Digital ethics by design–a comprehensive evaluation of the design for values approach in practice

Sattlegger, Antonia; Alleblas, Joost; van de Poel, Ibo

DOI
10.1080/23299460.2025.2534273

Publication date
2025

Document Version
Final published version

Published in
Journal of Responsible Innovation

Citation (APA)
Sattlegger, A., Alleblas, J., & van de Poel, I. (2025). Digital ethics by design–a comprehensive evaluation of the design for values approach in practice. *Journal of Responsible Innovation*, *12*(1), Article 2534273. https://doi.org/10.1080/23299460.2025.2534273

Routledge
Taylor & Francis Group

Journal of Responsible Innovation

ISSN: 2329-9460 (Print) 2329-9037 (Online) Journal homepage: www.tandfonline.com/journals/tjri20

Digital ethics by design – a comprehensive evaluation of the design for values approach in practice

Antonia Sattlegger, Joost Alleblas & Ibo van de Poel

To cite this article: Antonia Sattlegger, Joost Alleblas & Ibo van de Poel (2025) Digital ethics by design – a comprehensive evaluation of the design for values approach in practice, Journal of Responsible Innovation, 12:1, 2534273, DOI: 10.1080/23299460.2025.2534273

To link to this article: https://doi.org/10.1080/23299460.2025.2534273

Published online: 13 Oct 2025.

JOURNAL OF RESPONSIBLE INNOVATION
2025, VOL. 12, NO. 1, 2534273
https://doi.org/10.1080/23299460.2025.2534273

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Digital ethics by design – a comprehensive evaluation of the design for values approach in practice

Antonia Sattlegger, Joost Alleblas and Ibo van de Poel

Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands

ABSTRACT

Many guidelines outline ethical principles for designing and deploying emerging digital technologies, like AI, in public services, but there is a gap between such principles and practices. We evaluate whether an educational intervention can enable public sector professionals to close this gap and implement responsible innovation. The educational intervention was based on Design for Values, a responsible innovation approach to integrate values into the design process. We employ a systems perspective to evaluate the effects of the intervention. While the educational intervention helps foster techno-moral virtues and enhance accountability, its success depends on the broader organizational context. Future research should explore the long-term embedding of Design for Values in various settings, using comparative and longitudinal methods to understand better the factors that influence its effectiveness.

ARTICLE HISTORY
Received 31 March 2024
Accepted 10 July 2025

KEYWORDS
Designing for values; AI ethics; digital ethics; emerging digital technologies; public sector innovation

Introduction

There is no shortage of guidelines, principles, and manifestos articulating values for the responsible design of emerging digital technologies in the public sector, like artificial intelligence (AI). Jobin, Ienca, and Vayena (2019) found 84 documents pertaining to principles and guidelines for responsible AI by public sector organizations, research institutions, and private companies. Despite the wealth of these guidelines, they have little impact on the responsible design of AI in practice (McNamara, Smith, and Murphy-Hill 2018).

While the importance of responsible design of AI for public services is widely acknowledged, there is a disconnect between the emphasis on ethical principles and the societal impacts of the actual design and deployment of technology in public service delivery (James and Whelan 2022). Rather than focusing on these actual impacts, societal impacts are typically framed within ethical AI principles and implicit values, often discussed as ancillary benefits or challenges when prioritizing service and duty-oriented values (Madan and Ashok 2023).

For one, such principles and guidelines concerning, for instance, privacy, accountability, or fairness do not consider AI systems in a more holistic sense as systems within systems (Hagendorff 2020). These principles, therefore, fail to acknowledge the context and situatedness of AI systems while providing few technical explanations and examples (Hagendorff 2020). Second, while AI development and deployment might bear similarities to other professional fields in which normative principles and guidelines were successfully introduced, the differences seem more critical (Mittelstadt 2019). As examples of such differences, Mittelstadt (2019) mentions the absence of shared goals and professional history and the absence of proven methods to put principles into practice. Finally, the universal and uncompromising formulation of normative principles for AI neglects other normative and legal frameworks developers, operators, and contractors have to deal with when designing and deploying AI systems (McNamara, Smith, and Murphy-Hill 2018).

The responsible design and deployment of emerging digital technologies is crucial in the public sector because of the direct impact of these technologies on (vulnerable) citizens' lives (Wirtz, Weyerer, and Sturm 2020). Ensuring that technological design is based on values like fairness, transparency, and

CONTACT Antonia Sattlegger a.s.sattlegger@tudelft.nl TU Delft Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX, Delft, Netherlands

accountability is essential for maintaining public trust (Grimmelikhuijsen and Meijer 2022; Meijer and Grimmelikhuijsen 2020; Selten and Meijer 2021; Veale, Van Kleek, and Binns 2018). However, the increased adoption of AI in public service delivery in the Netherlands has produced some prominent negative examples, such as the use of biased algorithmic fraud detection systems in social welfare distribution (Amnesty International 2021; Giest and Klievink 2024; Sattlegger, van den Hoven, and Bharosa 2022) or the use of third-party algorithms for visa processing (Maleeyakul et al. 2024). Instead of enhancing the public good, these technologies have exacerbated inequalities and contributed to growing distrust in the government's use of digital technologies, particularly AI. This highlights the urgent need for a more responsible approach to designing and implementing AI in the public sector.

We are interested in how approaches from the field of Responsible Innovation (RI) may help to overcome these problems. RI has developed into a broad research field, and many RI approaches are available (Fisher et al. 2024). Though the RI literature has emphasized the need for a systemic approach (Stilgoe, Owen, and Macnaghten 2013), such a systemic approach was only exceptionally applied in subsequent empirical studies (Herzog and Blank 2024; Neudert, Smolka, and Böschen 2024). By applying a systemic lens of an educational intervention within a broader political-administrative (eco)system, we seek to contribute to this emerging body of literature.

Public sector organizations' responsible design of emerging digital technologies requires particular attention to the contextual conditions and organizational dynamics specific to or amplified in governmental organizations. Emerging digital technologies are designed, adopted, and governed as part of broader public service provision within the public-centered ecosystem (Naughton, Dopson, and Iakovleva 2023). Therefore, we take a system perspective by conceiving of the system in which a specific RI approach is implemented or supported as a socio-technical system or an innovation (eco)system (Kudina and van de Poel 2024; Smolka and Böschen 2023; Stahl 2021, 2023). This means that successful implementation will require more than a simple intervention, but rather a range of more substantive changes that align and change the system durably.

The intervention we study is a dedicated course on Design for Values (DfV) taught in-company to governmental professionals developing and designing IT systems. DfV is an approach similar to Value Sensitive Design (VSD), focused on fostering attention to values in the design process. A focus on education for professionals helps make them aware that a range of changes is required to implement an RI approach, like DfV, successfully.

Our guiding research question is: To what extent can an educational program support IT professionals in designing responsible emerging digital technologies for the public sector to help better address the contextual conditions and organizational dynamics specific to government organizations within a broader political-administrative ecosystem?

We answer this research question by evaluating the application of the DfV approach in a professional education course at a governmental IT organization in the Netherlands. The professionals were guided in applying the DfV approach to reflect on the values, value conflicts, and design requirements of proactive public service provision using an AI-based application. While we analyze an educational intervention for the responsible design of AI in the Dutch public sector, our findings extend beyond the domain of AI to the responsible design of emerging digital technologies in the public sector context. We evaluate whether the DfV approach helped course participants bridge the gap between abstract moral principles and concrete practice by integrating the DfV approach into their daily practices, including designing such systems, commissioning, and discussions with clients. While we use an educational program as a case, this research is not about evaluating educational training. Instead, it analyses the educational program to allow participants to apply DfV approaches in practice. It is not our intention to evaluate the educational intervention itself, as others have done before us (Bardone, Burget, and Pedaste 2023; Marschalek et al. 2017; Richter, Hale, and Archambault 2019; Spruit 2014; Stahl et al. 2023; Tomblin and Mogul 2020), but rather the impact of the intervention on participants ability to engage with values in the design of emerging digital technologies, as well as how the organizational dynamics specific to governmental contexts facilitate and impede this engagement.

