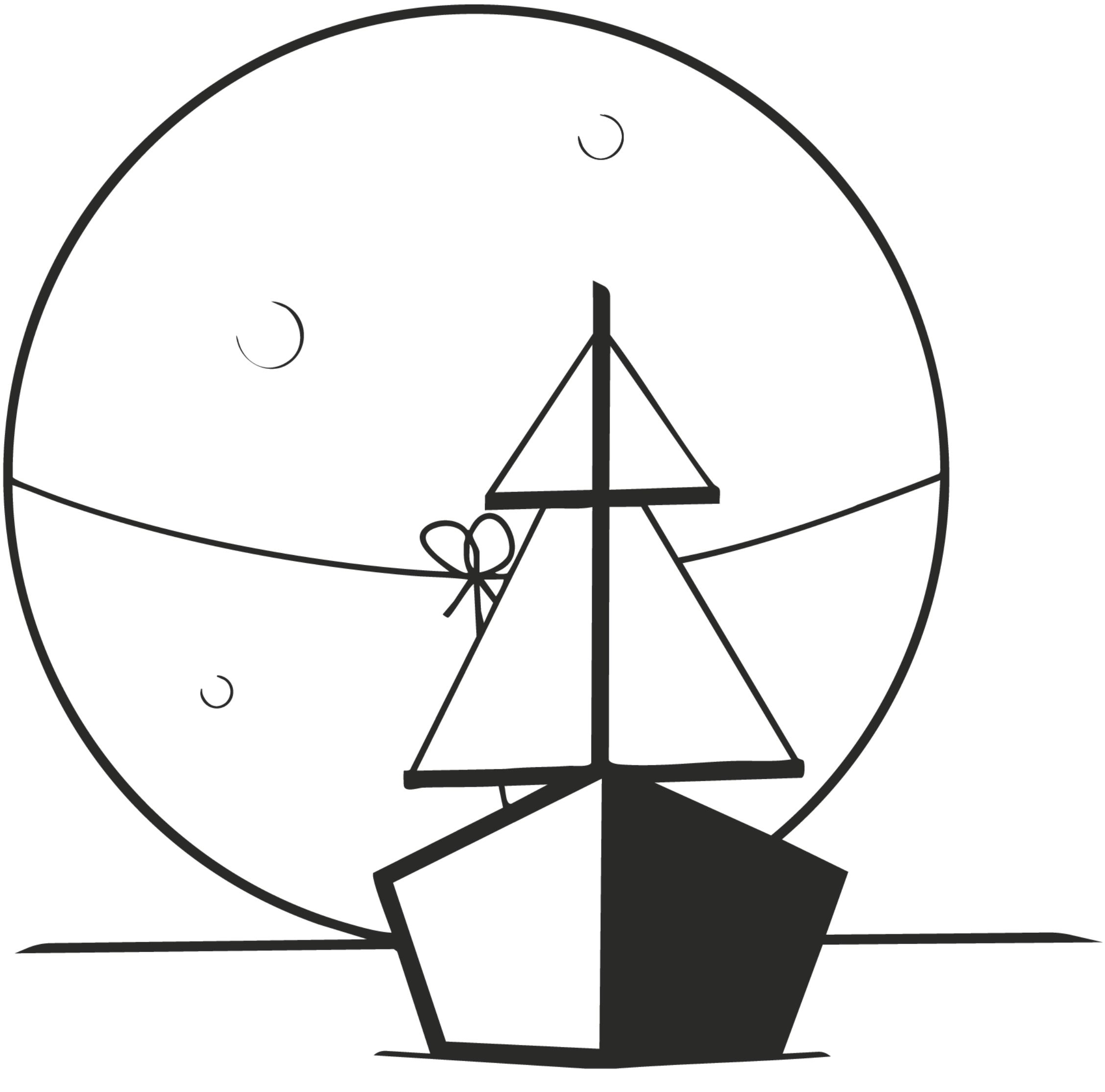
Navigating value tensions in the use of AI for policy preparation

Towards guidelines & a practical tool



David Mieras

Cover illustration

Let curiosity inspire reflection, casting light upon that which lies in the dark, pulling us towards responsibility.

Title

Navigating value tensions in the use of AI for policy preparation;

towards guidelines & a practical tool.

Master Thesis

MSc Strategic Product Design

Faculty of Industrial Design Engineering

Delft University of Technology

Year

2025

Author

David Lodewijk Mieras

Graduation Committee

Chair

Dr. ir. Lianne Simonse

Faculty of Industrial Design Engineering

Mentor

Dr. Kars Alfrink

Faculty of Industrial Design Engineering

Company mentor

[Name redacted]

[Organisation redacted]

Preface

Dear reader,

Before is send you off to read this thesis, I would like to express my gratitude to the people who have helped me so much in bringing forth this thesis.

Firstly, I would like to thank my supervisors Lianne, Kars and [Name redacted], and my additional informal company mentor [Name redacted] for their time, helpful feedback and support throughout the project.

Secondly, [redacted]

Thirdly, I would like to thank all the other inspiring people I have had the pleasure of meeting through this project. As well as those who inspired me to keep going.

Lastly, I want to thank my parents for pointing me in the right direction.

Best,

David

Executive summary

This thesis explores how the Dutch government can adopt artificial intelligence (AI) responsibly in the policy development process, a domain that has received little attention compared to AI use in policy execution. The project was conducted in collaboration with the governmental organisation.

The thesis identifies that while AI holds promise for improving policy quality, efficiency, and democratic engagement, it also introduces serious risks, such as depoliticisation, bias, loss of professional judgment, and declines in public trust. These risks, combined with organisational barriers like low AI literacy, limited capacity, and fragmented structures, have led to hesitant adoption within ministries.

