DOCTORAL CONSORTIUM

Developing Public Values Based AI Systems Using Value Sensitive Design

Erik Slingerland,[1] Guido Ongena,[1]
Marlies van Steenbergen[2]

[1] HU University of Applied Sciences, Research Group Process Innovation & Information Systems, Utrecht, Netherland
erik.slingerland@hu.nl, guido.ongena@hu.nl

[2] HU University of Applied Sciences, Research Group Digital Ethics, Utrecht, Netherland
marlies.vansteenbergen@hu.nl

The growing prevalence of AI systems in society, has also prompted a growth of AI systems in the public sector. There are however ethical concerns over the impact of AI on society and how this technology can impact public values. Previous works do not connect public values and the development of AI. To address this, a method is required to ensure that developers and public servants can signal possible ethical implications of an AI system and are assisted in creating systems that adhere to public values. Using the Research pathway model and Value Sensitive Design, we will develop a toolbox to assist in these challenges and gain insight into how public values can be embedded throughout the development of AI systems.

Keywords: public values, value sensitive design, digital ethics, artificial intelligence

University of Maribor Press

DOI https://doi.org/10.18690/um.fov.4.2024.50
ISBN 978-961-286-871-0

1 Introduction

Within the public sector, the growing reliance on digitalization has prompted the rise of e-government. A domain of research within public administration focussed on utilizing digital applications in various aspects of the public domain (Bannister & Connolly, 2014). The implementation of these digitalisation applications is not always successful and can have serious ethical implications. An example is the reveal of the NSA surveillance activities which sparked a global debate surrounding the balance between the values of privacy and (inter)national security (MacAskill et al., 2013). More recently in the Netherlands, the child benefit scandal surfaced, where the illegitimate use of personal information, led to parents incorrectly being classified as fraudulent by algorithms (Autoriteit Persoonsgegevens, 2020). The increasing potential of Artificial intelligence (AI) systems has prompted public servants to utilize this technology in the public domain but, as the example demonstrates, not always with positive outcomes for citizens.

In the proposed AI act, the European Parliament defined several requirements for the use of AI in Europe. These also include public values like equality, sustainability, and transparency (European Parliament , 2023). The demarcation of what constitutes a public value and how these values relate is ambiguous, for example in the relation and distinction between transparency and openness (Meijer, 2013; Whittlestone et al., 2019). Various researchers state that the development and use of technology contain underlying values. Technology is increasingly viewed as a socio-technical system, which focuses on the reciprocal interaction between humans and technology. (Bannister & Connolly, 2014; Flanagan et al., 2008). Achieving values like fairness in these socio-technical systems, is only possible when examining both the social and technical aspects of a system (Selbst et al., 2019).

Currently, there is a gap between the implementation of AI in the public domain and research into public values. To contribute to the implementation of AI systems that adhere to public values, this research aims to answer the following question:

How can public values be implemented and validated in Artificial Intelligence systems in the public domain?

This research question will be inquired from two perspectives. Firstly, the process of identifying and operationalizing public values for AI systems and secondly developing tools to implement these public values in AI systems or validate their presence.

2 Related Work

The concept of embodied values states that a digital application derives its ethical value from a combination of its designed properties and its usage (Flanagan et al., 2008; van de Poel, 2020). This is related is based on two ethical concepts. Firstly, normative ethics aims to judge morality and formulate recommendations about how to act or live. Secondly, value theory states that we can make evaluations of technical artefacts based on ethical values. These values are lasting convictions or matters that people feel 'ought to be.' (van de Poel & Royakkers, 2011). By using this as the basis of an ethical framework, a digital application can be examined on how it contributes to or disrupts the presence of a specific normative value. In the coming section, the concept of public values and a method for designing AI systems to adhere to values are explored.

Within this research, an AI system is defined according to Article 3 of the proposed AI act. This definition is useable within the context of e-government as it has political support and international recognition. 'An AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments' (Artificial Intelligence Act, 2024).

