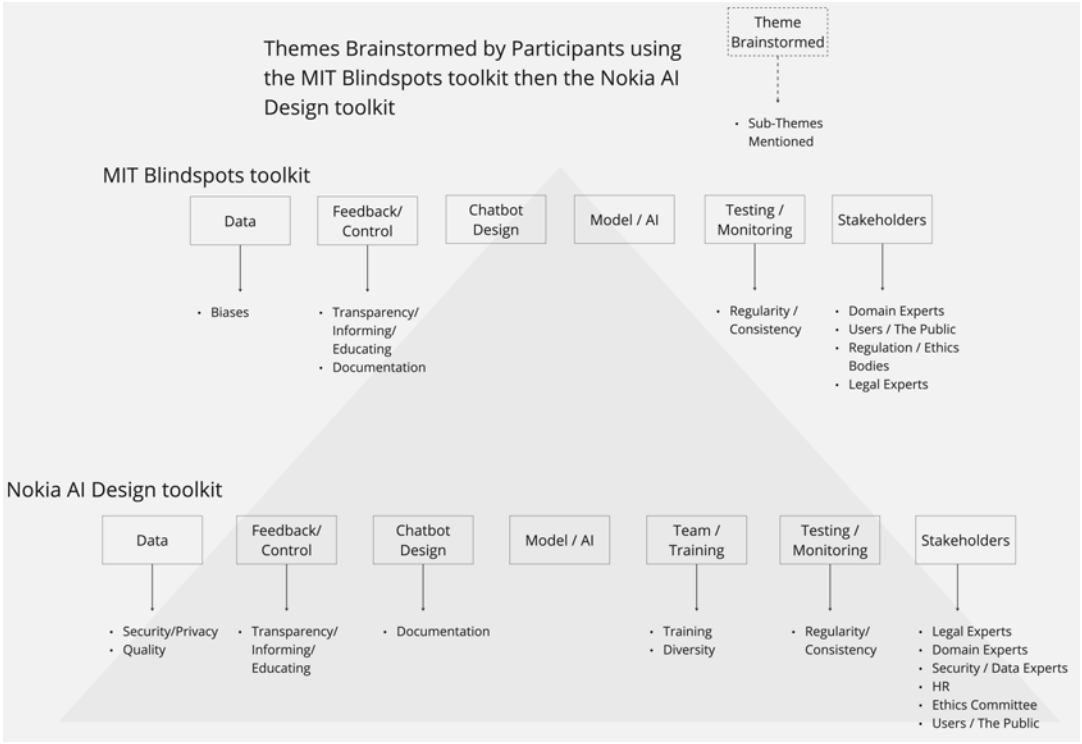
Six participants felt that the the Nokia AI Design toolkit was more suited for collaborations. One participant commented that *"[the] toolkit is good in team meetings because it's more open-ended and people can contribute more"* and that it was useful to *"get the ball rolling"*. Nine participants indicated that the MIT Blindspots toolkit was more suited for solo work whereas the Nokia AI Design toolkit was more suitable for team collaborations. They felt that the MIT Blindspots toolkit could be used as an *"education tool"*, mainly because it *"provides examples and some insights to people who might not be familiar with all the different aspects of RAI"*.

Participants discussed using the two toolkits together: *"[the MIT Blindspots toolkit] offers more concrete instructions where you can go back and flesh out your ideas"* after diverging with the Nokia AI Design toolkit since *"as one person you would need more cuing [than is available in the Nokia AI Design toolkit]"* to flesh out your ideas. Another participant also echoed this sentiment stating: *"in an individual setting, I felt like it [the Nokia AI Design toolkit] needed more direction."* Finally, another participant mentioned that: *"A good strategy would be to start with [the Nokia AI Design toolkit] to brainstorm and think about the issues, then use [the MIT Blindspots toolkit] to learn more about how to handle the issues, then go back to [the Nokia AI Design toolkit] to build on your ideas"*.

Participants also commented about toolkits ability to help them consider different stakeholders' perspectives. Five participants reported that the Nokia AI Design toolkit encouraged them to consider aspects from the perspectives of different stakeholders including users, the development team, the legal team, and ethicists.



(a) Ideas brainstormed by participants using the Nokia AI Design toolkit then the MIT Blindspots toolkit.



(b) Ideas brainstormed by participants using the MIT Blindspots toolkit then the Nokia AI Design toolkit.

Figure 6: Coding trees for both toolkits. (a) Participants who used the Nokia AI Design toolkit then the MIT Blindspots toolkit; (b) Participants who used the MIT Blindspots toolkit then the Nokia AI Design toolkit. Each box represents a main theme, and is then broken down into sub-codes that represent the ideas under each theme that participants touched upon in their brainstorming sessions.

They also mentioned perspectives in the sense of considering the system from a *"system's, sustainability, usability, trust, and privacy perspective"*.

Finally, the key strength of the MIT Blindspots toolkit lies in its provision of examples and case studies, as well as recommendations and suggestions fostering learning. All study participants showed appreciation towards these aspects of the toolkit. One participant even felt that brainstorming would be redundant as they felt that all the considerations they needed to think of were provided by the MIT Blindspots toolkit's recommendations already. Participants felt that the case studies and examples offered "directed action", whereas *"it was difficult to apply [the Nokia AI Design toolkit]'s questions to a specific use-case"* and that *"the cards are quite general and it required thinking about how this applied to our specific scenario"*. Another participant also stated that they had to read the Nokia AI Design toolkit's cards twice: once to understand them, then once to apply them to the workshop's specific scenario. Two participants also felt that if you did not understand one of the Nokia AI Design toolkit's cards, there were no examples to help. Finally, another participant also felt that there were too many assumptions they had to make to fit the Nokia AI Design toolkit to their application. Nonetheless, one participant also noted that the *"case study was useful but probably pigeon-holed you into thinking about the problem in that dimension only"*. Regarding providing recommendations and solutions, two participants felt that with the Nokia AI Design toolkit if you do not know how to solve a specific issue it offers no guidance, while the MIT Blindspots toolkit had case scenarios and examples you could build upon. One participant also felt that *"[the Nokia AI Design toolkit] just said 'have you thought about this or that' but there was nothing about what to do about it"* and another participant built upon that stating: *"I am ignorant [about legal aspects], and the cards did not provide any help except highlighting my ignorance [referring to the Nokia AI Design toolkit]"*.

Nevertheless, few participants found the MIT Blindspots toolkit's recommendations too vague and overwhelming. They did not want a *"you need to do this"* approach, but more practice-based help such as providing tool recommendations. One participant referenced Adobe tools showing you how to improve accessibility and another participant mentioned wanting a code analysis tool that flags problematic parts of the code as more helpful ways to provide support than the MIT Blindspots toolkit's recommendations. They mentioned that the suggestions felt very broad (e.g., documenting the security of all system components in consultation with experts) and almost like PhD projects (e.g., creating a tool to measure fairness). Interestingly, one participant mentioned that all the toolkit does is keep asking *"have you considered this?"*, continuing on to reply *"no, and I didn't have time to"*. This highlights participants' negative attitude and feeling of being overwhelmed because of this lack of practical support.