The thesis uses a constructive design research method. Answering research question by means of design. With a design project that uses an design approach based Frame Innovation, Vision in Product Design (ViP), and Value Sensitive Design (VSD). Resulting in a prototype tool that is evaluated with civil servants. The design balances encouragement and responsibility, aiming to stimulate AI curiosity, proposed as a key mechanism for learning and soft AI capacity-building, while reinforcing awareness of ethical and procedural boundaries.

Findings show that responsible AI adoption depends not only on technical safeguards but also on developing collective AI capacity, trust, and professional judgment. The tool incites reflection rather than prescription, helping users think critically, recognise dilemmas, and connect to existing support resources, which anchors quality assurance in the Dutch policy process.

This thesis contributes to bridging the gap between theoretical frameworks of responsible AI and practical application in policy preparation.

Reading Guide

Below is an explanation of some important terms and their meaning in this report.

Policy officers [Beleidsmedewerkers] = Civil servants responsible for researching, drafting, and providing recommendations on public policy.

Policymakers [Beleidsmakers] = Political- and senior civil leadership who make policy decisions.

Polity [Staatsbestel] = A system of government.

Institution = Depending on the context, the term refers to formal organisations, as understood in common parlance, or to institutions as understood in institutional theory, meaning the formal and informal 'rules of the policy game' (Hill & Varone, 2021),

Declaration of the use of AI

Because of the nature of the topic of the thesis, I set out to use AI conservatively, taking special notice of the secure treatment of sensitive data, and limiting the influence of AI on the ideas written in this thesis. Below is a list of ways AI was used in this thesis project:

Wispher (local) - The recordings of the evaluation sessions are transcribed using Wispher, to account for the security of the sensitive information in the recording, the model was run locally.

MarianMT model (local) - The Dutch text is translated to English locally using the MarianMT model (Helsinki-NLP, 2021), the model was run locally.

ChatGPT (limited use) - Limited use for: explaining and contextualising novel terms from literature during literature research, for my own understanding. Support in writing, for example, when struggling to make a sentence make sense, or to find a proper term or synonym. And a few times for inspiration on how to linguistically stitch together concluding notes in pre-drafts. And to contest my use of specific idiosyncratic terms.

Grammarly – Grammarly was used after the removal of sensitive information to correct clear writing mistakes.

Table of contents

1. Introduction

In recent years, incidents like the Dutch childcare scandal, which resulted in the destruction of the lives of innocent citizens and the resignation of the national government, have put the risks of using artificial intelligence (AI) in policy execution in the limelight (Peeters & Widlak, 2023). The use of artificial intelligence systems in the preparation of policy, or *'AI for policy'* has received far less attention (Kuziemski & Misuraca, 2020), but is fundamentally challenging to our democratic polity (Bullock, 2019; Newman et al., 2022).

The recent advancements in machine learning and the development of large language models have significantly broadened the scope of conceivable AI involvement in policy development, extending far beyond simple automation (Dafoe, 2018; Newman & Mintrom, 2023). At the same time, it is clear that the use of AI brings with it significant risks. To make responsible adoption of AI possible, ministries need practical and accessible guidelines for policy preparation professionals (Zuiderwijk et al., 2021).

In the wake of the rapid developments of AI, there have been calls for the Dutch government to invest in the adoption of AI, for instance, by the Netherlands Scientific Council for Government Policy (WRR, 2021). This design project seeks practical insight into the challenges the government faces in the adoption of responsible AI for policy.

1.1 Problem definition

Due to the novelty of the technologies, the effects of using AI in the development of policy are uncertain (WRR, 2021). What is certain is that AI affects the processes it is introduced into; AI is non-neutral (Janssen et al., 2022; Kuhn, 1970; Stinson, 2022). By using a non-human intelligent advisor, the policymaking process is fundamentally altered. Risking the depoliticisation of policy preparation leads society to drift based on unpolitical technocratic inertia. Another risk is the technocratic weaponisation of AI, when AI systems are used to legitimise unsupported policy. Prominent barriers hindering the adoption of AI for policy are fragmentation (with and between ministries), a lack of insight into AI opportunities, perceived uncertainty, risks involved, required effort, a lack of AI literacy, and a lack of AI capacity. The latter, caused by previous divestments in technical infrastructure and capabilities. In addition, the government has notorious difficulties in sourcing technical systems . Adoption is further hampered by a lack of processual perspective, ethical dilemmas that stretch beyond legality, and a lack of empirical validation within government. Worries about general risks of the use of AI, in combination with technical, organisational and political barriers, have resulted in limited adoption of AI systems in the process of policy development. Some of these barriers are exacerbated by a lack of a clear consensus on what constitutes "responsible AI for policy", and how it can be achieved in practice.

This is consistent with other countries where we see limited adoption in government, like Canada (Madan & Ashok, 2024), and AI for policy, highlighting reluctance, which is in sharp contrast to the public sector.

This thesis, and the design project that is part of it, are motivated by the combined needs from practice, in the form of the governmental organisation, and by gaps in academic literature. The main gap in the literature this project intends to contribute to the dissolution of is in the lack of translation of theoretical insights to the development of practical application, policy preparation practice (Zuiderwijk et al., 2021). Another gap exists around fundamental questions specific to the use of AI systems in the policy preparation phase. And there is a need for the development of tools that take the unique value logic of public organisations into account (Fatima et al., 2022).

The governmental organisation is presented with vast opportunities to use AI in the policy preparation stage, but it is under-prepared for the adoption of these novel technologies . In part because she lacks standards and guidance on the responsible, safe, and effective use of AI. While the governmental organisation, as an organisation, is aware of the general risks involved in the use of AI, the average policy official has little insight into these risks.