2.1 Public Values

Within the field of public administration, there has been a shift towards policy based on public values (Molina & McKeown, 2012). The previous economic approach to decision-making was criticised, as it did not account for the broader societal impact of policy. This slowly evolved into the concept of public interest and prompted policymakers to consider public values (Bozeman, 2002). A commonly cited definition of public values is: "…values providing normative consensus about (a) the rights, benefits, and prerogatives to which citizens should (and should not) be entitled; (b) the obligations of citizens to society, the state, and one another; and (c) the principles on which governments and policies should be based" (Fukumoto & Bozeman, 2018). This paper looks at public values based on the third aspect of this definition.

These public values also apply when we look at IT innovations in the public domain. Socio-technical systems have the potential influence and be influenced by values (Bannister & Connolly, 2014). Researchers therefore call upon the public sector to recognise that technology is not neutral and has underlying values in its usage. This is reflected in European regulations like the European Data Protection Act and AI Act, which add various ethical obligations (Royakkers et al., 2018). Some researchers also reject the notion that there is a one-dimensional set of public values that can be defined. Values can overlap, have different meanings depending on the context and derive their importance from the social context. A practical approach to deal with these conflicts and overlap is to define concrete and measurable conceptualisations of public values and make a context-dependent decision on which values to include in a system (Wal & Van Hout, 2009).

2.2 Value Sensitive Design

Value Sensitive Design (VSD) was developed as a theoretical approach for designing technology that accounts for human values in a principled and comprehensive manner through the design process. By investigating a design question from conceptual, empirical, and technical perspectives with various techniques, the developer can establish ethical requirements for an artefact and develop a plan on how to achieve them (Friedman et al., 2013). VSD contains various techniques like the stakeholder analysis and value source analysis that can be used in the conceptual and empirical investigation to establish stakeholders values and use them in the design of an artifact (Friedman et al., 2017). Applying VSD in the design of AI systems prompts unique challenges. For example, the capability of some AI systems to adjust their behaviour over time, can cause them to disembody a value for which it was designed (Tsamados et al., 2022; Umbrello & Yampolskiy, 2022). To account for this, a few design methods have been proposed (Sadek et al., 2023). An example is Umbrello & van de Poel who expand the scope of VSD to the entire lifecycle of an AI system. This method maps the investigations of VSD into four activities:

context analysis, value identification, design requirements and prototyping. These activities are a cyclical process that go through multiple iterations, as illustrated in Figure 1. By utilising these steps throughout the lifecycle of an AI system, the value-sensitive design process for AI technologies (VSD for AI) allows users to determine whether the system still adheres to normative values through its deployment and if necessary adjust the system (Umbrello & van de Poel, 2021).

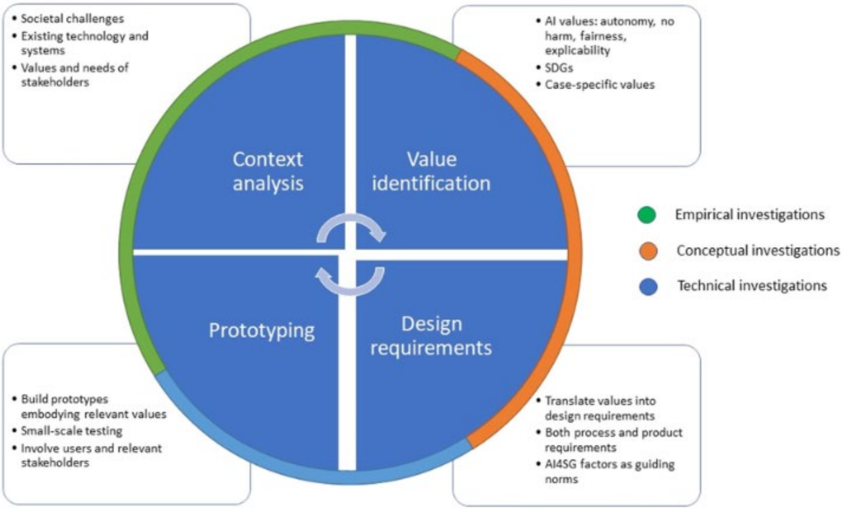
• Societal challenges
• Existing technology and systems
• Values and needs of stakeholders
• AI values: autonomy, no harm, fairness, explicability
• SDGs
• Case-specific values
Context analysis
Value identification
Prototyping
Design requirements
Empirical investigations
Conceptual investigations
Technical investigations
• Build prototypes embodying relevant values
• Small-scale testing
• Involve users and relevant stakeholders
• Translate values into design requirements
• Both process and product requirements
• AI4SG factors as guiding norms