In the coming sections, we proceed as follows: we first present the underlying RI theory, the DfV approach, and the systems perspective we employ; second, we present the case and evaluation method; third, we provide the results and subsequently discuss and conclude our findings.

Theory

RI aims to align technology with society's values, needs, and expectations (European Commission 2019). It challenges a linear view of technological development, in which technology is developed through technology push or market demand. Instead, it is aimed at democratizing the process of technology development.

Four core elements of RI are inclusiveness (including relevant stakeholders and values), anticipation (of the societal consequence of new technology), reflexivity (reflecting on underlying values, goals, and assumptions), and responsiveness (responding to societal needs and issues, and to new developments) (Stilgoe, Owen, and Macnaghten 2013). In this section, we explain the theories from the RI field that we have used. This is, first of all, DfV as the RI approach underlying our educational intervention, and second, a system perspective on RI to evaluate the effects of this intervention.

Design for values

RI builds on insights from Technology Assessment (TA), particularly TA approaches that aim to provide feedback insights and anticipations of social consequences in the development process, such as constructive technology assessment (Schot and Rip 1997) and real-time technology assessment (Guston and Sarewitz 2002). Over time, various RI approaches have developed, including Socio-Technical Integration Research (STIR), VSD, DfV, and ethics-by-design. All these approaches have a similar overall goal, i.e., making technological development and deployment more responsible, but their focus is different.

In our educational intervention, we used the DfV approach because we have hands-on experience implementing this approach in the Netherlands.

Before we proceed and explain the DfV approach in more detail, a word on terminology might be helpful. In this article, we conceive values as abstract expressions of what is good and desirable, such as justice, well-being, safety, and privacy. Moral principles, or simply principles, express certain (moral) obligations, such as 'do not lie' or 'do not discriminate' or 'treat people equally.' Such principles may be seen as specifications of more general values (van de Poel 2013), and the DfV approach allows for translating them into more specific design and system requirements. Moral or ethical virtues should be distinguished from both values and principles; they are human character traits that are morally good or praiseworthy, like honesty, perseverance, and integrity.

DfV builds on similar approaches, particularly VSD, developed by Batya Friedmann and colleagues in the late 1980s (Friedman and Hendry 2019). VSD and DfV assume that some ethical issues that technologies might raise during use and operation can be meaningfully addressed during the design phase, mainly through embedding specific positive values in technological design.

DfV processes may be depicted as going through four phases (Umbrello and van de Poel 2021):

(1) Context analysis
(2) Value identification
(3) Design
(4) Prototyping

These iterative phases proceed from an analysis of the context in which a new technology must function to the identification of relevant values to design for, to the actual design process, including the translation of values into design requirements and dealing with conflicting values, to the building and testing of a prototype. The latter phase might generate new insights that may feed into another process iteration.

Evaluating the effect of the intervention: a system perspective

While DfV is a promising approach to implementing values of ethical concern into the design process of emerging digital technologies, such as AI-based systems, there is little research that evaluates how interventions based on the DfV framework (or similar frameworks like VSD or ethics-by-design), such as an educational intervention, can affect socio-technical systems, such as in our case the Dutch governmental system of AI development and deployment.

Approaches like DfV and VSD have been applied to a wide range of technologies including, for example, wind turbine parks (Künneke et al. 2015; Oosterlaken 2014), civil healthcare drones (Cawthorne 2023), biorefineries (Parada 2020), digital government collaborative platforms for sustainability (Sapraz and Han 2021), and care robots (Poulsen 2022). However, the broader effects of such interventions on socio-technical systems have not been systemically considered.

Therefore, this study employs a system perspective on the design and deployment of specific emerging technologies that contribute to digitalization in the public sector. We mainly focus on a sub-system of a government IT service provider and government organizations that design digital technologies within a broader political-administrative ecosystem. This means neither problems nor solutions can be isolated as a concern for technological design alone. Instead, problems and solutions must be understood from an 'innovation ecosystem' perspective in which emerging technologies, such as responsible AI, can flourish (Stahl 2021). A wide range of stakeholders must be included in the design process, especially as salient values are selected (Werker 2020). These stakeholders are interdependent with other agents and operate under formal and informal institutions (Smolka and Böschen 2023; Stahl 2021; Werker 2020).

Furthermore, a system perspective considers the possibility and availability of a supporting framework or infrastructure for the proposed values incorporated in emerging technologies contributing to digitalization in the public sector. This means that support and deliberation must be organized at the ICT-organizational level and institutionally at the government level. Indeed, policymakers have an essential role in providing framework conditions for the success of a responsible design project, such as responsible AI (Smolka and Böschen 2023).

Organizational cultures, such as transformational leadership, creativity, innovation, institutional support, and openness to change, have been found to impact RI's translation into organizational practice critically (Owen et al. 2021; Pansera et al. 2020). However, in the public sector, the unique organizational dynamics – characterized by multiple conflicting accountability demands, hierarchical decision-making, and a strong emphasis on serving political principles (Crosby, 't Hart, and Torfing 2017) – can hinder the responsible design of emerging digital technologies. In such a context, where collective responsibilities to address public and moral values are often dispersed across various administrative subunits, individual responsibility may be diluted, making it challenging for practitioners to embrace RI's principles fully. Educational interventions are, therefore, crucial in cultivating the techno-moral virtues necessary for public sector professionals to navigate these complexities (Steen, Sand, and van de Poel 2021; van de Poel and Sand 2021). By fostering a reflexive mindset – encouraging openness to uncertainty, vulnerability, and critical questioning (Steen 2021) – and empowering practitioners to become proactive, responsible innovators (Smolka and Fisher 2024), education can play a pivotal role in ensuring that RI practices are not only adopted but are also effectively integrated into organizational norms and practices.

Further, adopting DfV approaches to design emerging digital technologies can address the need for accountability within public sector organizations. By providing a systematic and explainable process for translating values into design choices, DfV approaches offer a clear framework that guides responsible innovation and makes the decision-making process more transparent and accountable.

Method

To answer our research question of how an educational intervention based on DfV can bring about systemic change and support designers in moving from abstract principles to concrete practices for emerging digital technologies in the public sector, we evaluated the impact of a course designed to teach DfV to professionals in the IT domain.

Course and case design

The course focused on the responsible design of AI-based emerging digital technologies in the public sector. The participants systematically applied DfV methods in two hypothetical AI use cases:

First, an AI-driven smart electric vehicle charging system that optimizes charging schedules to promote renewable energy use and enhance grid stability. The system uses inputs like grid capacity, electricity prices, and solar energy availability alongside vehicle data, such as battery levels, shared or private vehicles, and user preferences.

Table 1. Course overview.

	Course Overview	Value	Exercise
1	Introduction to ethics of technology, responsible innovation, and designing for values	Introduction in ethics, DfV method, and case	Defining the design challenge. Identifying the relevant direct and indirect stakeholders, their interests, and the values behind those interests.
2	Value scenarios	Fairness	Working with value scenarios: Outline scenarios that address the design challenge you have identified. Discuss these scenarios using the Envisioning Cards.
3	Investigating values	Meaningful human control, responsibility	Conceptualization of value investigations. Defining conceptual, technical, and empirical investigations.
4	Translating values into design requirements	Transparency	Translating values into socio-technical design requirements: Identify a relevant value in your case. Construct a value hierarchy for that value.
5	Value conflicts	Transparency, privacy	Managing value conflicts – There are four strategies for dealing with value conflicts: (1) Calculative approaches, (2) Satisficing, (3) Respecification, and (4) Innovation.
6	Values in the use phase	Contestability, value change	Embedding designing for values methods in an organizational context.