Fundamental risks like depoliticisation as a result of AI adoption are largely overlooked and require practical translation to be put on the agenda. Given that when AI for policy implementation is done irresponsibly, citizens and democratic institutions will be harmed because of it inadvertently, in return, hampering further adoption. Furthermore, governmental policy on the use of AI is highly restrictive, and the expertise and capabilities needed for responsible AI implementation are difficult to find within the organisation . Making it difficult for individual policy officers and the organisation as a whole to learn how to use AI responsibly. This suggests that the governmental organisation needs guidelines and a practical tool to support the responsible use of AI by policy preparation professionals.

Informed by existing frameworks and insights from international organisations such as the *Recommendation of the Council for Agile Regulatory Governance to Harness Innovation* (OECD, 2021), the *AI Impact Assessment (AIIA)* (ECP, 2018), and relevant academic literature like the *circular value-based assessment framework* developed by Yurrita et al. (2022), and the model of *Features contributing to contestable* (Alfrink et al., 2022, p. 629), *AI five loops model* for the systemic integration of contestability in AI-based policy execution (Alfrink et al., 2023).

1.2 Design Project

This thesis takes a constructive design research approach (Koskinen, 2011), using the construction of a design prototype, in the context of a design project for a governmental orgnisation, as a means of creating knowledge.

Design Assignment
Based on the initial definition of the design challenge for the governmental organisation, the design assignment at the start of the project was the following: *Design a set of guidelines and a prototype tool for assisting professionals within Dutch ministries in the ethical and effective utilisation of AI in the policy preparation stage. The tool should fit the existing organisational environment and resonate with the practical experiences of these professionals, ensuring they can capture the benefits of AI while effectively mitigating associated risks.*

Figure 1 shows an initial map of the design challenge associated with this design assignment.

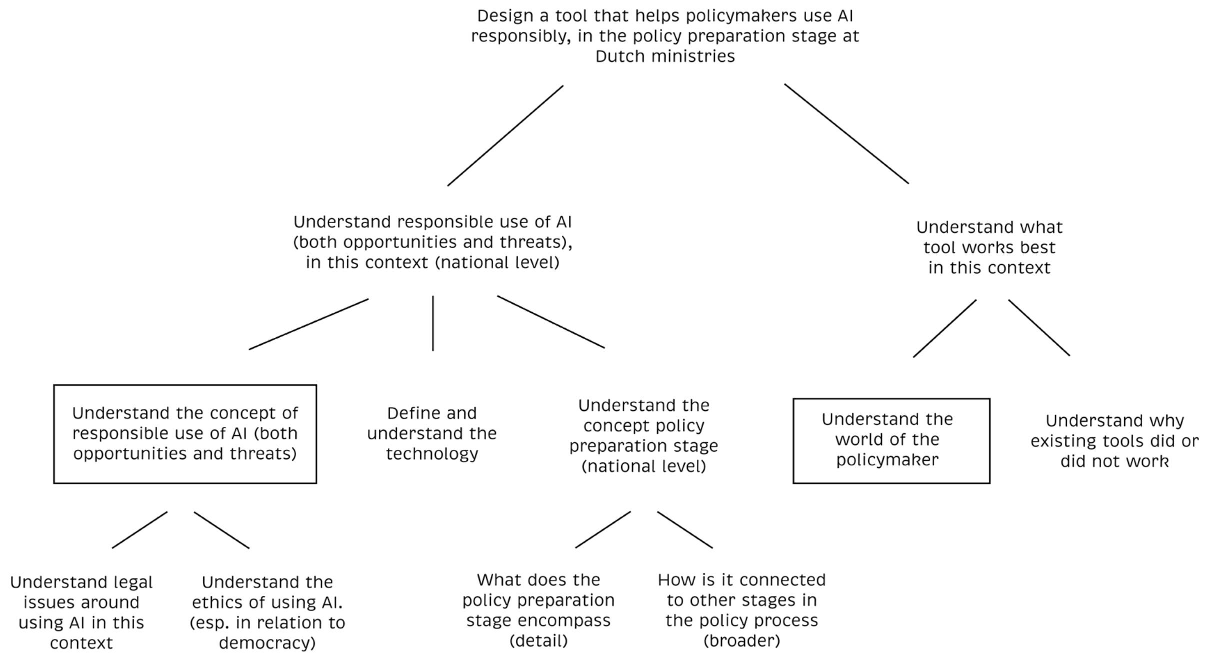

Design a tool that helps policymakers use AI responsibly, in the policy preparation stage at Dutch ministries
Understand responsible use of AI (both opportunities and threats), in this context (national level)
Understand what tool works best in this context
Understand the concept of responsible use of AI (both opportunities and threats)
Define and understand the technology
Understand the concept policy preparation stage (national level)
Understand the world of the policymaker
Understand why existing tools did or did not work
Understand legal issues around using AI in this context
Understand the ethics of using AI. (esp. in relation to democracy)
What does the policy preparation stage encompass (detail)
How is it connected to other stages in the policy process (broader)

Figure 1. Initial mapping of the design challenge

Research questions

The main research question (RQ 0) is derived in accordance with the design assignment. The two subsidiary research questions (RQ1 & RQ 2) are established based on the initial mapping of the design problem. Answering these subsidiary research questions is needed to answer the main research question. The research questions are:

RQ 0: *How can we design practical tools to guide the use of AI in the governmental policy development process?*

RQ 1: What does responsible use of AI entail in policy preparation?

RQ 2: How should a tool for policy preparation professionals be designed to effectively incorporate advice?