Summary. Overall, during the brainstorming session, participants were able to generate a greater breadth of ideas and go into more depth with each idea using the Nokia AI Design toolkit. They also felt it allowed them to consider more stakeholders' perspectives. Participants also enjoyed its flexibility, the ability to fill out answers directly in the tool, and the ability to save their responses. Participants felt the Nokia AI Design toolkit was more suited for collaborative work, whereas the MIT Blindspots toolkit was better suited for solo work and education. The Nokia AI Design toolkit's open-ended cuing supported its use during different design phases and more divergent brainstorming. Finally, participants felt that the MIT Blindspots toolkit provided more information and found its provision of examples and case studies educational, but some participants felt that the recommendations provided were too vague and general.

5 DISCUSSION

By conducting four workshops with 17 AI researchers, we established that VSD and RAI values align to a great extent, and, as such, we explored the effects of toolkits' design features regarding collaboration and learning. We identified a number of links between toolkits' design choices, which resulted in differences in the ways participants perceived and interacted with them. First, our participants generally found the MIT Blindspots toolkit more suitable for individual work and the Nokia AI Design toolkit better for collaboration due to its generalisability and open-endedness. The Nokia AI Design toolkit facilitated broader ideation, evident in the quantity and variety of ideas generated and the breadth of categories in its coding trees. Conversely, the MIT Blindspots toolkit's provision of examples, case studies, solutions, and recommendations was a key discussion point during the workshops. In contrast, the absence of such elements in the Nokia AI Design toolkit required participants to engage more deeply and spend more time understanding its content. Finally, participants had mixed reactions to design features unrelated to content such as the order and number of cards, and navigation options. Non-linear navigation was appreciated for its flexibility, while a linear approach ensured thorough consideration of all cards.

Next, we synthesise these results into a number of theoretical implications in terms of links between the toolkits' design features and their support of VSD (§5.1), and practical implications revolving around the toolkits' ability to operationalise VSD by supporting collaboration and learning (§5.2).

5.1 Theoretical Implications

Discrepancies between RAI and VSD. Overall, the RAI values matched the VSD values closely. However, our workshop participants and the study researchers did not reach consensus on two values: *accountability* and *transparency*. For accountability, the lack of consensus might be explained by recent empirical evidence illustrating that different groups of people define (and prioritize) responsible AI values differently [39]. This suggests that seemingly similar sets of values should not be used interchangeably without a thorough understanding of their fundamental differences. For transparency, which VSD often refers to as trust, the picture was slightly different. While transparency and trust are certainly intertwined [86], transparency has been found to both enable and violate trust depending on contextual factors [91]. For example, revealing an AI model's low confidence score for its prediction might reduce trust in its competence, while increasing trust in its honesty. VSD has also been used as a facilitator for transparency [20], despite not including the value explicitly in the original set provided. Given the significance of transparency in AI systems [46] and its

established distinction from trust, we suggest that VSD's values "often implicated in system design" [26] should be updated to reflect these findings. While the VSD methodology based on conceptual, empirical, and technical investigations has been recently adapted to AI systems [82, 93], updating the core set of values included in the VSD framework has been largely remained unexplored.

Framing RAI Toolkits in Research Taxonomies. Wong et al. (2022) [88] discuss how different RAI toolkits frame ethics, and the discourse they use around ethical concepts. For example, some toolkits choose to focus on risks and negative outcomes, while others highlight the benefits and positive outcomes of building responsible AI. We propose adding a new dimension regarding the framing of the toolkits themselves (e.g., as educational, collaboration, or reflection tools). Framing the types of support that RAI toolkits offer, or the activities they can facilitate, can help their users to select appropriate toolkits more effectively, especially given the large number and variety available [88]. Our research shows that this framing is not always necessarily intended by the toolkit creators, it tends to be more implicit and depends on the design features of the toolkit. For example, participants felt that the Nokia AI Design toolkit's ability to support "organic" brainstorming through its more open-ended cues was conductive of discussions and collaboration in team settings. Previous studies have also established the role of open-ended discussions in supporting collaboration [53]. On the other hand, the MIT Blindspots toolkit was perceived as an educational tool because of its detailed examples, case studies, and recommendations.

Low Actionability in RAI Toolkit Recommendations. While previous work has shown that the use of examples and analogies can help establish empathy [37] and lower the effort needed during learning [4], participants still felt that the MIT Blindspots toolkit's recommendations fell short of being practically meaningful. A large body of work has recently surfaced discussing the limitations of general recommendations for responsible AI [33]. These works echo the sentiments of participants and also call for more practical tools and frameworks to improve actionability [57]. In this study, participants thought the recommendations provided were too general or too large in scope, making participants feel "overwhelmed" and even "ignorant". Participants' suggestions on how to improve these recommendations all revolved around analysing their code, providing links to specific tools that can be used, and giving more customised feedback. Similar to Wong et al. (2022) [88]'s finding that most RAI toolkits recommend involving stakeholders but offer no practical guidance on how to do that, this study also adds that this lack of practical guidance for advocated actions extends beyond involving stakeholders and generalises to a number of different recommendations provided by these toolkits.

Considering the Role of Non-Content-Based Features of RAI Toolkits. Participants were affected by a number of design decisions unrelated to the toolkits' content. Participants frequently separated their comments on toolkits' content (e.g., the cards given, ideas expressed, and text used), the presentation of that content and the interactions the toolkits afforded. While Wong et al. (2022) [88] mention the work practices that toolkits explicitly envision, we suggest that the toolkits' design decisions relating to content presentation and interaction modalities can also impact the work practices they support. For example, participants felt that the MIT Blindspots toolkit's provision of non-sequential navigation, where they could jump in and out of different parts, supported more iterative work processes. Our findings also show that toolkits' design decisions can further exacerbate a "decontextualized approach to ethics" [p. 14][88]. For example, in the Nokia AI Design toolkit, while participants were able to input their answers directly into the tool, they were expecting custom or interactive responses that addressed what they had written. They felt that the outputs of the tool were too generic and even misleading in that they relied too heavily on toolkit users' ability to describe the system accurately and their integrity to describe it honestly.

Overall, our findings support several strands of previous research and extend them to new aspects and paradigms, as well as offering an understanding of how VSD can impact RAI toolkit design and how VSD values and RAI values align.