Second, an AI-driven app to proactively notify citizens about their eligibility for public services, such as social benefits. The underlying algorithms identify and predict citizens' eligibility for social services based on citizens' personal and socio-economic information, such as income, employment status, household composition, or geographic location. Various value-laden design choices determine the impact of this AI application on proactive public services. Notably, the degree of proactivity is determined by the citizen's effort required in the eligibility determination and the delivery process (Bharosa et al. 2021; Scholta and Lindgren 2023).

The course aims were twofold. One is to raise awareness among the participants about the significance of values in the digital domain. Two, to enable participants to integrate DfV approaches into their daily practices to identify and effectively manage conflicting values in designing emerging digital technologies to deliver public services. Professionals were provided with methodological knowledge of DfV in the context of responsible design for emerging digital technologies.

The authors were responsible for developing the course design, facilitating the training, and guiding the evaluation. Two iterations of the course were conducted between March 2022 and March 2024. Each course consisted of six sessions over three months, each lasting four hours. Each session consisted of a teaching part: (1) a methods section introducing the DfV approach and instruments, and (2) a theory section on ethical conceptions of different values in the digital domain, such as fairness, transparency, responsibility, or contestability. Based on this input, participants entered a design part. They applied the DfV approaches to the practical case in sub-groups through exercises. Each session ended with a reflection: the participants were guided in reflection-on-action to collectively reflect on the applicability of the DfV approaches on the case and their projects and organization. An overview of this course can be found in Table 1.

Participants

The training was given at a governmental IT organization in the Netherlands. The organization provides digital solutions for public sector entities. It collaborates with various government agencies and bodies to provide strategic advice on developing and implementing IT projects to improve efficiency, service delivery,

Table 2. List of workshop participants anonymized.

	Participants
1	Project Manager
2	Advisor Data Science & AI
3	Strategic Advisor/ Enterprise Architect
4	Interaction Designer
5	eGovernment Architect
6	(Strategic) Advisor and Project Leader
7	Strategic Advisor
8	Organizational Advisor
9	Government Information Advisor
10	Project Manager
11	ICT Architect
12	Director

and digital transformation within the public sector. As such, the participants collaborate with various government stakeholders within a broader political-administrative ecosystem. The participants were a diverse group of data scientists, strategic advisors, and managers. An overview of the anonymized participants can be found in Table 2. Participation in the course was voluntary and thus self-selected, though some participants were proactively encouraged by their supervisors.

This research involving human participants has undergone ethical review and received approval from the Human Research Ethics Committee to ensure compliance with ethical standards and safeguard the well-being and rights of all participants.[1] Prior to their involvement, all participants were fully informed about the nature and purpose of the research. They voluntarily provided their informed and written consent to participate in this research and subsequent publication.

Data collection and analysis

We conducted several evaluations to research whether the course impacted the designers' ability to design for values and to close the gap between principles and practices. First, this research draws on a participatory evaluation approach. The authors observed the sessions in the formative evaluation, particularly the design exercises. Second, after every session, the participants were asked to reflect collectively on the substantial and methodological input and exercises and to relate these to their work routines. The authors guided the reflection. Second, in the summative evaluation, we conducted a survey directly after the course. Third, after one year for one group and several months for the other group, we conducted a focus group with the participants to evaluate the course and its (enduring) impact on the participants. Because of the elapsed time, we expected to discern whether the course had a more lasting impact or not. The evaluative workshop consisted of five main parts. An overview of the summative evaluation is provided in Table 3. The data was collected through material produced by the participants, such as canvases used in the exercises, field notes taken by the authors, and recordings of the evaluative workshop.

Individual reflection: After a presentation to recall the course, participants were asked to reflect on the course and their application.

Initial assessment: Participants were presented with thirteen statements concerning the course and its perceived impact. For each statement, they were asked to indicate agreement or disagreement by raising either a red or green card. In addition to this binary choice, participants were invited to briefly comment on their response. In several cases, participants raised both cards, offering reasoning for both agreeing and disagreeing with the statements. This was not an indication of neutrality or uncertainty, but rather to reflect ambivalence or context- or statement-specific reasoning. Importantly, discussion among participants was not permitted during this phase to ensure that all initial responses reflected individual perspectives. A summary of the statements and results can be found in Table 4.

Reflective dialogue: Participants were paired up based on divergent answers to the statements. In these pairs, participants reflected on and discussed the following questions:

- When have you applied insights from the course in practice?
- When were you able to apply the methods, or are you not able to?
- What obstacles hinder application?

After this reflective dialogue, the results were shared and discussed collectively.

Visioning and Ideation: Participants were grouped to envision an ideal learning intervention based on their reflections from the previous step. They were asked to reflect on their learning needs and desires regarding applying DfV approaches.

Table 3. Overview summative evaluation.

	Summative evaluation steps
1	Individual open reflection: Note three specific insights, learning experiences, or activities that were most valuable to you.
2	Initial assessment by testing evaluative statements: Do you agree or disagree with the following statements?
3	Open reflection in pairs: When have you applied insights from the course in practice? When have you not applied insights? What impedes the application in practice?
4	Open reflection in quadruplets: What does the ideal masterclass on designing for values look like?
5	Collective discussion.

Table 4. Evaluative statements.

	Statement	Agree	Both	Disagree
1	Through the course, I have the feeling that I know more about the ethical aspects that play a role in the development of data applications.	7	1	1
2	The course has given me practical tools for my daily tasks.	2	3	4
3	I integrate aspects of the course into the adoption and/or execution of assignments.	6	2	1
4	I can raise ethical aspects with the client and articulate what I need from them.	6	3	0
5	I can raise ethical aspects within my organization – whether with colleagues or within my team – and articulate what I need from them.	7	2	0
6	As a result of the course, I feel more responsible for the ethical aspects of assignments or projects.	3	2	4
7	I have control or influence over ethical aspects in assignments or projects and can assume this responsibility as part of my role.	6	3	0
8	As a result of the course, I now find myself asking questions I did not use to ask – such as critical questions about underlying values – more often.	3	1	5
9	As a result of the course, I now consult more frequently with my team or client about ethical values.	2	0	7
10	In my opinion, several essential matters have not been adequately addressed within the course.	0	4	5
11	I need more knowledge about digital ethics, for example, through a refresher course.	2	3	4
12	I need more practical experience (learning by doing) to effectively apply the methods from the course.	6	0	3
13	I need more support from my organization, my supervisors, or colleagues to apply what I've learned to my tasks.	0	2	7

Collective Discussion: The workshop concluded with a collective discussion on the usability, desirability, limits, and opportunities of the DfV approach. Participants shared insights, feedback, and suggestions for improvement, fostering mutual learning and understanding.

Overall, the workshop allowed participants to assess their understanding and application of DfV approaches, engage in reflective dialogue, envision future learning interventions, and collectively discuss the strengths and challenges of the DfV approach.

Our finding must be seen in the light of several limitations. Participation in this course was voluntary, and participants were primarily self-selected. Therefore, both the awareness of moral values and value conflicts in the design of digital applications were high, as was the motivation to address and apply the DfV approach in practice. Therefore, our findings must be considered based on a self-selection bias. Further, we, the authors, were involved in this course's design, implementation, and evaluation. This provided us with in-depth insights, interaction with the participants, and opportunities for ongoing formative evaluation. However, this involvement shapes our perspective and the reflection of the participants. A further limitation is the researchers' bias regarding the salience of values for responsible digitalization in the public sector. This bias was most present in the selection of theoretical lectures presented to the participants. These lectures often focused on one or two values deemed relevant for the case study the participants worked with. Although these values were presented as examples of values potentially relevant to the case and exemplifying how to move from conceptual questions to design decisions, these lectures might have influenced participants in their discussions and selection of salient values during their case study work.