The design approach (see Figure 3) for the project is based on three theoretical design approaches: Frame Innovation (Dorst, 2015b), Vision in Product Design (ViP) (Hekkert & van Dijk, 2011) and Value Sensitive Design (VSD) (Friedman & Hendry, 2019). These approaches together cover the wider field of challenges that were to be expected throughout the project. Adding Value Sensitive Design due to the large role values play in the design challenge at hand.

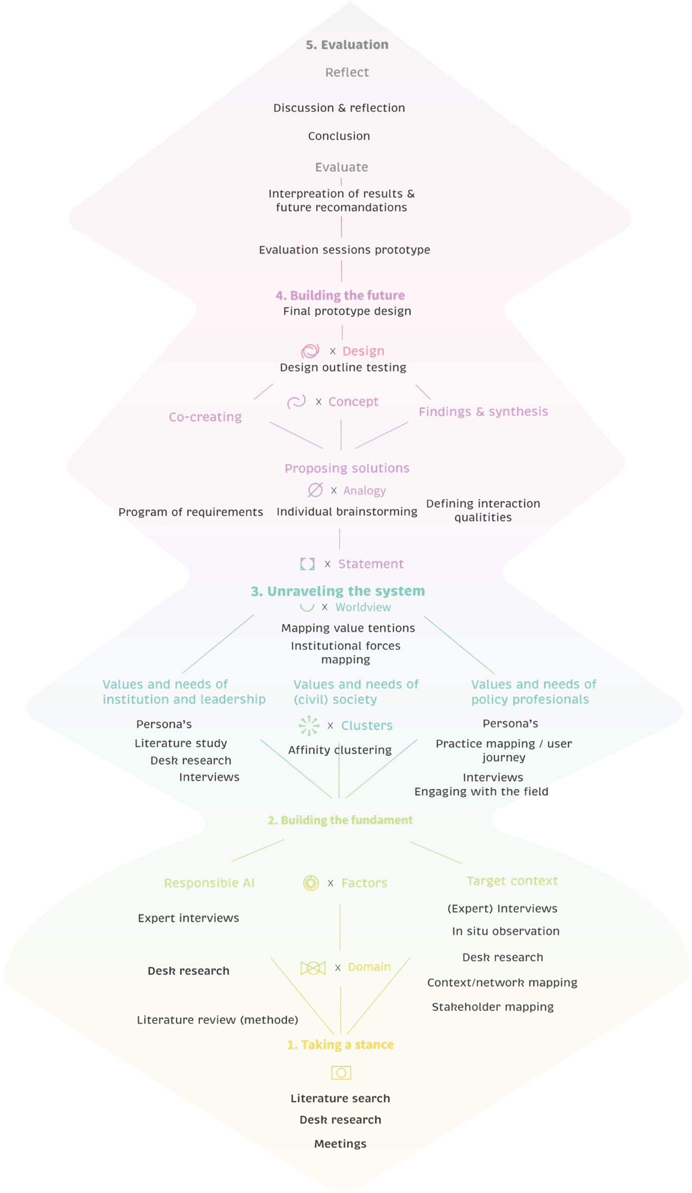


Figure 2. Overview of the design process.

1.4 Relevance and Contributions

The original design assignment includes a reference to a *set of guidelines* and *a prototype*. Because the desk research in the project resulted in finding existing ones, including one specifically developed for the Dutch government (Meijer & Ruijer, 2021), reducing the need for a novel set produced as part of this project. The project's focus their for shifted to mechanisms to realise responsible AI in practice, which is a distinct and complex problem of its own.

The design project resulted in a prototype that was tested with participants from the governmental organisation.

The evaluation results support the notion that there is a tension between the values and needs of leadership and the organisation, and those of users, the policy officers and others involved in policy development. Political leadership and upper management are mostly concerned with avoiding risks and ensuring proper regulation before the use of AI. Policy professionals, on the other hand, only see themselves responsibly implementing AI when they are taken by the hand and supported through all steps of the process. Not necessarily because they disagree with the ideal of broad AI literacy and full accountability, but because they experience a misalignment between that ideal and the reality of their work.

The prototype is able to incite curiosity in users and, at the same time, provoke hesitation. Some users become interested and want to explore AI's potential in their own practice. Others are more cautious, especially when the risks are made prominent, and the responsibility for implementation remains unclear. This difference in effect, between motivation and restraint, reflects the core design principle of the tool, to balance encouragement with responsibility. This expectation is in line with the general goal as expressed by the governmental organisation and the intended role of the Policy Compass. The experience of a slightly restrictive or tempering effect might be a necessary condition for inviting users to explore critically, instead of rushing into adoption. Finding tension between what is desirable from a responsible AI perspective and user preference.

This thesis contributes to the existing literature by responding to the call for empirical, situated work in the context of AI in public governance (Zuiderwijk et al., 2021). It aims to help close the gap between theoretical insights and practical application in policy preparation practice. There is limited testing or validation of responsible AI tools in situ, especially in the early phases of policy development. This design project explores what happens when you introduce a tool in that space. A key contribution is the idea of AI curiosity as a precursor to AI capability. The prototype does not aim to teach everything at once, but to prompt questions, dialogue, and individual motivation, as a basis for distributed capacity-building. This connects to earlier work on strategic alignment (van Noordt & Tangi, 2023), but shifts the focus to soft capabilities, including user engagement and reflective decision-making. The design reflects a practical response to the gap around responsible AI in the preparatory phase of policy, where implementation questions and institutional effects are still underexplored. The findings suggest that tools like this can help bridge the gap between central AI strategies and day-to-day policy work, not by providing answers, but by creating space to think, ask, and discuss.