5.2 Practical Implications

Looking specifically at the toolkits' ability to operationalise the core concepts of VSD (i.e., working with stakeholders and embedding values into their work), the toolkits' design features support these aspects in various ways. We synthesized six design recommendation for creators of RAI toolkits. These recommendations are summarised in Figure 8 in the Appendix.

Encouraging Collaboration with Stakeholders through Open-Ended Cuing. Despite the majority of RAI tools targeting technical practitioners [2, 41, 67, 92], several existing tools can support the inclusion of and collaboration with external or non-technical stakeholders through their design decisions. Supporting collaboration versus solo work has been a main point of discussion across the study. The ways in which the Nokia AI Design toolkit supports collaboration is through the open-ended and general nature of its cards, affording broad and unbiased ideation and opening several avenues for discussions. This can both support collaboration across teams, but can also lower the barrier-to-entry for non-technical stakeholders more explicitly and practically [88] as the cards can help spark new ideas they have not thought of before without being extremely technical or specific and thus less intimidating. While this open-endedness had identifiable disadvantages, such as an increased cognitive load to apply the toolkit to specific scenarios or technologies, it was the main design feature that supported collaboration found in the study.

Increasing Empathy through Examples, Case Studies and Mentioning Stakeholders. Empathy is crucial for ethical decision-making in engineering contexts [36]. Two main design features supported an increase in empathy and the consideration of diverse stakeholders' perspectives: the MIT Blindspots toolkit's provision of examples and case studies, and the Nokia AI Design toolkit's mention of numerous stakeholders in its cards. In the former, this led to participants' mentioning an improved ability to empathise and reflect on users' and stakeholders' experiences as they did not have to spend extra mental effort translating the information to their specific scenario. In the latter, participants were able to brainstorm considerations and steps that need to take place to build responsible AI that take into account a wider range of stakeholders and perspectives. Empathy and collaboration go hand-in-hand as collaboration

fosters empathy which then fosters a more user-centred mindset in practitioners [38].

Supporting Iteration through Generalisability and Navigation. Our study surfaced two design features that could support iterative development (as is an inherent part of and recommended by VSD adapted to build AI systems [82]). Firstly, participants mentioned that the Nokia AI Design toolkit's open-ended nature made it suitable for both early-stage ideation during planning stages, and late-stage testing and evaluation phases. The toolkit could be used several time throughout the design process and the fact that results can be reloaded into the tool and compared can support constant improvement and iteration. This finding has parallels in previous work on AI where interpretations could differ depending on the stage a practitioner was involved in [58]. Secondly, participants mentioned that the MIT Blindspots toolkit's non-linear navigation allows them to jump back and forth as many times as needed between different phases without having to sequentially go through the cards every time. While this style of navigation might mean that certain cards are overlooked if toolkit users deem them irrelevant, it can support iterating through phases more practically.

Supporting Reflectivity & Meaningful Outcomes through Responsiveness and Feedback. Previous work has advocated for practitioner reflections [60] and identified RAI toolkits that explicitly call for such reflections [88]. In this study, we found that toolkits' design decisions, such as providing response and adaptive feedback to users' responses, can affect their ability and willingness to reflect. The provision of customised and adaptive feedback would make the outputs of such toolkits more meaningful and would allow toolkit users to reflect on their responses and improve their practices. The current lack of responsive feedback could lead to misleading outcomes where toolkit users feel their work is sufficient but it does not actually respect the required values. This deficiency also means that the effectiveness of the tool relies on its use and the discretion of its users in reporting their work. In the context of AI systems, previous works have already been weary of leaving too much up to the discretion of practitioners [87], and without providing actionable and customised feedback, RAI toolkits risk following suit.

Supporting Actionability & Shared Knowledge through Creating Accessible Outcomes. Participants valued the Nokia AI Design toolkit providing them with accessible outcomes that they could refer back to easily and use in their work moving forward. It is rare that RAI toolkits provide outcomes in this form, despite recent findings showing that practitioners use AI ethics resources and their outcomes in a number of actionable ways [89]. With decks of cards especially and other toolkits such as the MIT Blindspots toolkit, users would have to provide their answers or outcomes in a separate form (e.g., on paper, on a Miro board) and there would be extra work needed to summarise or make sense of these outcomes. Given technical practitioners' resistance or reluctance to engage in value-based and ethics-related work [49], seeing it as a burden or additional load [60], providing accessible and actionable outcomes might encourage them to engage more. Providing outcomes in an easy-to-read form that is understandable by a variety of toolkit users can also serve as shared knowledge that helps teams establish a shared mental model of the outcomes produced [13] and meaningful discourse [74] instead of appealing to one type of practitioner over the other (e.g., code for developers or sticky notes for designers).

Reducing Cognitive Load through Designing with Values in Mind. Finally, mapping the Nokia AI Design toolkit to VSD values shows that RAI toolkits are able to implicitly respect a number of VSD values without being explicitly designed with these values in mind. Participants appreciated not having to think of more human-centred values during their work, rather preferring that the toolkit they are using considers these values for them implicitly in the recommendations and solutions it provides and the ideas it sparks. Such an approach could therefore help overcome practitioners' reluctance to engage in such work [49, 78] and having to take on responsibilities outside of their roles to bridge disciplinary gaps across stakeholders [19]. While other interventions such as training practitioners to consider these values and understand their implications are certainly needed, implicitly supporting these values until practitioners are capable or willing to explicitly do so themselves can be extremely helpful. From the findings of this study where participants' did not reach a consensus with experts on certain cards, and previous work [39], it becomes clear that ensuring toolkit users are aware of value definitions is crucial to avoid misunderstandings.

5.3 Limitations and Future Work

This study has three limitations that call for future research efforts. Firstly, the limited size of our participant sample reduces the generalisability of our results. Future work would benefit from testing the generalisability of these findings on larger samples. It is also worth noting that given recent findings that different groups prioritise and perceive values differently [39], replicating this study with a different cohort besides early-career researchers as they might perceive the toolkits differently and react in other ways.

Secondly, we also acknowledge the relative homogeneity in researchers' backgrounds and the study's research context and realise that results might differ across different disciplines and regions. It is worth testing whether these links and implications also apply within other socio-cultural contexts and specific domains.

Finally, we opted in to test two RAI toolkits for the study practicalities. However, other RAI toolkits might be applicable. As such, future work should include: *i)* a wider study with a larger number of RAI toolkits, *ii)* a quantification of the exact effects of the different links established and their influence on each other, *iii)* an exploration of how the presentation form or medium used by toolkits impacts their effects on users and the outcomes produced, and *iv)* an exploration of how other theoretical frameworks besides VSD are operationalised. It is important to note that while steps were taken to improve comparability between the two toolkits used in this study, it is challenging to directly compare toolkits with different content and delivery mediums.