Results

In this section, we analyzed how the achievement of the objectives of the course – raising awareness and teaching participants to apply the methods in practice – was experienced by participants. We related this experience to success factors in a didactic and organizational sense. The course was developed to facilitate two main objectives: First, to raise awareness among the participants of the importance of values in the digital domain. Second, to teach participants how to use DfV approaches in their daily work, identify and operationalize values, and deal with value conflicts in designing digital applications to deliver public services. Building on the formative evaluation of the course, we operationalized these goals into evaluative statements reflecting the potential impact the course may have (had) on the participants. We formulated thirteen such statements (see Table 4). Participants were asked to agree or disagree with the statements and, occasionally, provided explanations. Below, we provide these explanations. For reasons of anonymity, participants received the pronouns they/them.

Individual awareness and reflection

Conceptual knowledge – A prerequisite to raising the participants' awareness was conceptual knowledge about values in the digital domains. Participants agreed that they became more knowledgeable about ethical

aspects in the development of data applications (answers to statement 1 (7A|1B|1D), see Table 4). One participant (participant 2) disagreed on the scope of 'data applications'; instead, they emphasized that the ethical aspects and actions were relevant to them beyond the digital domain. Another participant emphasized that the ethical aspects, dilemmas, and methods were relevant in governance, political, and societal debates. These features were relevant to them as professionals but more so as individuals (participant 10). Another participant (participant 11) emphasized that beyond 'just knowledge,' the course provided them with 'more awareness' of ethical aspects about which they already had 'suspicions' or 'a gut feeling' (participant 5). They said these ethical aspects were 'more explicit' to them and that they encountered them 'at more different places' (participant 11). We validated these answers by asking whether the participants felt they required additional knowledge concerning ethical aspects after the course (answers to statements 10 (0A|4B|5D) and 11 (2A|3B|4D), see Table 4). Generally, the participants agreed that they did not need additional conceptual knowledge. However, one participant (participant 3) said they would have liked more insights into (meta-)philosophical debates about the nature of values.

Individual reflection – Asked whether the course stimulated the participants to ask themselves more critical questions than before about underlying values in their work, a slight majority disagreed (answers to statement 8 (3A|1B|5D), see Table 4). The participants emphasized that they already critically reflected on their work before the course. However, one participant remarked that they became more aware of the scope and impact of ethical dilemmas (participant 12).

Effect on perceived responsibility – We expected increased awareness and reflection to impact participants' perceived responsibility (answers to statement 6 (3A|2B|4D), see Table 4). While half of the participants agreed, others mentioned they already felt responsible before the course. One participant reflected on their role in reflecting on values and ethical issues. They struggled to differentiate when they were a facilitator, helping clients elucidate their values and identify possible value conflicts, as opposed to imposing their own or their organization's values in design projects. The participant remarked: 'Whose values? Your role is different. I would like to see that separated. I do not have a clear understanding of that, yet. I need to understand the difference.' (participant 2).

Collective awareness and reflection

Reflection collective awareness – While participants observed that ethical aspects were discussed more frequently within their teams since the course, they either disagreed or expressed uncertainty about whether this change was attributable to the course itself or to a broader, government and organization-wide shift in attention to ethical issues in digital technologies (answers to statement 9 (2A|0B|7D), see Table 4). The reflection on ethical issues with the client varied according to the project phase and the client's willingness. The early phases of a project, particularly the intake phase, offered participants the most space to reflect proactively on ethical issues. (participants 5, 6, and 2). However, in cases where a legal framework was already available, discussions about values were crowded out (participant 6).

Communication and shared language – The participants felt most strongly that the course had provided them with a language to articulate values and value conflicts (answers to statements 4 (6A|3B|0D) and 5 (7A|2B|0D), see Table 4). This shared language facilitated a discussion about values and values conflicts with colleagues and commissioners. This language also made it easier for discussions with the client to elucidate underlying values and value conflicts (participant 5). However, one participant remarked, 'Discussing value conflicts is much easier with those who have followed the course rather than with others who have not' (participant 2). One participant appreciated using other course participants as sounding boards to discuss ethical questions they had encountered in their work (participant 11).

Articulating the elephant in the room – The need for a shared language and a framework to reflect on values and value conflicts was illustrated by a participant. The DfV approach, they argued, provided 'a way to elucidate the elephant in the room.' (participant 6, building on participant 12).

Implementation in practice

Subsequently, we analyzed whether individual awareness and collective reflection translated into actionable results. We were interested in whether the course impacted the participants' (daily) practices, mainly whether they applied the DfV approach in design processes.

Procedural knowledge – When asked to reflect on the course, the participants recalled the sessions about the value hierarchy, translation of value into design requirements, and the strategies to deal with value conflicts as the most helpful means to reflect on and reduce complexity (answers to statement 1 (7A|1B|1D), see Table 4). However, participants disagreed as to whether the course provided the necessary procedural knowledge – practical tools to apply in their daily activities (answers to statement 2 (2A|3B|4D), see Table 4). Making ethical design decisions or supporting clients in doing so was not perceived to be part of the participant's 'daily activities' (participant 2), yet 'should a situation arise' (participant 8), they knew where to find the DfV approaches to deal with value conflicts. DfV approaches were then applied as frameworks in the early phases of projects, such as the intake with the client (answers to statement 3 (6A|2B|1D), see Table 4). Participants generally agreed they required more practical experience working with the DfV approaches (answers to statement 12 (6A|0B|3D), see Table 4). One participant disagreed: through the course, they recognized that 'similar issues [were] being discussed in scientific discourse,' which supported them in 'justifying why certain choices are made' (participant 1). Values like user autonomy were frequently discussed in the design process, though not always clearly communicated. However, it did not trigger or enable this participant to make those choices in the first place.

Application in design – In reflecting on whether they applied the DfV approaches in practice, the participants discussed these methods' scope and application area. During the course, the participants were guided by the two examples of AI-based technologies we provided. Initially, they engaged with algorithmic design choices, such as which indicators should decide the EV-charging capacity, or which personal information should trigger proactive public service delivery. Nudged by sessions on (dealing with) value conflicts, the participants increasingly reflected on socio-technical design choices and the broader socio-political context. The design artifact became increasingly broad or non-technical, moving away from AI-based or digital applications.

One participant (participant 1) reflected on their involvement in the design process of an application aimed at proactively informing citizens about subsidies and services they were eligible for, similar to the case we had been using in the educational intervention. While the team considered values such as privacy, transparency, and fairness – particularly in algorithmic design and data migration – there was limited attention to the broader societal context of use. The design was based on the assumption that citizens did not access subsidies and services due to a lack of awareness. However, the participant acknowledged that other factors, such as distrust in government among certain societal groups, might have played a role. Consequently, they reflected, a technological solution might not have addressed the underlying problem effectively. This reflection aligns with the RI ambition to challenge a linear view on technology push and market demand.

Several participants remarked that AI, in particular, provides unprecedented and transformative means to public sector organizations, which may exacerbate value conflicts. However, this was emphasized not to be is not distinctive to AI but broadly emerging digital technologies.

In reflecting on their application in practice, the participants emphasized not applying the DfV approaches to translate values into (technical) design requirements. Rather, they applied the DfV approach to reflect on underlying values and value conflicts early in developing new systems or applications. One participant remarked: 'The design of a system does not begin with a question located in the digital domain [… .] It begins with the question of what is actually required' (participant 12). This opened a broader discussion on the possible design artifacts. One participant, who is a strategic advisor, remarked: 'Design, for me, is also about designing a process. (…) About the position of an AI system in the organization. When I think of design, I don't just think of a piece of software. For me, it [DfV approach] is extremely useful in many other domains besides software development.' (participant 2).

Experienced success factors

In the evaluation, we discussed the success factors necessary to facilitate individual awareness, collective reflection, and the practice of DfV approaches. These success factors concerned didactic aspects of the course itself and organizational conditions to facilitate the uptake of DfV approaches.