1.5 Structure of the Report

This thesis is structured as follows: Firstly, Chapter 2 provides the background on responsible AI for policy based on literature. Discussing AI and government, AI for policy, responsible AI, and curiosity. Next, situating the design challenge, Chapter 3 gives an overview of the design context. Including an exploration of systems involved in policy development in Dutch democracy, looking at Dutch governments' relation with AI, and mapping the people and organisations in the design context. Subsequently, moving to the design project: method, design approach, design process, final design prototype, evaluation and results are discussed in Chapters 4, 5, and 6. Finally, the thesis reflects on the findings and the overall design process, presents a synthesis including proposed framework sketches, and relates these back to the literature in chapter 7.

2. Background

This chapter discusses the theoretical and empirical background based on the academic literature. Positioning this thesis and providing the theoretical foundation for the design context and the design project, including the resulting proposed prototype design. These are discussed in the subsequent chapters.

The chapter is structured as follows: starting with a discussion of the AI and government. Followed by a more elaborate exploration of AI for policy, discussing its use cases, benefits, risks and barriers to adoption. The next section looks at responsible AI, first by framing the concept theoretically, before discussing a variety of different proposed approaches. The final section introduces curiosity, which takes a key role in the proposed prototype design. The large sections on AI for policy and responsible AI include a 'Conclusion' subsection that synthesises the key findings and relates them to the design project, forming the theoretical building block of the design.

2.1 AI and Government

Governments have a dual role in the adoption of Artificial intelligence in society. Firstly, they need to develop rules and regulations to protect society against damage from irresponsible or outright dangerous uses of the technologies (Cath et al., 2017; Kuziemski & Misuraca, 2020). Secondly, governments can stimulate the development of the technologies, for instance, in the form of investments in the required education or infrastructure, or through the provision of direct investment and subsidies (Guenduez & Mettler, 2023), and public-private collaboration with the sector. Engendering the private sector to develop AI activities, -knowledge and -infrastructure (van Noordt et al., 2023), to capture economic growth and other societal benefits, and to stay competitive in the fierce global competition for AI dominance (Guenduez & Mettler, 2023; Satariano & Mozur, 2024).

In addition to regulating and facilitating AI in society, governments themselves can use AI. The development and utilisation of AI in the public sector has lagged behind that in the private sector, as has attention for AI in government and academia (Desouza et al., 2020). However, in recent years, the field has grown substantially on the coattails of rapid developments around AI in the public sector, and the publication of governmental AI policy documents (van Noordt, 2023). AI is seeing increasing use in governments, and academia has theorised many more use cases for the public sector (Madan & Ashok, 2023; Wirtz et al., 2019; Zuiderwijk et al., 2021).

Continuing, this section discusses the definition of AI, based on an exploration of definitions used in academic and government-related practice. And giving an overview of the use cases of AI in government.

2.1.1 Definition of AI

Many different definitions of artificial intelligence are used in academia and practice. In general, the term refers to a variety of "intelligent" computational technologies. Herein, the meaning ascribed to the term intelligent varies among fields and is continually evolving.

AI definitions have historically been characterised by terms relating to human- or rational thinking and actions (Russell & Norvig, 2010). An example of humanlike intelligence framing is found in UNESCO's (2021) definition, referencing technologies with "a capacity to learn and to perform cognitive tasks" (p. 10). Other definitional approaches focus on the technical characteristics, for example, the definition by Campion et al. (2022) that prescribes "utilizing ML and big administrative data" (p. 2). Studying policymakers- and academic AI definitions, Krafft et al. (2020) found a tendency of policy documents to favour human-thinking or behaviour characteristics, whereas AI researchers used more technical specification-heavy definitions. The rigidity of the discriminatory principles of a definition can vary equally. Definitions used in regulation and law have to be rigidly discriminatory to be enforceable (Krafft et al., 2020). Definitions used by designers can be more open, to allow for greater conceptual flexibility and the creation of novel insights and designs. For instance, using metaphors to conceptualise a design problem around public AI (Alfrink et al., 2024), adding a great degree of conceptual framing without defining.

Constructing rigid definitions of AI is made challenging by the rapid developments of AI technologies. Anticipatory definitions in regulatory practice are increasingly agile (OECD, 2021), and technology-neutral (European Parliament, 2023; European Union, 2024). Moving away from definitions based on listed technologies and focusing on a system's functions, intentions, or outcomes. Making the regulations more adaptive in the face of developments (Mul & Werkhorst, 2020).

This report follows the definition of the Organisation for Economic Co-operation and Development (OECD, 2024), as it aligns well with the understanding of AI in academic discourse and policy-facing practice (Krafft et al., 2020). A selection of recent AI definitions in governmental contexts, including the OECD (2024) The definition can be found in Table 1.

Source	Definition
(UNESCO, 2021, p. 10)	"AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments."
(Campion et al., 2022, p. 2)	"…an operational definition of AI in the public sector as a set of technologies, solutions, and processes designed to augment policy makers' decision making by utilizing ML and big administrative data."
Article 3 (1) of the EU AI Act	"'AI system' means a machine-based system that is 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;"
OECD, 2024	An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Table 1. Selection of definitions of AI in governmental contexts

2.1.2 AI in Government

Within government, AI can be used to support the delivery of public services, internal management, and policy-making (van Noordt & Misuraca, 2022). The use of AI for the delivery of public services can lead to better information provision, more understandable information for the public, and innovative services, for instance, through personalisation. Improving how citizens experience services (Zuiderwijk et al., 2021). Internal management use cases are often focused on making operations more efficient, but they can also contribute to increased transparency and consistent decision-making (van Noordt & Misuraca, 2022).