6 CONCLUSION

The aim of this study was to explore (i) the extent with which Responsible AI toolkits advocate for Value-Sensitive Design values in their content and recommendations and (ii) the extent with which different design features of these toolkits affects their ability to support VSD by promoting stakeholder collaboration and toolkit

user learning. Through a qualitative approach involving workshops with 17 AI researchers using RAI toolkits, we highlighted relationships between RAI toolkits and VSD values, and explored the design features influencing stakeholder collaboration and user learning in RAI toolkits. Key findings include the facilitation of collaboration through open-ended cuing, increased empathy via examples and case studies, support for iteration through generalisability and navigation, meaningful outcomes through responsiveness and feedback, actionability and shared knowledge through accessible outcomes, and reduced cognitive load by implicitly integrating values in toolkit recommendations. These insights contribute to understanding the operationalisation of theoretical frameworks like Value Sensitive Design in Responsible AI toolkits, addressing the need for practical and user-friendly tools in the design of Responsible AI.

ACKNOWLEDGMENTS

This work is partially funded by the Leverhulme Trust through the Leverhulme Centre for the Future of Intelligence [Award Number: RC-2015-067]. This work was also supported by The Alan Turing Institute's Enrichment Scheme.

REFERENCES

[1] T. Aarts, L. Gabrielaitis, L. de Jong, R. Noortman, E. van Zoelen, S. Kotea, S. Cazacu, L. Lock, and P. Markopoulos. 2020. Design Card Sets: Systematic Literature Survey and Card Sorting Study. Proceedings of the ACM Designing Interactive Systems Conference (DIS). https://doi.org/10.1145/3357236.3395516

[2] A. Balayn, M. Yurrita, J. Yang, and U. Gadiraju. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors That Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 482–495. https://doi.org/10.1145/3600211.3604674

[3] A. Berger, A. Kurze, A. Bischof, J. Benjamin, R. Wong, and N. Merrill. 2023. Accidentally Evil: On Questionable Values in Smart Home Co-Design. Proceedings of the Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/3544548.3581504

[4] S. Bolkan and A. Goodboy. 2019. Examples and the facilitation of student learning: should instructors provide examples or should students generate their own? *Communication Education* 68, 3 (2019), 287–307. https://doi.org/10.1080/03634523.2019.1602275

[5] A. Borning, B. Friedman, J. Davis, and P. Lin. 2005. Informing Public Deliberation: Value Sensitive Design of Indicators for a Large-Scale Urban Simulation. ECSCW 2005. https://doi.org/10.1007/1-4020-4023-7_23

[6] A. Borning and M. Muller. 2012. Next Steps for Value Sensitive Design, Vol. 2. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/2207676.2208560

[7] K. Boyd. 2022. Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3531146.3534626

[8] V. Braun and V. Clark. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[9] V. Braun and V. Clark. 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health* 13, 2 (2021), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

[10] V. Braun and V. Clarke. 2012. *Thematic analysis.* American Psychological Association, 57–71. https://doi.org/10.1037/13620-004

[11] J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability and Transparency (FAccT).

[12] A. Calderon, D. Taber, H. Qu, and J. Wen. n.d.. AI Blindspot: A discovery process for spotting unconscious biases and structural inequalities in AI systems. https://aiblindspot.media.mit.edu/

[13] J. Cannon-Bowers, E. Salas, and S. Converse. 1993. *Shared mental models in expert team decision making.* Psychology Press.

[14] V. Chen, J. Li, J. Kim, G. Plumb, and A. Talwalkar. 2022. Interpretable Machine Learning: Moving from mythos to diagnostics. *ACMQueue* 19, 6 (2022). https://doi.org/10.1145/3546036

[15] M. Cherubini. [n. d.]. Ethical Autonomous Algorithms. https://medium.com/@mchrbn/ethical-autonomous-algorithms-5ad07c311bcc

[16] M. Constantinides. 2022. Making Responsible AI development simple through prompt cards. https://www.bell-labs.com/institute/blog/making-responsible-ai-development-simple-through-prompt-cards/

[17] C. Custis. 2021. Operationalizing AI Ethics through Documentation: ABOUT ML in 2020 and Beyond. https://partnershiponai.org/about-ml-2021/

[18] F. Delgado, S. Barocas, and K. Levy. 2022. An Uncommon Task: Participatory Design in Legal AI. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022). https://doi.org/10.1145/3512898

[19] W.H. Deng, N. Yildirim, M. Chang, M. Eslami, K. Holstein, and M. Madaio. 2023. Investigating Practices and Opportunities for Cross-Functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 705–716. https://doi.org/10.1145/3593013.3594037

[20] J. Dexe, U. Franke, A.A. Nöu, and A. Rad. 2020. *Towards Increased Transparency with Value Sensitive Design.* Springer. https://doi.org/10.1007/978-3-030-50334-5_1

[21] DigitalCatapult. 2020. Lessons in Practical AI Ethics.

[22] M. Dolata and G. Schwabe. 2019. Translation and Adoption: Exploring Vocabulary Work in Expert-Layperson Encounters. *Computer Supported Cooperative Work (CSCW)* 28 (2019), 685–722. https://doi.org/10.1007/s10606-019-09358-9

[23] S. Elsayed-Ali, S. Berger, V. Santana, and J. Becerra-Sandoval. 2023. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI.

[24] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020). https://doi.org/10.2139/ssrn.3518482

[25] S. Fox, S. Shorey, D. Kang, E.and Valle, and E. Rodriguez. 2023. Patchwork: The Hidden, Human Labor of AI Integration within Essential Work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023). https://doi.org/10.1145/3579514

[26] B. Friedman, P. Kahn, and A. Borning. 2002. Value Sensitive Design: Theory & Methods.

[27] B. Friedman, P. Kahn, J. Hagman, R. Severson, and B. Gill. 2006. The Watcher and the Watched: Social Judgments About Privacy in a Public Place. *Human-Computer Interaction* 21 (2006), 235–272. https://doi.org/10.1207/s15327051hci2102_3

[28] H. Frluckaj, L. Dabbish, D. Widder, H. Qiu, and J. Herbsleb. 2022. Gender and Participation in Open Source Software Development, Vol. 6. Proceedings of the ACM on Human-Computer Interaction. https://doi.org/10.1145/3555190

[29] V. Fuchsberger, C. Moser, and M. Tscheligi. 2012. Values in Action (ViA): Combining Usability, User Experience and User Acceptance. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/2212776.2223711

[30] I. Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (2020), 411–437.

[31] J. Garst, V. Blok, L. Jansen, and O. Omta. 2022. From Value Sensitive Design to values absorption – building an instrument to analyze organizational capabilities for value-sensitive innovation. *Journal of Responsible Innovation* (2022).