Figure 1: Values sensitive design process for AI technologies
Source: adapted from Umbrello & van de Poel, 2021

There is a gap in linking public values and VSD. There are various papers related to the development of AI systems that adhere to values like fairness and transparency, but there is little research with a focus on public values. Approaching public values as context-dependent phenomenon, allows for VSD to operationalise these public values with techniques like stakeholder and value source analysis. This could create a practical approach to the development of AI systems that adhere to public values.

3 Methodology

To structure the research, the research pathway model is used. In this model, the trajectory is positioned in theoretical, conceptual and practical contexts. This ensures both scientific rigour and practical relevance. In these contexts; creation, exploration and delivery activities are employed (van Beest et al., 2021).

Public values are approached as normative values that systems can be tested against. To guide developers and public servants through the actions in VSD for AI, a toolbox will be developed. The toolbox will also include an instrument to measure the degree to which an AI system embodies various public values. This will allow developers and public servants to examine the AI systems periodically and signal whether the system still embodies the intended values. The toolbox will consist of three main components:

1. A method for mapping relevant public values during the design of an AI system
2. A library of code chunks and design patterns to assist during development.
3. An instrument for testing and evaluating an AI system on public values.

In the following section, the structure from the Research Pathway Model will be used to examine the development of the toolbox. This is also visualised in Figure 2.

Creation phase

As the problem has been identified, the project starts by investigating current state-of-the-art knowledge and assessing the needs of stakeholders. For this task, five activities have been identified. To gain insight into the theoretical context surrounding public values, a literature review on public values is conducted. This will be used as input for a Delphi study. This Delphi study with domain experts is used to create an initial prioritization of public values to include in the theoretical framework. Based on this prioritization, the values will be conceptualized so norms and measures can be identified for each value. In the conceptual context, the prototype of the toolbox will be developed with a focus group using techniques and principles from VSD as inspiration. Lastly in the practice context, interviews will be conducted with AI developers and public servants in the public domain to gain a deeper understanding of the context in which the toolbox will be deployed.

Exploration phase

The exploration phase consists of an iterative process with three main activities. The phase starts by using the input from the creation phase to form the prototype of the toolbox. Secondly, the prototype is tested as an experiment with a test and control group (Mettler et al., 2014). This experiment will be evaluated on two main measures. Firstly, the participants will be interviewed to establish their awareness of the ethical implications before and after using the toolbox. Secondly, the final AI system will be examined by a focus group of ethics and AI experts to examine whether the developed AI system embodies the values that were defined at the start of the project. By doing this for the test group who utilized the toolbox and the control group who did not, it is possible to establish the validity of the toolbox and examine whether participants have an increased ethical awareness surrounding AI systems. Lastly, the framework of public values is reevaluated and redefined where necessary. The new framework iteration and the outcome of the experiment are used to redesign the toolbox for a new iteration. This process is repeated until a final version is reached. The protocol for these experiments and evaluations is being developed.