Success factors at the course level

We identified didactic aspects necessary for the course's success based on the participants' answers.

Contextualization – The participants argued that further contextualizing the course and the DfV approach would increase their understanding. They identified two aspects relevant to further conceptualization. First, relating the DfV approach to other approaches for responsible innovation. One participant mentioned grounding the DfV approach in 'fundamental ethics' to clarify the underlying assumptions (participant 3). This would have helped them to elucidate underlying assumptions of the DfV approach, and, therefore, provided guidelines as to when this approach is beneficial, requires adaptations, or should be amended or replaced with other approaches. One participant remarked that while a 'technocratic design method,' such as DfV, fit well with the organization and the mindset and background of employees, they would assume that in specific contexts, other, more heuristic approaches facilitated 'neutral or nonviolent communication' (participant 12). Relating those different approaches and comparing the underlying assumptions would have helped to embed these methods in practice. Dutch public sector organizations have been developing frameworks, tools, and guidelines for data ethics, such as Guidance Ethics or a Human Rights Impact Assessment. While the participants appreciated the 'conceptual' (participant 8) and 'high-level reflection' (participant 5) on values and value conflicts, integration with existing initiatives in practice would have been helpful. One participant remarked feeling occasionally overwhelmed by the number of frameworks, tools, and guidelines for data ethics (participant 6).

Selection of participants – The selection of participants was perceived as crucial. Participants' motivation is crucial for a successful collective learning journey. This course was based on self-selection. On the one hand, most participants were motivated to learn about the subject matter and to apply what they had learned in practice. On the other hand, this meant we encountered participants who were already aware and felt responsible for data ethics. When asked whether their awareness had increased (answers to statement 8 (3A|1B|5D), see Table 4) or whether they felt more responsible after the course (answers to statement 6 (3A|2B|4D), see Table 4), most participants disagreed, arguing they had already felt that way. Potential participants who were not yet aware or did not feel responsible were probably not participating in the course. A participant remarked that they had no such responsibility (participant 10). According to the participants, the responsible project owners, managers, and commissioners should participate in the course as they are the actors making and steering ethical design decisions in projects or the organization. These actors were also identified as crucial and sometimes hampering the regard for values and DfV approaches in practice.

Learning-by-doing – The participants appreciated using cases closely related to their work in the course. Providing participants with individual and collective assignments throughout the course was perceived as helpful by the participants.

Expectation management – An additional important success factor was the management of expectations at the beginning of the course, as the participants found the course complex and sometimes dense.

Additional knowledge – A few participants would have found a list with values and their operationalization helpful.

Success factors at the organizational level

Whether or not participants found themselves able to implement the DfV approaches in practice was found to be impacted by several organizational aspects.

Organizational culture – First, the participants emphasized that the organizational culture needed to facilitate value reflections. The participants generally agreed that their intraorganizational organization's culture enabled such reflections. This openness was also illustrated by the initiatives of several participants after the course, such as integrating methods in contracting documents, designing an organization-wide 'value compass,' or setting up a workgroup on AI with a particular focus on ethics. Participants remarked that it was essential to proactively provide the space and time for such reflections (participant 3 & 5). Others perceived there was room and time to address values and value conflicts, though they were unsure of 'the sensitivity to the message' (participant 12).

Organizational commitment – Second, participants emphasized the importance of organizational commitment to the course. As a means of showing commitment, for example, an initial introduction to the course should have been given by the leadership of the organization (participant 6). This would have communicated commitment and ownership to the participants. The organization's leadership participated in the initial course, positively contributing to discussions.

Organizational guidance – Third, participants needed organizational guidance to identify and operationalize values into design requirements. One participant argued, 'What I miss at [our organization] is, what values are you now testing against? What do we stand for as an organization? Providing a direction and the values we stand for as an organization.' (participant 3). Throughout the course, the participants reflected upon this point. There was a shared need for organizational guidance on identifying and operationalizing relevant values. Some participants argued that an inventory and operationalization of organizational values, or value compass, would be helpful. Some participants took the initiative to re-examine and amend an existing value compass that they considered insufficient. Further, organizational guidance was perceived to be required regarding the operationalization of values, as two participants emphasized: 'It is difficult to indicate when something is sufficiently translated into a norm. What is a good norm?' (participants 3 & 5).

Organization as design artifact – Fourth, reflecting on the issues raised above, the participants argued that the organization, its processes, and practices should also become a design artifact to be designed following a DfV approach.

Political-administrative environment – Lastly, the participants critically reflected upon the institutional context in which they were embedded, particularly regarding clients. As a public company providing services to other public sector organizations, they found themselves in a political-administrative environment that was not always conducive to normative discussions. A participant illustrated this point by referring to the strategies to deal with value conflicts (participant 11). They argued, 'Within the government when discussing value conflicts, we quickly arrive at "satisficing." Good is good enough, but who determines that? And then often you end up at that situation, at which money and time have run out.' (participant 11). Another participant (participant 8) added that in a political-administrative environment, preferences and interests could change rather quickly. They added, 'we are generally working in the satisficing modus in a political-administrative environment […]. As a technocrat, it's nice that we can provide a method, facilitate a conversation. Help clients to have a conversation and provide insights. But tomorrow it could be different again. So, I think it's very helpful. But I'm sure that in practice it can be quite challenging, because of administrative rationality and interests. […]. It helps to have an ideal, but it's not necessarily what ultimately happens.' (participant 8).

Discussion

This study employs a system perspective on the responsible design of emerging digital technologies. This means that more is needed than the responsible design of a technological artifact to achieve responsible AI in the public sector. The discussion employs this system perspective by reflecting on the success factors we mentioned in the results section and by identifying the strengths and weaknesses of the DfV approach in this respect.

Success criteria for implementing DfV in a public sector organization

Based on our results, we identify four success factors that seem critical for our intention to have a more durable and lasting effect on the governmental system of AI development: 1) the need for individual responsibility, 2) the need for embedding the DfV approach in organizational procedures and practices, 3) the need for designers to have agency in a political-administrative system, and 4) the need for broader political support for the DfV approach. The second part of the discussion addresses the strengths and weaknesses of DfV as they came forward in the results.

Individual responsibility

The findings underscore the importance of professionals taking individual responsibility as a critical success factor in the responsible design of emerging digital technologies (van de Poel and Sand 2021). Participants in the study highlighted the necessity of cultivating the skills of a reflexive practitioner – a skill that requires continuous practice and adaptation across diverse contexts. This reflects the broader need for professionals to engage with DfV approaches within their spheres of influence actively. While it is easier to implement RI methods with open-minded, enthusiastic innovators (Fisher et al. 2024; Smolka and Fisher 2024), the emphasis on individual responsibility suggests that educational interventions can foster systemic long-

term change. Our findings contrast with research with other professionals, such as scientists. Glerup, Davies, and Horst (2017) find that while scientists working on emerging technologies perform 'bottom-up responsibilities,' they reject what they perceive as 'extra' responsibilities promoted in RI discourse. Rather than 'top-down' efforts at 'responsibilisation' (Glerup, Davies, and Horst 2017), the RRI scholarship needs to be grounded in local language and experience (330) to 'develop shared interpretations and practice of responsibility among RI scholars and scientists' (330). The public sector participants perceived responsible design as their individual and collective responsibility and generally displayed techno-moral virtues. They were willing to take responsibility for techno-moral design choices. However, this must be seen in light of the aforementioned self-selection bias. They also perceive themselves as part of a broader system, be it professional, organizational, political-administrative, or societal, which is perceived as limiting their efforts. The DfV approach, they generally agree, provides them with the language to better understand, communicate, and fulfill these responsibilities in practice, as we will elaborate on later.