[32] T. Hagendorff. 2020. The ethics of AI ethics: an evaluation of guidelines. *Minds and Machines* 30 (2020), 99–120. https://doi.org/10.1007/s11023-020-09517-8

[33] M. Harbers and A. Overdiek. 2022. Towards a living lab for responsible applied AI. In *Proceedings of the Design Research Society Conference (DRS)*. https://doi.org/10.21606/drs.2022.422

[34] G Harboe and E. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/2702123.2702561

[35] L. Havens, M. Terras, B. Bach, and B. Alex. 2020. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. https://doi.org/10.48550/arXiv.2011.05911

[36] Beever J. Strobel J. Brightman A.O. Hess, J.L. 2017. Empathic Perspective-Taking and Ethical Decision-Making in Engineering Ethics Education. *Philosophy of Engineering and Technology* 26 (2017). https://doi.org/10.1007/978-3-319-45193-0_13

[37] J. Holden. 2018. Improving Nursing Student Empathy With Experiential Learning. Proceedings of the Nursing Education Research Conference. http://hdl.handle.net/10755/623970

[38] J. Howcroft, K. Mercer, and J. Boger. 2021. DEVELOPING ETHICAL ENGINEERS WITH EMPATHY. Proceedings of the Canadian Engineering Education Association (CEEA). https://doi.org/10.24908/pceea.vi0.14856

[39] M. Jakesch, Z. Buçinca, S. Amershi, and A. Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3531146.3533097

[40] A. Jobin, M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Natural Machine Intelligence* 1 (2019), 389–399. https://doi.org/10.

1038/s42256-019-0088-2
[41] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219
[42] S. Khan. 2014. Qualitative Research Method: Grounded Theory. *International Journal of Business and Management* 9 (2014). https://doi.org/10.5539/ijbm.v9n11p224
[43] Stanford Digital Civil Society Lab. n.d.. Digital Impact Toolkit. https://digitalimpact.io/toolkit/
[44] M. Lee, D. Kusbit, A. Kahng, J. Kim, X. Yuan, D. Chan, A. See, R. Noothigattu, S. Lee, A. Psomas, and A. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI. https://doi.org/10.1145/3359283
[45] V. Liao and S. Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3531146.3533182
[46] J. Lindley, H. Akmal, F. Pilling, and P. Coulton. 2020. Researching AI Legibility through Design. Proceedings of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376792
[47] M. Luria. 2018. Designing Robot Personality Based on Fictional Sidekick Characters. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3173386.3176912
[48] A. Maillet. 2019. Introducing the Human-Centered AI Canvas. https://medium.com/@albmllt/introducing-the-human-centered-ai-canvas-a4c9d2fc127e
[49] N. Manders-Huits and M. Zimmer. 2009. Values and pragmatic action: The challenges of introducing ethical intelligence in technical design communities. *The International Review of Information Ethics* 10 (2009), 37–44. https://doi.org/10.29173/irie87
[50] A. Marinho, A. Poulsen, M. Kroesen, and C. Chorus. 2021. Perspectives about Artificial Moral Agents. *AI and Ethics* 1 (2021), 477–490. https://doi.org/10.1007/s43681-021-00055-2
[51] B. Martin and B. Hanington. 2012. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers.
[52] K. Martin. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160, 4 (2019), 835–850.
[53] T. Mattelmäki, E. Brandt, and K. Vaajakallio. 2011. On designing open-ended interpretations for collaborative design exploration. *CoDesign* 7, 2 (2011), 79–93. https://doi.org/10.1080/15710882.2011.609891
[54] A. McNamara, J. Smith, , and E. Murphy-Hill. 2018. Does ACM's code of ethics change ethical decision making in software development? Proceedings of the ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, 729 – 733. https://doi.org/10.1145/3236024.3264833
[55] J. Metcalf, E. Moss, E. Watkins, R. Singh, and M. Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3442188.3445935
[56] Microsoft. 2022. Judgment Call. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall
[57] C. Miller and R. Coldicott. 2019. People, power and technology: The tech workers' view. https://doteveryone.org.uk/report/workersview/
[58] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, D.I. Raji, and T. Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596
[59] S. Mora, F. Gianni, and M. Divitini. 2017. Tiles: A Card-Based Ideation Toolkit for the Internet of Things. Proceedings of the Conference on Designing Interactive Systems (DIS). https://doi.org/10.1145/3064663.3064699
[60] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & Society* (2021). https://doi.org/10.1007/s00146-021-01308-8
[61] E. Moss, E. Watkins, J. Metcalf, R. Singh, and M. Elish. 2021. Governing with Algorithmic Impact Assessments: Six Observations. Proceedings of the ACM Conference on Artificial Intelligence, Ethics and Society (AIES). https://doi.org/10.1145/3461702.3462580
[62] D. Murray-Rust, I. Nicenboim, and D. Lockton. 2022. Metaphors for designers working with AI. In *Proceedings of the Design Research Society Conference (DRS)*. https://doi.org/10.21606/drs.2022.667
[63] NISt. 2016. *AI RISK MANAGEMENT FRAMEWORK*. Retrieved Dec 2023 from https://www.nist.gov/itl/ai-risk-management-framework
[64] E. Olmsted-Hawala. 2006. Card Sorting, Information Architecture And Usability: Adding in Our Users' Perspective to Re-Design the Census Bureau Web Site. Proceedings of the International Professional Communication Conference, 27 – 33. https://doi.org/10.1109/IPCC.2006.320360
[65] A. Palmer and D. Schwan. 2023. More Process, Less Principles: The Ethics of Deploying AI and Robotics in Medicine. *Cambridge Quarterly of Healthcare Ethics* (2023), 1–14. https://doi.org/10.1017/S0963180123000087
[66] D. Peters, L. Loke, and N. Ahmadpour. 2021. Toolkits, cards and games – a review of analogue tools for collaborative ideation. *CoDesign* 17, 4 (2021), 410–434. https://doi.org/10.1080/15710882.2020.1715444
[67] D. Piorkowski, S. Park, A.Y. Wang, D. Wang, M. Muller, and F. Portnoy. 2021. How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 131 (apr 2021), 25 pages. https://doi.org/10.1145/3449205
[68] M. Rahimi, J. Guo, S. Kolaky, and M. Chechik. 2019. Towards Requirements Specification for Machine-Learned Components. Proceedings of the IEEE International Requirements Engineering Conference. https://doi.org/10.1109/REW.2019.00049
[69] M. Ribeiro, S. Singh, and C. Guestrin. n.d.. Lime. https://github.com/marcotcr/lime
[70] R. Roy and J. Warren. 2019. Card-based design tools: a review and analysis of 155 card decks for designers and designing. *Design Studies* 63 (2019), 125–154. https://doi.org/10.1016/j.destud.2019.04.002
[71] M. Sadek, R.A. Calvo, and C. Mougenot. 2023. Co-designing conversational agents: A comprehensive review and recommendations for best practices. *Design Studies* 89 (2023). https://doi.org/10.1016/j.destud.2023.101230
[72] M. Sadek, R.A. Calvo, and C. Mougenot. 2023. Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. *AI & Ethics* (2023). https://doi.org/10.1007/s43681-023-00373-7
[73] C. Sanderson, D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson, and D> Hansen. 2023. AI Ethics Principles in Practice: Perspectives of Designers and Developers. *IEEE Transactions on Technology and Society* 4, 2 (2023), 171–187. https://api.semanticscholar.org/CorpusID:245131189
[74] K. Sekiguchi and K. Hori. 2021. Designing ethical artifacts has resulted in creative design. *AI & Society* 36 (2021), 101−−148. https://doi.org/10.1007/s00146-020-01043-6
[75] Knowledge Center Data & Society. n.d.. Tools. https://data-en-maatschappij.ai/en/tools
[76] N. Stoimenova and M. Kleinsmann. 2020. Identifying and addressing unintended values when designing (with) Artificial Intelligence(with) Artificial Intelligence. Proceedings of the Design Research Society (DRS) International Conference. https://doi.org/10.21606/drs.2020.222
[77] L. Strikwerda, M. van Steenbergen, A. van Gorp, C. Timmers, and J. van Grondelle. 2022. The value sensitive design of a preventive health check app. *Ethics and Information Technology* 24, 38 (2022).
[78] van Merrienboer J.J.G. & Paas F.G.W.C. Sweller, J. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10 (1998), 251−−296. https://doi.org/10.1023/A:1022193728205
[79] M. Tahaei, M. Constantinides, D. Quercia, S. Kennedy, M. Muller, S. Stumpf, V. Liao, R. Baeza-Yates, L. Aroyo, and J. Holbrook. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1–4. https://doi.org/10.1145/3544549.3583178
[80] S. Umbrello. 2022. The Role of Engineers in Harmonising Human Values for AI Systems Design. *Journal of Responsible Technology* 10 (2022).
[81] S. Umbrello and O. Gambelin. 2022. *Agile as a Vehicle for Values: A Value Sensitive Design Toolkit*. Springer.
[82] S. Umbrello and I. Van de Poek. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics* 1, 3 (2021), 283–296. https://doi.org/10.1007/s43681-021-00038-3
[83] I. van de Poel. 2020. Embedding Values in Artificial Intelligence (AI) Systems. *Mind and Machines* 30 (2020), 385–409.
[84] Q. Vera and M. Muller. 2019. Enabling Value Sensitive AI Systems through Participatory Design Fictions. *arXiv preprint arXiv:1912.07381v1* (2019). https://arxiv.org/abs/1912.07381
[85] P. Verbeek, C. Sin, D. Hardoon, M. Johns, and V. Evers. 2021. AI and Ethics: The Key to a Successful Human-AI Relation. https://www.sginnovate.com/events/ai-and-ethics-key-successful-human-ai-relation
[86] O. Vereschak, G. Bailly, and B. Caramiaux. 2022. What AI Practitioners Say about Human-AI Trust: Its Role, Importance, and Factors That Affect It. International Conference on Hybrid Human-Artificial Intelligence.
[87] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. https://doi.org/10.1145/3306618.3314289
[88] R. Wong, M. Madaio, and M. Nick. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023). https://doi.org/10.1145/3579621
[89] N. Yildirim, M. Pushkarna, N. Goyal, M. Wattenberg, and F. Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook. In Proceedings of the CHI Conference on Human Factors in