Compared to the private sector, governments face a more complex set of challenges in the adoption of AI. Governments deal with a broad network of stakeholders satisfied by a variety of different types of value, far beyond the financial, as governmental AI is held to a strict standard of advancing the public good (Cath et al., 2017). Another difference with the private sector is the power of the state, which increases the potential damage that can be done as a result of the irresponsible use of AI (Peeters & Widlak, 2023). This contributes to the need for governments to implement high levels of transparency and accountability (Desouza et al., 2020). Additionally, governments are highly scrutinised, for instance, by the media (Desouza et al., 2020), who amplify mistakes or harm resulting from governmental use of AI to greater proportions than those in the private sector. This leads to a greater impact on the trust of citizens in the government (Margetts & Dorobantu, 2019).

Continuing, in the following section, the use of AI in the policy development process, or *AI for policy*, is explored in detail.

2.2 AI for Policy

In this thesis, '*policy'* is used to refer to both policy goals, what the government wants to happen, and policy instruments, the measures devised to achieve the policy goal (Linder & Peters, 1989), for example, government programs and legislation. The term '*AI for policy*', proposed by this thesis, refers to the use of AI in the policy development process. This use of AI for policy is the most recent evolution in greater trend towards evidence based policy making, where "factual" evidence in de development op policy theories as a basis for policy (Sanderson, 2009), for instance, using big data (McNeely & Hahm, 2014), improving policy quality by reducing arbitrariness and increasing the certainty of analyses and projected effects of policy (Vydra & Klievink, 2019). At least in theory, views critical of this paradigm are discussed in the subsection on responsible AI. AI can be used in processes throughout the policy cycle, for example, by aiding rapid detection of problems in society to advance agenda-setting. And improved policy making through better predicting policy effects. Monitoring and evaluation of policy and its implementation. And increase citizen participation in the policy process (Valle-Cruz et al., 2020; van Noordt & Misuraca, 2022).

The term 'AI for policy' is not widely used in the related field, but it is a useful term that succinctly captures the underlying concept. It is in line with terms used in the literature on governmental AI, for example, "AI for policy making" (van Noordt & Misuraca, 2022, p. 5), describing a similar concept. And it uses a similar structure and logic as the terms "analysis *for* policy" and "analysis *of* policy" used in the adjacent field of policy analysis (Hill & Varone, 2021, p. 5).

AI for policy can be distinguished from other concepts that relate AI to policy. Like 'AI policy', 'AI in policy' and 'AI for the execution of policy'. For an overview of these concepts, including a description, see Table 1.

AI & policy relation	**Description**
AI for policy	The use of AI in the development process of policy.
AI policy	Laws, regulations, strategies, and public programs on the development, implementation and use of AI.
AI in policy	Provisions on AI in non-AI-related policy. For example, the inclusion of a legal basis for the use of AI in the enforcement of policy.
AI for the execution of policy	The use of AI to execute policy, for example, for the automation of service delivery and policy enforcement.

Table 2. Typological overview of relations between AI and policy

AI for policy is a small subset of both the use of AI in practice (van Noordt & Misuraca, 2022) and of the academic discourse on governmental AI. Limited work on AI in policy preparation exists. AI for policy is more often discussed as a part of governmental AI at large (e.g. van Noordt & Misuraca, 2022; Zuiderwijk et al., 2021). Furthermore, literature on AI for policy is in large part explorative, lacking empirical validation . This is due to the lack of adoption and the difficulty of studying the policy process due to its complexity , among other reasons. This means that the characteristics and dynamics of AI for policy are poorly studied, and have to be extrapolated from work on related subjects and practices..

2.2.1 Technologies for AI for Policy

As discussed in the subsection on definitions of AI, the exact bounds of what technologies count as AI are somewhat opaque and contested. More concrete are the technologies discussed in the context of AI for policy.

The current attention for AI in popular culture is fuelled by the rapid advancements in generative AI and the resulting increased commodification of AI. However, these models are complex and only applicable to a small subsection of AI for policy use cases. Many AI technologies that can be used in AI for policy are much simpler, narrow AI technologies (Samoili et al., 2020). Which can be more differentiated (Dwivedi et al., 2021), tailored to fit specific use cases and context. This is reflected in Zuiderwijk et al.'s (2021) summary of technologies that fall within the scope of the governmental definitions of AI:

> "Approaches and technologies that comprise an AI system may include, but are not limited to: machine learning, including supervised and unsupervised learning (Smola & Vishwanathan, 2008; UNESCO, 2020); Artificial Neural Networks (Krenker, Bester, & Kos, 2011); fuzzy logic (Klir & Yuan, 1995; Yen & Langari, 1999); case-based reasoning (Cort́es & Sanchez-Marre, 1999); natural language processing (Liddy, 2001); cognitive mapping (Eden, 1988; Golledge, 1999); multi-agent systems (Ferber & Weiss, 1999); machine reasoning (Bottou, 2014), including planning, predictive analytics, knowledge representation and reasoning, search, scheduling, and optimization; and, finally, cyber-physical systems (Baheti & Gill, 2011; Lee, 2008; Radanliev, De Roure, Van Kleek, Santos, & Ani, 2020), including internet-of-things and robotics, computer vision, human-

> computer interfaces, image and facial recognition, speech recognition, virtual assistants, and autonomous machines and vehicles.“ (p. 2)

A meta-empirical study on European governmental AI projects, by Van Noordt & Misuraca (2022), concurs with the relevance of more narrow AI for policy making. Finding many instances of the uses of computer vision and identity recognition, predictive analytics and threat intelligence, some uses of machine learning/deep learning, natural language processing/ text mining/speech analytics, AI-empowered knowledge management and one instance of security and threat intelligence.

In conclusion, the advancements in generative AI have created a lot of buzz for AI. Many other AI technologies have (potential) use cases in AI for policy. Generative technologies create novel opportunities for the policy development process, and accompanying attention for AI for policy. In this, it can potentially serve as a beachhead for the adoption of more situationally differentiated, but less glamorous and technically demanding, narrow AI technologies.