Organizational Embedding

Organizational support and organizational culture were found to be essential drivers for the integration of DfV in organizational procedures and individual daily practice. This aligns with the literature on the importance of a systematic perspective on RI implementation (Kroes et al. 2006; Smolka and Böschen 2023; Stahl 2023). The shared organization-wide reflection on values and shared knowledge of DfV in practice appear fundamental. Participants expressed the importance of substantial organizational embedding, for example, by developing an organizational value framework encompassing and legitimizing values in practice and procedural embedding through integrating moral reflections in procedures, standards, templates, and continued training. The participants perceived this organizational embedding primarily as a responsibility of organizational leadership. These findings confirm the findings from other domains, such as Pansera et al. (2020, 402–403) research in biotech companies emphasizing the importance of legitimation, institutional entrepreneurship and leadership, and persistent 'institutional work.'

Agency in a political-administrative system

A recurrent issue during the courses and the evaluations concerned the (perceived) degree of executive power and moral agency of the IT professionals within a political-administrative ecosystem. As previously highlighted, the RI literature challenges a linear perspective of technological development, where innovation is driven by technology push or market demand. Instead, it advocates democratizing the technology development process through anticipation, reflexivity, responsiveness, and inclusiveness. Participants argue that they can feel constrained in the political-administrative context. Rather than democratizing technology development and innovation through an iterative and collaborative design process, in a political-administrative context, a democratic mandate is conveyed through a political directive and the administrative hierarchy. Despite this perceived diminished agency, participants reported feeling enabled by a shared language and structured approach to raising these procedural and substantial norms.

Leadership support and commitment

A fourth critical success factor was top-down support and commitment by organizational leadership. The political sensitivity to the responsible design of AI was repeatedly emphasized by reference to a recent scandal surrounding the Dutch government's use of algorithms. The Dutch childcare benefits scandal involved the Tax and Customs Administration's use of algorithms to falsely accuse thousands of families of fraud, resulting in severe financial hardship and the collective resignation of the Dutch government in early 2021. Considering this prominent scandal, participants emphasized the need for a top-down (legitimized) operationalization of values for the responsible design of AI, such as the value framework NORA (Nederlandse Overheid Referentie Architectuur) developed by the Dutch government. Here, we see a more comprehensive need for political recognition of the critical decisions made, alignment between the value frameworks of clients and organizational values, and individual considerations on the contractor's side. The government must become a partner in the design of digital applications rather than just a client. In this idea of co-creation, in which responsibility is shared, the DfV approach allows the contractor to bring potential issues to the table in a concise (and perhaps technocratic) language while paying tribute to the importance of addressing values of ethical concern.

These four success criteria reflect the participant's desire for value alignment between them personally, the organizational context, the client's requirements, and the broader political-administrative environment. Broader societal values were either perceived as mediated through the political leadership or universal, including broader government value frameworks. While the participants emphasized their responsibility for DfV, ultimately, they understood themselves as 'public servants.' As such, they understood their responsibility as the identification, articulation, and, possibly, mediation of value conflicts. However, in persistent value misalignments, they emphasized the importance of DfV as a means of accountability. By systematically applying the DfV approaches transparently and in collaboration with the client, the participants could articulate and explain design choices, thereby addressing the need for accountability in a broader political-administrative context.

Our findings thus corroborate the need for a system perspective when evaluating and effecting RI approaches. Notably, our second, third, and fourth success factors relate to a broader governmental AI socio-technical system, mainly its political-administrative culture, in which changes must be effectuated. At the same time, our findings suggest that such systematic change may start at the individual and company levels with educational intervention. The fact that our (voluntary) course attracted 'believers' may, in this respect, also be seen as an advantage because these people can start to act as change agents in the broader organization and system.

Strengths and Weaknesses of the DfV approach

Shared language

The DfV approach gave the participants a shared language to communicate and reflect on value issues. In a political-administrative environment, such a language can provide a helpful tool to identify and communicate shared values and value conflicts. In this manner, the DfV approach gave a 'lingua franca' for discussing values and value conflicts in and outside the range of specific projects and project groups. As we argued earlier, this conflicts with what Glerup, Davies, and Horst (2017) phrase as 'top-down' efforts at 'responsibilisation.' Instead, they emphasize that RRI scholarship should be rooted in the local language to foster the application of RRI in practice. In contrast, the public sector participants perceived the DfV as providing a shared language to articulate better and share ethical design considerations.

This is a significant strength since it diminishes idiosyncrasy and allows participants to return to shared grounds through questions that ask for elucidation or the restatement of specific positions and individual valuations. Furthermore, the different ways to deal with value conflicts (maximizing, satisficing, respecification, and innovation) allow for a shared reflection on possible solutions to contextualized problems without necessarily committing beforehand to one solution.

Contextual focus

A second strength of DfV is its inherent focus on context. Contextual awareness of the many specific ethical guidelines developed in the client's and contractor's sphere – in addition to general AI guidelines and principles – is pivotal. While DfV aims to offer a grassroots and bottom-up approach to select values of ethical concern in the design of artifacts and systems, participants were already aware of a wide range of such concerns these artifacts and systems might raise. The question was rather one of navigation and selection than of awareness. Navigational liberties depended on communicating concerns to stakeholders (clients, co-workers, supervisors) and the extent to which these stakeholders were aware of DfV. Indeed, while the participants admitted that a definitive selection of salient values for a design brief was not something they could decide in isolation, the articulation of contextualized concerns to the client was facilitated through the course. Their assessment of colleagues who had not followed the course reflected such an 'epistemic empowerment.'

Technocratic bias

We identified one specific weakness: The DfV approach has its roots in engineering. The technocratic aspect of the approach was perceived as valuable for identifying and communicating shared values and talking about design rather than platonic ideals. However, in a political-administrative environment, this approach also has its limitations. It remains silent on issues of power and procedural justice. These procedural issues do not surface, such as who should be meaningfully included in the design process. This is particularly

problematic from a systems perspective. While educating the DfV approach might be a helpful intervention from a system perspective, as we have seen above, the DfV approach itself does not provide participants with the language or tools to reflect on the need for broader systematic change.

Integrating the DfV approach with other theoretical perspectives, tools, and practices and providing more guidance in procedural matters could, therefore, be beneficial. This weakness relates to a broader threat to the Dutch government. After various scandals, there has been a heightened public sensitivity to the government's use of artificial intelligence and other emerging technologies. The course addresses the participants' perceived needs, the organizational leadership, and the broader Dutch government landscape for more guidance on ethical issues in developing emerging digital technologies, such as AI. Guidelines and checklists are perceived as means to reduce complexity, ambiguity, moral overload, and depoliticization. However, these prescriptive approaches to standardizing values and value operationalization also threaten the individual and collective reflection on ethics and moral agency, with the danger of leaving ethical questions to experts (Hollanek 2025).

Conclusion

By analyzing and reflecting on the participants' experiences, we aimed to answer whether the DfV approach can support IT professionals in translating AI ethical principles into practice. We expected the course first to have an impact on the awareness among the participants about the significance of moral values in the digital domain. Second, to enable participants to integrate DfV approaches in their daily practices to identify and effectively manage conflicting values in the design of digital technologies to deliver public services. Though we address a gap between principles and practice identified in the AI ethics scholarship, our findings extend to emerging digital technologies in the public sector.

We find that teaching DfV methods contributed to cultivating ethical virtues in the participants. However, the DfV methodology needs to pay more attention to the governance of data ethics. More than a one-off or ad-hoc consulting on the design of cases, practitioners and their respective organizations benefit from iterative or ongoing reflection and guidance. While the interaction strengthened the individual responsibilities, the participants perceived a gap in the organizational embedding. The design and development of emerging digital technologies is a complex inter-organizational process. The participants perceived a significant dependence on the goodwill of their clients and questioned their responsibilities as developers of the technology. The DfV approach provides a lingua franca to identify and communicate shared values and value conflicts. In a political-administrative environment, the participants perceived the DfV approach as a means to articulate 'the elephant in the room' (participant 12). Embedding this shared language in the organization can contribute to sustainable integration in practice as the participants translate such abstract ethical principles through shared articulation. However, the fact that the approach is somewhat technocratic was also perceived as a significant limitation of the DfV approach, as it provides little guidance on procedural issues.