Delivery phase

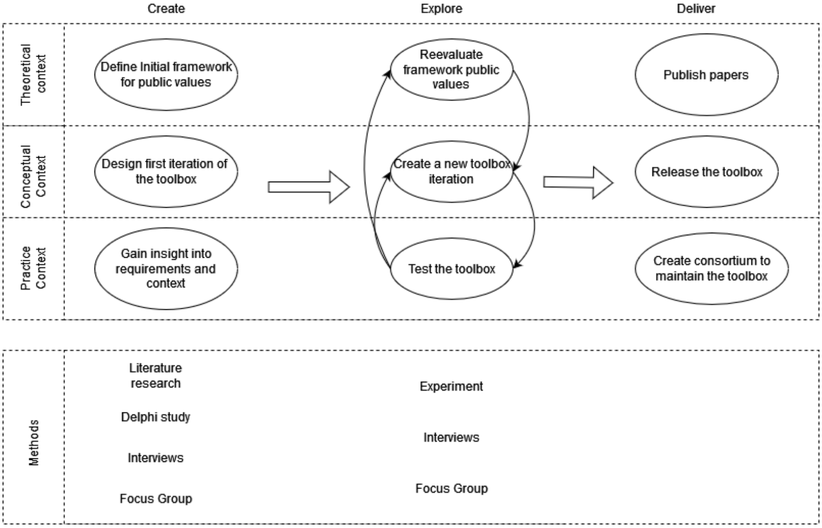
Create
Explore
Deliver
Theoretical context
Conceptual Context
Practice Context
Define Initial framework for public values
Design first iteration of the toolbox
Gain insight into requirements and context
Reevaluate framework public values
Create a new toolbox iteration
Test the toolbox
Publish papers
Release the toolbox
Create consortium to maintain the toolbox
Methods
Literature research
Delphi study
Interviews
Focus Group
Experiment
Interviews
Focus Group

Figure 2: Visualisation of the methodology based on the research pathway model with the methods used.

The delivery phase is concerned with communicating the research results to the public. Three main activities are defined. Firstly, the framework of public values in AI development will be published as a paper to be used in further research. Secondly, the toolbox will be released as an open-source application and be accompanied by a paper detailing the development process. Lastly, to ensure that the toolbox will be properly maintained, a consortium of partners will be realized from the actors involved during the research. This consortium will be tasked with maintaining the toolbox and organizing workshops to instruct new parties in how to use the toolbox once it has been released.

4 Future development and next steps

The first step is to create an initial framework for public values. Currently, the literature review is completed, and a Delphi study is being prepared. In a Delphi Study, domain experts are asked about their opinion anonymously about an issue. This is done by seeking consensus between the experts over a series of rounds (von der Gracht, 2012). Our Delphi study will be conducted with experts in ethics, digital ethics, AI and e-government. It consists of two phases. In the first phase, the participants receive a list of public values distilled from the literature review. The participants are tasked with eliminating overlapping values. Here consensus will be defined per public value where the majority must agree on its inclusion in at least two consecutive rounds. In the second phase, the participants are tasked with ranking the remaining values on importance. This will be done using a tournament ranking system, based on the Q methodology (Brown, 1996). Each participant sees sets of two public values. For every set, the participant specifies which value is more important. Here consensus is reached when after two consecutive rounds, the ranking does not shift. The resulting list of public values is used to create the initial framework. To operationalize these values, they need to be conceptualized. This involves specifying the values to a concrete norm and defining requirements that it can be tested against (Veluwenkamp & Van Den Hoven, 2023). To define the norms and requirements, a combination of literature, (inter)national laws, interviews with domain experts and industry standards will be used. An example would be taking the value of privacy, using a norm from the European Data Protection Act and linking these to an ISO requirement. This will result in a structured approach to operationalizing public values into measurable requirements that AI systems can be evaluated against.