2.2.2 Use Cases of AI for Policy

Policy development is a multifaceted process that consists of a constellation of formalised processes and activities, with informal and ad hoc interactions mixed between (Fischer et al., 2007), carried out by actors inside, working for, and around the government (van den Berg et al., 2015). The variety of processes involved creates many potential avenues for the use of AI. This is reflected in the literature, which theorises and describes use cases ranging from improving the efficiency of the policy-making process (Zuiderwijk et al., 2021), to solving problems that were unsolvable before, for example, the increasingly complex crises society is faced with, such as climate change (Coeckelbergh & Sætra, 2023; Februari, 2023).

A variety of different types of use cases of AI for policy are described in the literature, although most are embedded within AI in government typologies (e.g. van Noordt & Misuraca, 2022; Zuiderwijk et al., 2021). And not many AI for policy-specific categorisations of use cases are used. Valle-Cruz et al. (2020) use the structure of the policy cycle, to categorise AI for policy use cases using its phases: a*genda-setting, policy formulation and decision-making, implementation,* and *policy evaluation*. This categorisation is therefore used in the discussion of use cases that follows, with the addition of a category of *Throughout the policy-making process* (illustrated in Figure 4), completing the typology as discussed in the conclusion to this section.

In the *Agenda-setting phase*, AI can aid the collection of data, for example, from social media (Loukis et al., 2017), and help in analysis by identifying patterns in large and complex data (Desouza & Jacob, 2017). This can support faster detection of societal issues (van Noordt & Misuraca, 2022), as well as aiding the accuracy of problem identification (Zuiderwijk et al., 2021), encouraging collaboration and improving legitimacy (Valle-Cruz et al., 2020).

In the *policy formulation phase*. AI can assist in the generation and evaluation of policy options (Valle-Cruz et al., 2020). For instance, by providing improved forecasting and the simulation of policy options (Margetts & Dorobantu, 2019). Additionally, AI is thought to be able to improve citizen participation (van Noordt & Misuraca, 2022), by facilitating communication with citizens and processing public input during consultations (Zuiderwijk et al., 2021)..

In the *decision-making phase,* judgments are in part based on analyses done in the policy formulation phase; the increased quality of analyses and simulations in that stage finds its use similarly in the decision-making stage. Supporting decision-making by predicting policy outcomes (Valle-Cruz et al., 2020), and aiding the accuracy of decisions (Zuiderwijk et al., 2021).

Furthermore, AI could help build trust and accountability by making decisions accessible to the public (Valle-Cruz et al., 2020).

In the *implementation phase*, AI supports automation, better use of resources, and streamlining of processes (Valle-Cruz et al., 2020). Part of the implementation phase is the communication of the policy to the general public and other stakeholders; here, AI can support the tailoring of the communications (Androutsopoulou et al., 2019), especially for vulnerable groups who encounter challenges in understanding the complex language typically used by governments. Furthermore, AI can help in the monitoring of the implementation of the novel policy (van Noordt & Misuraca, 2022), allowing for faster evaluation and adjustment when necessary.

In the *policy evaluation phase*, AI opens up new ways to analyse feedback, monitor impact in real-time (Valle-Cruz et al., 2020). Allowing for the development of more dynamic and responsive evaluation approaches, supporting the evaluation of existing and novel policies (van Noordt & Misuraca, 2022) and helping identify where action is needed (Zuiderwijk et al., 2021), feeding pack to the agenda-setting phase of the policy cycle.

Some use cases of AI for policy are applicable *throughout the policy-making process.* This includes general applications for administrative streamlining, like improving processes or digital coordination (Madan & Ashok, 2023), to reduce administrative burden and increase efficiency (Zuiderwijk et al., 2021). These use cases are especially relevant given the importance of efficiency as a motive behind AI projects in the government context at large (Madan & Ashok, 2022). Ideally, these use cases free up civil servants' time and attention for more substantive tasks (Madan & Ashok, 2023), contributing to higher policy quality.

In conclusion, there is a wide variety of AI use cases that have the potential to improve the policy-making process. For the categorisation of use cases of AI for Policy, the thesis follows Valle-Cruz et al. (2020) in using the phases of the policy cycle[1]: a*genda-setting, policy formulation and decision-making, implementation,* and *policy evaluation*. This is a useful way to categorise the use cases, given the familiarity of people in the policy chain with the policy cycle model, and the alignment of the model with practice. However, several general use cases are identified in the literature that can provide support t*hroughout the policy-making process,* making this a meaningful addition as a separate category. The AI for policy typology used by this thesis is illustrated in Figure 4.

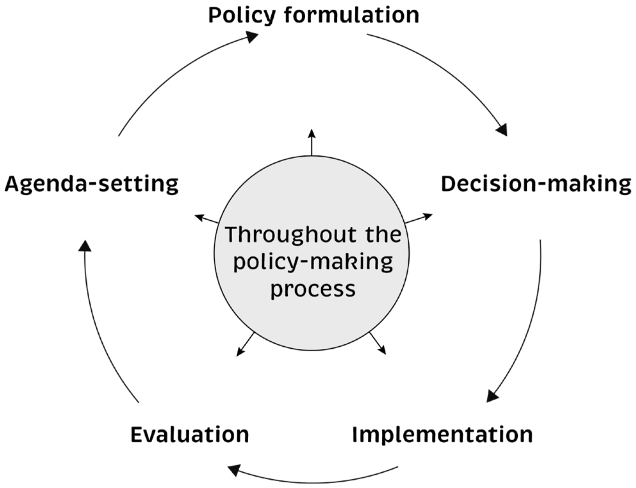


Figure 3. Types of AI use cases in the policy-cycle

Over the phases of the policy process, AI for policy can support the collection and analysis of data to identify policy issues more accurately and faster, allow more participation and help in generating and evaluating policy options, and aid decision-making through improved forecasting. As well as supporting the implementation of policy by automating

[1] In the context of the Dutch government and the Policy Compass, a version of the policy cycle is used that includes an additional sixth phase of reorientation after the implementation (KCBR, n.d.; van der Staaij & Sneller, 2023). This reorientation phase is a special type of agenda setting, however AI for policy use cases for are more connected to the evaluation phase, of which reorientation is an extension.

processes and improving communication with citizens. In the evaluation, creating capacity for real-time monitoring for dynamic feedback. Some use cases have potential use cases throughout the entire policy process, such as administrative streamlining and coordination, reducing administrative burdens and allowing civil servants to spend more time on substantive tasks.