Moreover, the DfV approach may not be able to provide the language to articulate broader systematic and political issues. The DfV approach should, therefore, be applied sensitively to its underlying engineering assumptions. Otherwise, instead of DfV through continuous dialogue, collective reflection, and negotiation among various stakeholders, one may risk falling prey to ethics as a 'deontologically inspired tick-box exercise' (Hagendorff 2020, 112), crowding out virtue ethics.

Future research should continue systematically evaluating the impact of DfV and other RI approaches in different contexts. Notably, the enduring embedding of such approaches in various contexts could provide valuable insights into the contextual factors shaping the application and development of such methods. Empirical research should mainly focus on comparative and longitudinal methods.

Note

1. Human Research Ethics Committee TU Delft (Reference 3943).

Acknowledgements

We thank the participants of the course for their collaboration.

Author contributions

CRediT: **Antonia Sattlegger:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing; **Joost Alleblas:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing; **Ibo van de Poel:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Ibo van de Poel's contribution is part of the research programs ValueChange (ERC grant number 788321) and Ethics of Socially Disruptive Technology (NWO grant number 024.004.031). Joost Alleblas is part of the research program ValueChange (ERC grant number 788321).

Notes on contributors

Antonia Sattlegger is a PhD candidate at TU Delft. She works on moral responsibility in the design of artificial intelligence in the public sector. She is affiliated with Digicampus and TU Delft Digital Ethics Centre, where she contributes to professional training on Digital Ethics by Design.

Joost Alleblas is a PhD candidate at TU Delft. He investigates how visions shape transitions, especially energy transitions. He is part of an ERC-funded project that focuses on changing values in society and the implications for the design of socio-technical systems. He has participated in multiple workshops for private and public organizations that used the methodology of Design for Values to bring into focus the ethical issues surrounding emerging technologies.

Ibo van de Poel is Professor in Ethics and Technology at TU Delft. His research focuses on value change, ethics of disruptive technologies, ethics of technological risks, design for values, responsible innovation, and moral responsibility.

ORCID

Antonia Sattlegger http://orcid.org/0009-0000-1154-655X
Joost Alleblas http://orcid.org/0000-0002-2420-7597
Ibo van de Poel http://orcid.org/0000-0002-9553-5651

References

Amnesty International. 2021. Xenophobic Machines Discrimination through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal. https://www.amnesty.nl/content/uploads/2021/10/20211014_FINAL_Xenophobic-Machines.pdf?x25503.

Bardone, E., M. Burget, and M. Pedaste. 2023. "The RRI map: Making Sense of Responsible Research and Innovation in Science Education." *Journal of Responsible Innovation* 10 (1): 1. https://doi.org/10.1080/23299460.2023.2198183.

Bharosa, N., B. Oude Luttighuis, F. Spoelstra, H. van der Voort, and M. Janssen. 2021. Inclusion Through Proactive Public Services: Findings from the Netherlands. *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, 242–251. https://doi.org/10.1145/3463677.3463707.

Cawthorne, D. 2023. *The Ethics of Drone Design: How Value-Sensitive Design Can Create Better Technologies.* 1st ed. New York: Routledge. https://doi.org/10.4324/9781003372721.

Crosby, B. C., P. 't Hart, and J. Torfing. 2017. "Public Value Creation through Collaborative Innovation." *Public Management Review* 19 (5): 655–669. https://doi.org/10.1080/14719037.2016.1192165.

European Commission. 2019. *Rome Declaration on Responsible Research and Innovation in Europe*. https://digital-strategy.ec.europa.eu/en/library/rome-declaration-responsible-research-and-innovation-europe.

Fisher, E., M. Smolka, R. Owen, M. Pansera, D. H. Guston, A. Grunwald, J. P. Nelson, et al. 2024. "Responsible Innovation Scholarship: Normative, Empirical, Theoretical, and Engaged." *Journal of Responsible Innovation* 11 (1): 1. https://doi.org/10.1080/23299460.2024.2309060.

Friedman, B., and D. G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001.

Giest, S. N., and B. Klievink. 2024. "More Than a Digital System: How AI is Changing the Role of Bureaucrats in Different Organizational Contexts." *Public Management Review* 26 (2): 379–398. https://doi.org/10.1080/14719037.2022.2095001.

Glerup, C., S. R. Davies, and M. Horst. 2017. "'Nothing Really Responsible Goes on Here': Scientists' Experience and Practice of Responsibility." *Journal of Responsible Innovation* 4 (3): 319–336. https://doi.org/10.1080/23299460.2017.1378462.

Grimmelikhuijsen, S., and A. Meijer. 2022. "Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response." *Perspectives on Public Management and Governance* 5 (3): 232–242. https://doi.org/10.1093/ppmgov/gvac008.

Guston, D. H., and D. Sarewitz. 2002. "Real-time Technology Assessment." *Technology in Society* 24 (1-2): 93–109. https://doi.org/10.1016/S0160-791X(01)00047-1.

Hagendorff, T. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8.

Herzog, C., and S. Blank. 2024. "A Systemic Perspective on Bridging the Principles-to-Practice gap in Creating Ethical Artificial Intelligence Solutions – a Critique of Dominant Narratives and Proposal for a Collaborative way Forward." *Journal of Responsible Innovation* 11 (1): 1. https://doi.org/10.1080/23299460.2024.2431350.

Hollanek, T. 2025. "The Ethico-Politics of Design Toolkits: Responsible AI Tools, from Big Tech Guidelines to Feminist Ideation Cards." *AI and Ethics* 5: 2165–2174. https://doi.org/10.1007/s43681-024-00545-z.

James, A., and A. Whelan. 2022. ""Ethical" Artificial Intelligence in the Welfare State: Discourse and Discrepancy in Australian Social Services." *Critical Social Policy* 42 (1): 22–42. https://doi.org/10.1177/0261018320985463.

Jobin, A., M. Ienca, and E. Vayena. 2019. "Artificial Intelligence: The Global Landscape of Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kroes, P., M. Franssen, I. van de Poel, and M. Ottens. 2006. "Treating Socio-Technical Systems as Engineering Systems: Some Conceptual Problems." *Systems Research and Behavioral Science* 23 (6): 803–814. https://doi.org/10.1002/sres.703.

Kudina, O., and I. van de Poel. 2024. "A Sociotechnical System Perspective on AI." *Minds and Machines* 34 (3): 21. https://doi.org/10.1007/s11023-024-09680-2.

Künneke, R., D. C. Mehos, R. Hillerbrand, and K. Hemmes. 2015. "Understanding Values Embedded in Offshore Wind Energy Systems: Toward a Purposeful Institutional and Technological Design." *Environmental Science & Policy* 53:118–129. https://doi.org/10.1016/j.envsci.2015.06.013.

Madan, R., and M. Ashok. 2023. "AI Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda." *Government Information Quarterly* 40 (1): 740–624. https://doi.org/10.1016/j.giq.2022.101774.

Maleeyakul, N., C. Houtekamer, M. Rengers, G. Geiger, K. van Dijken, D. Howden, C. Black, and A. Papagapitos. 2024. *Ethnic Profiling*. Lighthouse Reports. https://www.lighthousereports.com/investigation/ethnic-profiling/.

Marschalek, I., M. Schrammel, E. Unterfrauner, and M. Hofer. 2017. "Interactive Reflection Trainings on RRI for Multiple Stakeholder Groups." *Journal of Responsible Innovation* 4 (2): 295–311. https://doi.org/10.1080/23299460.2017.1326262.