References

Autoriteit Persoonsgegevens. (2020). De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag [Report]. Autoriteit Persoonsgegevens. https://www.rijksoverheid.nl/documenten/rapporten/2020/07/17/de-verwerking-van-de-nationaliteit-van-aanvragers-van-kinderopvangtoeslag

Bannister, F., & Connolly, R. (2014). ICT, Public Values and Transformative Government: A Framework and Programme for Research. Government Information Quarterly, 31. https://doi.org/10.1016/j.giq.2013.06.002

Bozeman, B. (2002). Public-Value Failure: When Efficient Markets May Not Do. Public Administration Review, 62, 145–161. https://doi.org/10.1111/0033-3352.00165

Brown, S. R. (1996). Q Methodology and Qualitative Research. Qualitative Health Research - QUAL HEALTH RES, 6, 561–567. https://doi.org/10.1177/104973239600600408

European Parliament. (2023, June 8). EU AI Act: First regulation on artificial intelligence. EU AI Act: First Regulation on Artificial Intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. Van Den Hoven & J. Weckert (Eds.), Information Technology and Moral Philosophy (1st ed., pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. Foundations and Trends® in Human–Computer Interaction, 11(2), 63–125. https://doi.org/10.1561/1100000015

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), Early engagement and new technologies: Opening up the laboratory (pp. 55–95). Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4

Fukumoto, E., & Bozeman, B. (2018). Public Values Theory: What Is Missing? The American Review of Public Administration, 49, 027507401881424. https://doi.org/10.1177/0275074018814244

MacAskill, E., Dance, G., Cage, F., Chen, G., & Popovich, N. (2013, November 1). NSA files decoded: Edward Snowden's surveillance revelations explained. The Guardian. http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded

Meijer, A. (2013). Understanding the Complex Dynamics of Transparency. Public Administration Review, 73(3), 429–439. https://doi.org/10.1111/puar.12032

Mettler, T., Eurich, M., & Winter, R. (2014). On the Use of Experiments in Design Science Research: A Proposition of an Evaluation Framework. Communications of the Association for Information Systems, 34, 223–240. https://doi.org/10.17705/1CAIS.03410

Molina, A., & McKeown, C. (2012). The Heart of the Profession: Understanding Public Service Values. Journal of Public Affairs Education, 18, 375–396. https://doi.org/10.2307/23208659

Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Council of the European Union (2024). https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Royakkers, L., Timmer, J., Kool, L., & Van Est, R. (2018). Societal and ethical issues of digitization. Ethics and Information Technology, 20(2), 127–142. https://doi.org/10.1007/s10676-018-9452-x

Sadek, M., Calvo, R. A., & Mougenot, C. (2023). Designing value-sensitive AI: A critical review and recommendations for socio-technical design processes. AI and Ethics. https://doi.org/10.1007/s43681-023-00373-7

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. AI & SOCIETY, 37. https://doi.org/10.1007/s00146-021-01154-8

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1, 3. https://doi.org/10.1007/s43681-021-00038-3

Umbrello, S., & Yampolskiy, R. V. (2022). Designing AI for Explainability and Verifiability: A Value Sensitive Design Approach to Avoid Artificial Stupidity in Autonomous Vehicles. International Journal of Social Robotics, 14(2), 313–322. https://doi.org/10.1007/s12369-021-00790-w

van Beest, W., Boon, W., Andriessen, D., Pol, H., van der Veen, G., & Moors, E. (2021). A Research Pathway Model for evaluating the implementation of practice-based research: The case of self-management health innovations. Research Evaluation, 31. https://doi.org/10.1093/reseval/rvab023

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. Minds and Machines, 30(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

van de Poel, I., & Royakkers, L. M. M. (2011). Ethics, technology, and engineering: An introduction. Wiley-Blackwell.

Veluwenkamp, H., & Van Den Hoven, J. (2023). Design for values and conceptual engineering. Ethics and Information Technology, 25(1), 2. https://doi.org/10.1007/s10676-022-09675-6

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. Technological Forecasting and Social Change, 79(8), 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013

Wal, Z., & Van Hout, E. (2009). Is Public Value Pluralism Paramount? The Intrinsic Multiplicity and Hybridity of Public Values. Intl Journal of Public Administration, 32, 220–231. https://doi.org/10.1080/01900690902732681

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. 195–200. https://doi.org/10.1145/3306618.3314289