Computing Systems,.
[90] D. Yoo, A. Huldtgren, J. Woelfer, D. Hendry, and B. Friedman. 2013. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/2470654.2470715
[91] L. Yu and Y. Li. 2022. Artificial Intelligence Decision-Making Transparency and Employees Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behavioral Sciences* 12, 5 (2022). https://doi.org/10.3390/bs12050127
[92] A.X. Zhang, M. Muller, and D. Wang. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 22 (may 2020), 23 pages. https://doi.org/10.1145/3392826
[93] H. Zhu, B. Yu, A. Halfaker, and L. Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons, Vol. CSCW2. Proceedings of the ACM Conference on Human-Computer Interactions. https://doi.org/10.1145/3274463

A SURVEY QUESTIONS

A.1 Pre-Workshop Survey

(1) Please enter your name.
(2) Please enter your age.
(3) Please select your gender.
(4) Please select your role.
(5) How accurate are the following statements?
   - I am familiar with how at least one type of AI-based system works
   - I am familiar with how to build at least one type of AI-based systems
(6) Which of the following applies to you (you can select more than one)?
   - I've created an AI-based system before
   - I've learned about AI-based systems through formal education (i.e. school or university)
   - I've self-learned about AI-based systems through taking a course, studying online, or other activities for at least 6 months
   - It has been my job to create AI-based systems for at least 6 months
   - None of the above.
   - Other (please specify)
(7) How many years of experience do you have with AI-based systems (knowledge-based experience or hands-on experience)?
(8) If you would like to add any comments or extra information about your selections above then please do so here:

A.2 Post-Workshop Open-Ended Survey

(1) Please describe how you incorporated the toolkits in brainstorming activities/steps/considerations for building responsible AI. Did you use any specific strategies?
(2) When and how would you see yourself using the toolkits during your current workflows/tasks/studies?
(3) Did the toolkits help you overcome any specific challenges?
(4) Were there any challenges where the toolkits did not help, or any support you needed that they did not provide?
(5) Do you have suggestions or areas for improvement with the toolkits?

B TOOLKITS CONSIDERED FOR INCLUSION

B.1 List of Toolkits Considered

The toolkits considered from inclusion were obtained from Wong et al. (2022) [88]'s previous work and an online repository of practical tools for building responsible AI [75], in addition to applicable RAI toolkits created in 2023, and are shown in Table 3.

B.2 Toolkit Exclusion Process

Figure 7 shows a list of toolkits considered during each round of the exclusion process, along with the total number of toolkits considered during each round and which toolkits were excluded.

C DESIGN RECOMMENDATIONS

Figure 8 summarises the six design recommendations for creators of RAI toolkits.