McNamara, A., J. Smith, and E. Murphy-Hill. 2018. Does ACM's Code of Ethics Change Ethical Decision Making in Software Develop-ment. *Proceedings of the 26th ACM Joint European Software EngineeringConference and Symposium on the Foundations of Software Engineering(ESEC/FSE' 18), November 4–9, 2018, Lake Buena Vista, FL, USA.ACM,New York, NY, USA*. https://doi.org/10.1145/3236024.3264833.

Meijer, A., and S. Grimmelikhuijsen. 2020. "Responsible and Accountable Algorithmization: How to Generate Citizen Trust in Governmental Usage of Algorithms." In *The Algorithmic Society: Technology, Power, and Knowledge*, edited by M. Schuilenburg and R. Peeters, 53–66. London: Routledge. https://doi.org/10.4324/9780429261404.

Mittelstadt, B. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1 (11): 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Naughton, B., S. Dopson, and T. Iakovleva. 2023. "Responsible Impact and the Reinforcement of Responsible Innovation in the Public Sector Ecosystem: Cases of Digital Health Innovation." *Journal of Responsible Innovation* 10 (1): 1. https://doi.org/10.1080/23299460.2023.2211870.

Neudert, P., M. Smolka, and S. Böschen. 2024. "Towards Transformative Innovation Ecosystems: A Systemic Approach to Responsible Innovation." *Journal of Responsible Innovation* 11 (1): 1. https://doi.org/10.1080/23299460.2024.2414482.

Oosterlaken, I. 2014. "Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An Exploration." *Science and Engineering Ethics* 21 (2): 359–379. https://doi.org/10.1007/s11948-014-9536-x.

Owen, R., M. Pansera, P. Macnaghten, and S. Randles. 2021. "Organisational Institutionalisation of Responsible Innovation." *Research Policy* 50 (1): 1. https://doi.org/10.1016/j.respol.2020.104132.

Pansera, M., R. Owen, D. Meacham, and V. Kuh. 2020. *Embedding Responsible Innovation within Synthetic Biology Research and Innovation: Insights from a UK Multi-disciplinary Research Centre*. https://doi.org/10.1080/23299460.2020.1785678.

Parada, M. P. 2020. *Biorefinery Design in Context: Integrating Stakeholder Considerations in the Design of Biorefineries*. https://doi.org/10.4233/UUID:72E0C6E1-9C17-4B8C-B5CE-FB6A8E2ABF20.

Poulsen, A. 2022. *The Investigation of a New Care Robot Design Approach for Alleviating LGBT+ Elderly Loneliness*. https://researchoutput.csu.edu.au/en/publications/the-investigation-of-a-new-care-robot-design-approach-for-allevia.

Richter, J., A. E. Hale, and L. M. Archambault. 2019. "Responsible Innovation and Education: Integrating Values and Technology in the Classroom." *Journal of Responsible Innovation* 6 (1): 98–103. https://doi.org/10.1080/23299460.2018.1510713.

Sapraz, M., and S. Han. 2021. "Improving Collaboration between Government and Citizens for Environmental Issues: Lessons Learned from a Case in Sri Lanka." In *ICT Systems and Sustainability*, edited by M. Tuba, S. Akashe, and A. Joshi, 343–353. Advances in Intelligent Systems and Computing, vol. 1270. Singapore: Springer. https://doi.org/10.1007/978-981-15-8289-9_28.

Sattlegger, A., J. van den Hoven, and N. Bharosa. 2022. Designing for Responsibility. *DG.O 2022: The 23rd Annual International Conference on Digital Government Research*, *12*, 214–225. https://doi.org/10.1145/3543434.3543581.

Scholta, H., and I. Lindgren. 2023. "Proactivity in Digital Public Services: A Conceptual Analysis." *Government Information Quarterly* 40 (3): 3. https://doi.org/10.1016/j.giq.2023.101832.

Schot, J., and A. Rip. 1997. "The Past and Future of Constructive Technology Assessment." *Technological Forecasting and Social Change* 54 (2-3): 251–268. https://doi.org/10.1016/S0040-1625(96)00180-1.

Selten, F., and A. Meijer. 2021. "Managing Algorithms for Public Value." *International Journal of Public Administration in the Digital Age* 8 (1): 1. https://doi.org/10.4018/IJPADA.20210101.oa9.

Smolka, M., and E. Fisher. 2024. "Testing Reflexive Practitioner Dialogues: Capacities for Socio-Technical Integration in Meditation Research." *NanoEthics* 18 (1): 1. https://doi.org/10.1007/s11569-023-00450-5.

Smolka, M., and S. Böschen. 2023. "Responsible Innovation Ecosystem Governance: Socio-Technical Integration Research for Systems-Level Capacity Building." *Journal of Responsible Innovation* 10 (1): 1. https://doi.org/10.1080/23299460.2023.2207937.

Spruit, S. 2014. "Responsible Innovation through Ethics Education: Educating to Change Research Practice." *Journal of Responsible Innovation* 1 (2): 246–247. https://doi.org/10.1080/23299460.2014.922344.

Stahl, B. C. 2021. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies.* Cham: Springer. https://doi.org/10.1007/978-3-030-69978-9.

Stahl, B. C. 2023. "Embedding Responsibility in Intelligent Systems: From AI Ethics to Responsible AI Ecosystems." *Scientific Reports* 13 (1): 7586. https://doi.org/10.1038/s41598-023-34622-w.

Stahl, B. C., C. Aicardi, L. Brooks, P. J. Craigon, M. Cunden, S. D. Burton, M. De Heaver, et al. 2023. "Assessing Responsible Innovation Training." *Journal of Responsible Technology* 16:100063. https://doi.org/10.1016/j.jrt.2023.100063.

Steen, M. 2021. "Slow Innovation: The Need for Reflexivity in Responsible Innovation (RI)." *Journal of Responsible Innovation* 8 (2): 254–260. https://doi.org/10.1080/23299460.2021.1904346.

Steen, M., M. Sand, and I. van de Poel. 2021. "Virtue Ethics for Responsible Innovation." *Business and Professional Ethics Journal* 40 (2): 243–268. https://doi.org/10.5840/bpej2021319108.

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008.

Tomblin, D., and N. Mogul. 2020. "STS Postures: Responsible Innovation and Research in Undergraduate STEM Education." *Journal of Responsible Innovation* 7 (sup1): 117–127. https://doi.org/10.1080/23299460.2020.1839230.

Umbrello, S., and I. van de Poel. 2021. "Mapping Value Sensitive Design onto AI for Social Good Principles." *AI and Ethics* 1 (3): 283–296. https://doi.org/10.1007/s43681-021-00038-3.

van de Poel, I. 2013. "Translating Values into Design Requirements." In *Philosophy and Engineering: Reflections on Practice, Principles and Process (Philosophy of Engineering and Technology)*, edited by D. Michelfelder, N. McCarthy, and D. Goldberg, Vol. 15, 253–266. Dordrecht: Springer.

van de Poel, I., and M. Sand. 2021. "Varieties of Responsibility: Two Problems of Responsible Innovation." *Synthese* 198 (S19): 4769–4787. https://doi.org/10.1007/s11229-018-01951-7.

Veale, M., M. Van Kleek, and R. Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Conference on Human Factors in Computing Systems - Proceedings, 2018-April*. https://doi.org/10.1145/3173574.3174014.

Werker, C. 2020. "Assessing Responsible Research and Innovation (RRI) Systems in the Digital Age." In *Assessment of Responsible Innovation in the Digital Age*, edited by E. Yaghmaei and I. van de Poel, 275–292. London: Routledge. https://doi.org/10.4324/9780429298998.

Wirtz, B. W., J. C. Weyerer, and B. J. Sturm. 2020. "The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration." *International Journal of Public Administration* 43 (9): 818–829. https://doi.org/10.1080/01900692.2020.1749851.