Table 3: The list of toolkits that were reviewed as part of the study, along with their source (at the time of writing)

Number	Toolkit	Source
1	Aequitas	http://aequitas.dssg.io/
2	AI Assessment Tool	https://altai.ai4belgium.be/nl
3	AI Ethics Cards	https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers
4	AI Explainability 360 Open Source Toolkit	http://aix360.mybluemix.net/
5	AI Fairness 360	https://aif360.mybluemix.net/
6	AI Maturity Tool	https://ai.digimaturity.vtt.fi/?lang=en
7	AI Meets Design Toolkit	https://www.aixdesign.co/toolkit
8	AI System Ethics Self-Assessment Tool	https://www.smartdubai.ae/self-assessment
9	Algorithmic Accountability Policy Toolkit	https://ainowinstitute.org/aap-toolkit.pdf
10	Algorithmic Equity Toolkit (AEKit)	https://www.aclu-wa.org/AEKit
11	Artifical Intelligence Impact Assessment	https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf
12	Audit AI	https://github.com/pymetrics/audit-ai
13	Building an Algorithm Tool	https://www.cdt.info/ddtool/
14	Cards for Humanity	https://cardsforhumanity.idean.com/
15	Community Jury	https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/
16	Consequence Scanning Kit	https://www.doteveryone.org.uk/project/consequence-scanning/
17	Create your Own Datawalk	https://data-en-maatschappij.ai/uploads/VUB-Datawalk-gids-ENG-v3-digitaal-paginas.pdf
18	Data Cards Playbook	https://pair-code.github.io/datacardsplaybook/
19	Data Collection Bias Assessment	https://data-en-maatschappij.ai/en/tools/tool-data-collection-bias-assessment-form
20	Data Ethics Canvas	https://theodi.org/article/data-ethics-canvas/
21	Data Ethics Decision Aid	https://dataschool.nl/deda/?lang=en
22	Data Ethics Framework	https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework
23	Data Ethics Guide	https://www.cigref.fr/wp/wp-content/uploads/2019/02/Cigref-Syntec-Digital-Ethics-Guide-for-Professionals-of-Digital-Age-2018-October-EN.pdf
24	Deon Ethics Checklist	http://deon.drivendata.org/
25	Design Ethically Toolkit	https://www.designethically.com/toolkit
26	The Digital Ethics Compass	https://ddc.dk/tools/toolkit-the-digital-ethics-compass/#compass
27	Digital Impact Toolkit	https://digitalimpact.io/toolkit/
28	Digital Inclusion Card Game	https://data-en-maatschappij.ai/index.php?p=actions/asset-count/count&id=184990
29	Dynamics of AI Principles	https://aiethicslab.com/big-picture/
30	Ethical Explorers Pack	https://ethicalexplorer.org
31	Ethical OS Toolkit	https://ethicalos.org/
32	Ethics & Algorithms Toolkit	https://ethicstoolkit.ai/
33	Ethics Framework van Machine Intelligence Garage	https://futurescope.digicatapult.org.uk/wp-content/uploads/2023/04/DC_AI_Ethics_Framework-2021.pdf
34	Ethics Inc.: A Design Game for Ethical AI	https://www.ethicsinc-ontwerpspel.nl/ethisch-ontwerpspel-voor-ai/
35	Ethics Kit	http://ethicskit.org/tools.html
36	Fairlearn	https://fairlearn.github.io/
37	Guidance Ethics	https://ecp.nl/wp-content/uploads/2019/11/060-001-Boek-Aanpak-begeleidingsethiek-240165-binnenwerk-digitaal.pdf
38	Harms Modeling	https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/
39	HAX Workbook and Playbook	https://www.microsoft.com/en-us/haxtoolkit/workbook/
40	Intelligence Augmentation Design Toolkit	https://futurice.com/ia-design-kit
41	InterpretML	https://github.com/interpretml/interpret
42	Judgment Call	https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall
43	Lime	https://github.com/marcotcr/lime
44	LinkedIn Fairness Toolkit (LiFT)	https://github.com/linkedin/LiFT , https://engineering.linkedin.com/blog/2020/lift-addressing-bias-in-large-scale-ai-applications
45	MIT AI Blindspots	https://aiblindspot.media.mit.edu/
46	Model Cards	https://modelcards.withgoogle.com/about
47	Nokia AI Design Toolkit	https://bell-labs.com/rai-prompts/
48	NLP CheckList	https://github.com/marcotcr/checklist
49	People+AI Guidebook	https://pair.withgoogle.com/guidebook/
50	Principles for Accountable Algorithms en Social Impact Statement for Algorithms	https://www.fatml.org/resources/principles-for-accountable-algorithms
51	Product Impact Tool	https://productimpacttool.org/nl/portal/
52	RAI Toolkit	https://rai.tradewindai.com/
53	Responsible AI in Consumer Enterprise	https://static1.squarespace.com/static/5d387c126be524000116bbdbt/5d77e37092c6df3a5151c866/1568138185862/Ethics-of-artificial-intelligence.pdf
54	Responsible AI Diagnostic	https://pwc.qualtrics.com/jfe/form/SV_0UF8EgBJdAnV8fr
55	SageMaker Clarify	https://sagemaker-examples.readthedocs.io/en/latest/sagemaker_processing/fairness_and_explainability/fairness_and_explainability.html
56	SDoC for AI/AI Servcie FactSheets	Arnold et al. (2019), "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," in IBM Journal of Research and Development, vol. 63, no. 4/5
57	The Tarot Cards of Tech	https://www.artefactgroup.com/wp-content/uploads/2018/10/Artefact-Tarot-Cards-of-Tech_downloadable.pdf
58	TensorFlow Fairness Indicators	https://github.com/tensorflow/fairness-indicators
59	Unbias Toolkit	https://unbias.wp.horizon.ac.uk/fairness-toolkit/
60	Weights and Biases	https://wandb.ai/site
61	What If Tool	https://pair-code.github.io/what-if-tool/

Original Corpus (63 toolkits)	Round 1 - Target Users (38 toolkits)	Round 2 - Focus on Regulation (21 toolkits)	Round 3 - Indication of Use (6 toolkits)	Round 4 - Comparability (2 toolkits)
Aequitas	MIT AI Blindspots	Data Cards Playbook	MIT AI Blindsponts	MIT AI Blindsponts
AI Assessment Tool	Aequitas	AI Explainability 360 Open Source Toolkit	HAX Workbook and Playbook	Nokia AI Design Toolkit
AI Ethics Cards	AI Assessment Tool	AI Fairness 360	Judgment Call	
AI Explainability 360 Open Source Toolkit	AI Explainability 360 Open Source Toolkit	Audit AI	Nokia AI Design Toolkit	
AI Fairness 360	AI Fairness 360	Cards for Humanity	The Digital Ethics Compass	
AI Maturity Tool	AI Meets Design Toolkit	Community Jury	The Tarot Cards of Tech	
AI Meets Design Toolkit	AI System Ethics Self-Assessment Tool	Data Cards Playbook		
AI System Ethics Self-Assessment Tool	Artifical Intelligence Impact Assessment	Deon Ethics Checklist		
Algorithmic Accountability Policy Toolkit	Audit AI	Digital Inclusion Card Game		
Algorithmic Equity Toolkit (AEKit)	Building an Algorithm Tool	Harms Modeling		
Artifical Intelligence Impact Assessment	Cards for Humanity	HAX Workbook and Playbook		
Audit AI	Community Jury	Intelligence Augmentation Design Toolkit		
Building an Algorithm Tool	Data Cards Playbook	InterpretML		
Cards for Humanity	Data Collection Bias Assessment	Judgment Call		
Community Jury	Data Ethics Canvas	LinkedIn Fairness Toolkit (LiFT)		
Consequence Scanning Kit	Data Ethics Decision Aid	MIT AI Blindspots		
Create your Own Datawalk	Data Ethics Framework	Model Cards		
Data Cards Playbook	Data Ethics Guide	Nokia AI Design Toolkit		
Data Collection Bias Assessment	Deon Ethics Checklist	TensorFlow Fairness Indicators		
Data Ethics Canvas	Digital Inclusion Card Game	The Digital Ethics Compass		
Data Ethics Decision Aid	Ethical OS Toolkit	The Tarot Cards of Tech		
Data Ethics Framework	Ethics & Algorithms Toolkit			
Data Ethics Guide	Fairlearn			
Deon Ethics Checklist	Harms Modeling			
Design Ethically Toolkit	HAX Workbook and Playbook			
Digital Impact Toolkit	Intelligence Augmentation Design Toolkit			
Digital Inclusion Card Game	InterpretML			
Dynamics of AI Principles	Judgment Call			
Ethical Explorers Pack	Lime			
Ethical OS Toolkit	LinkedIn Fairness Toolkit (LiFT)			
Ethics & Algorithms Toolkit	Model Cards			
Ethics Framework van Machine Intelligence Garage	Nokia AI Design Toolkit			
Ethics Inc.: A Design Game for Ethical AI	SDoC for AI/AI Servcie FactSheets			
Ethics Kit	TensorFlow Fairness Indicators			
Fairlearn	The Digital Ethics Compass			
Guidance Ethics	The Tarot Cards of Tech			
Harms Modeling	Weights and Biases			
HAX Workbook and Playbook	What If Tool			
Intelligence Augmentation Design Toolkit				
InterpretML				
Judgment Call				
Lime				
LinkedIn Fairness Toolkit (LiFT)				
MIT AI Blindspots				
Model Cards				
NLP CheckList				
Nokia AI Design Toolkit				
People+AI Guidebook				
Principles for Accountable Algorithms en				
Social Impact Statement for Algorithms				
Product Impact Tool				
RAI Toolkit				
Responsible AI Diagnostic				
Responsible AI in Consumer Enterprise				
SageMaker Clarify				
SDoC for AI/AI Servcie FactSheets				
TensorFlow Fairness Indicators				
The Digital Ethics Compass				
The Tarot Cards of Tech				
Unbias Toolkit				
Weights and Biases				
What If Tool				

Figure 7: Lists of the toolkits considered during each round of the exclusion process as described in Section 3. Each column shows the toolkits considered during that round. Toolkits highlighted in grey are the ones that were excluded in each round.

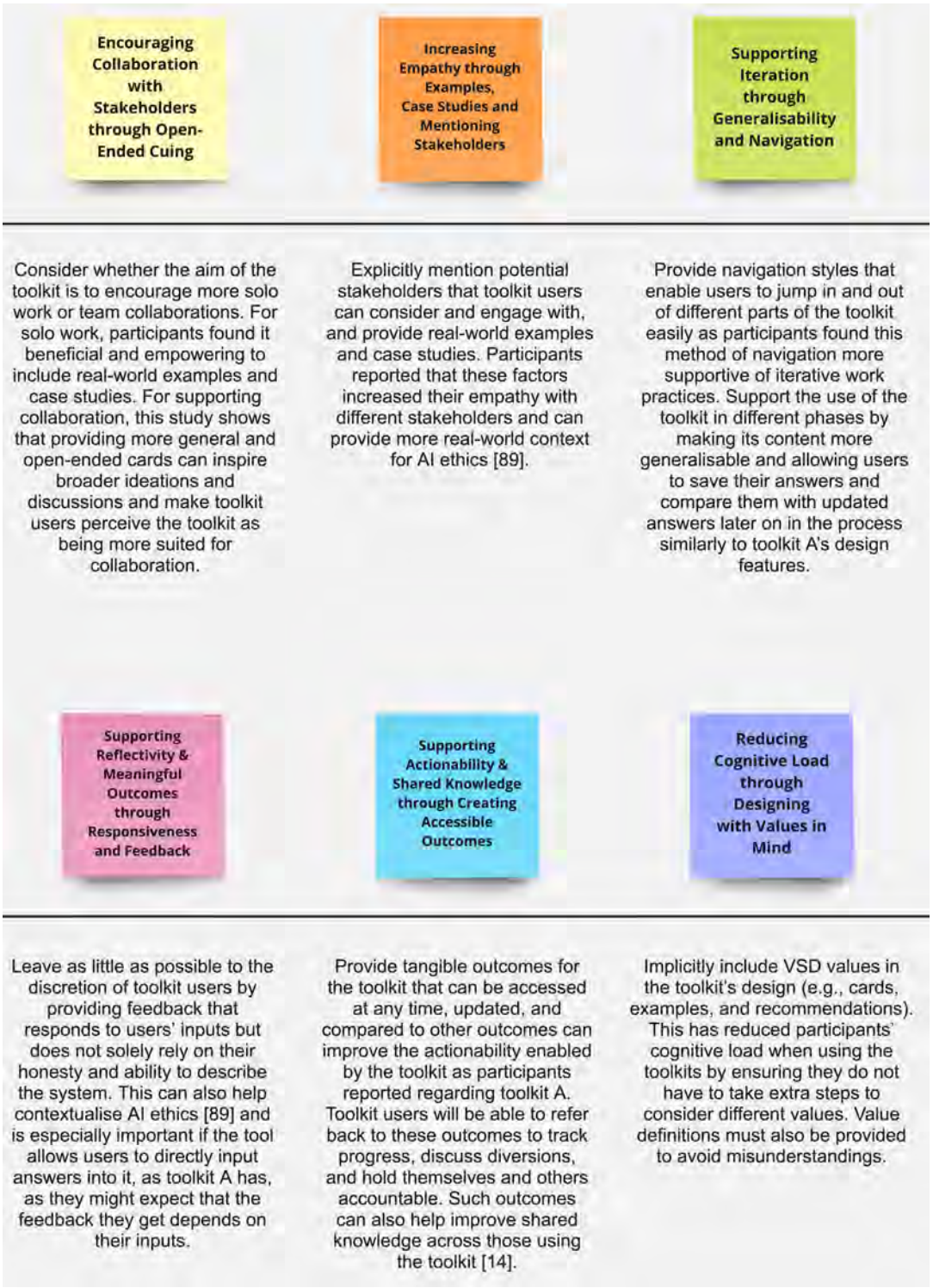


Figure 8: Six design recommendations we synthesised from our research. Each design recommendation is included in a sticky note with a detailed description included in the text under each sticky note.