That AI has potential responsible use cases in the policy-making process is not overly contentious, outside of critical discourses that take issue with AI for policy in principle, as discussed in the section on responsible AI. Practically speaking, the main contentions concern the extent to which AI can actually improve the policy-making process, and how much of the theorised possibilities hold up when faced with technical limitations and the complex nature of public administration .

2.2.3 Benefits of AI for Policy

AI for policy is promised to bring a variety of benefits. Similar to the use cases to which the benefits are related, the benefits of AI for policy are enjoyed throughout the policy cycle. And the benefits can continue during the lifetime of the developed policy, where the improved quality policy provides public value (van Noordt & Misuraca, 2022). Academic work on benefits specific to AI for policy is limited. Overviews and typologies that include AI for policy benefits are mostly broader, including various forms of AI for government benefits. For example, Madan & Ashok (2023) describe AI in government outcomes grouped around public values and how government organisations are structured and operate. Zuiderwijk et al. (2021) outline efficiency and performance, risk identification and monitoring, economic gains, and improved data and information processing. They point to broader benefits for society, decision-making, engagement and interaction, and sustainability. Continuing, this subsection discusses the benefits of AI for policy, structured according to a typology proposed by the me, described in the concluding paragraph of this subsection (see Figure 5), namely *policy quality benefits, democratic benefits*, and *efficiency benefits*.

Policy quality benefits
The use of AI for policy can benefit policy quality, both substantively (Zuiderwijk et al., 2021) and by improving responsiveness (König & Wenzelburger, 2020). These benefits result from improvements to the various processes of the policy process. Additionally, AI may be able to facilitate new ways of doing, to solve problems that were not solvable before, like complex challenges such as climate change (Coeckelbergh & Sætra, 2023). AI can support efficient information handling and higher-quality analysis (Zuiderwijk et al., 2021), benefiting policy quality through improved problem definition. For example, in the identification of social bottlenecks (van Noordt & Misuraca, 2022). Or, by processing data about opinions and behaviour, it provides a better understanding of what citizens think and need (Loukis et al., 2017) . AI models can aid the development and selection of policy options by improving the modelling and forecasting of policy options and new policies (Margetts & Dorobantu, 2019), improving policy option formulation and more structured and better informed decision-making , and reducing uncertainty and leading to better policy options and decisions (Vydra & Klievink, 2019). Alongside these benefits to policy quality due to the substantive advancements, AI for policy can aid policy responsiveness by supporting faster problem identification (König & Wenzelburger, 2020; van Noordt & Misuraca, 2022) and accelerating development. Allowing issues to be addressed quickly, before they escalate.

Democratic benefits
The democratic benefits of AI for policy include improvements to democratic values like openness, fairness and equity, and benefits to the functioning of the democratic polity.
AI can support democratic functions, such as transparency and accountability, by making it easier to keep track of decisions and retrieve records (Chen et al., 2023). Openness can be aided by helping to make public sector information more widely available or more easily searchable (Twizeyimana & Andersson, 2019). AI for policy could also benefit the democratic nature of policy development by supporting increased public participation or co-creation with citizens, by facilitating the involvement of large groups of citizens in policy-making (van Noordt, 2023). It can improve governments' interaction with citizens. Altogether, AI can help make policy accessible, improve responsiveness, and support transparency, factors that can contribute to building and maintaining trust (Madan & Ashok, 2023), as may the improvements in decision-making (Zuiderwijk et al., 2021). Furthermore, AI for policy can help improve policy from the perspective of the rule of law *[Rechtstatelijkheid]*, for instance, by helping to select the most appropriate policy instrument, and supporting policy officials in understanding the legal context, improving policy alignment with existing legal norms like treaties and laws.

Efficiency benefits
In the literature on the benefits of AI for policy, efficiency takes a less prominent place than in discourses about the private sector (Cath et al., 2017). However, efficiency is an important consideration in policy development , and an important motivation behind AI projects in the government context at large (Madan & Ashok, 2022). Among the efficiency benefits are increased (labour) productivity, with AI making it easier to gather information and develop policy options (Zuiderwijk et al., 2021). Additionally, administrative and repetitive "boring" work can often be taken over by AI . Reducing the man-hours invested and potentially improving the work experience of policy officers. Efficiency benefits can, in turn, lead to reinvestment of these gains into activities that add to policy quality and the democratic nature of the policy development process. Organisationally, AI can improve efficiency by streamlining internal workflows (Madan & Ashok, 2023).

In conclusion, AI for policy is expected to bring a wide range of benefits that span all phases of the policy cycle and extend into the lifetime of the policies themselves, where better policies can generate greater public value. Because AI for policy-specific frameworks remain limited, this thesis uses an original categorisation. Based on a synthesis of the benefits described in the literature. Derived by filtering for AI for policy use cases and related benefits, and clustering similar types of benefits, the benefits can be grouped into the following three categories (illustrated in Figure 5):

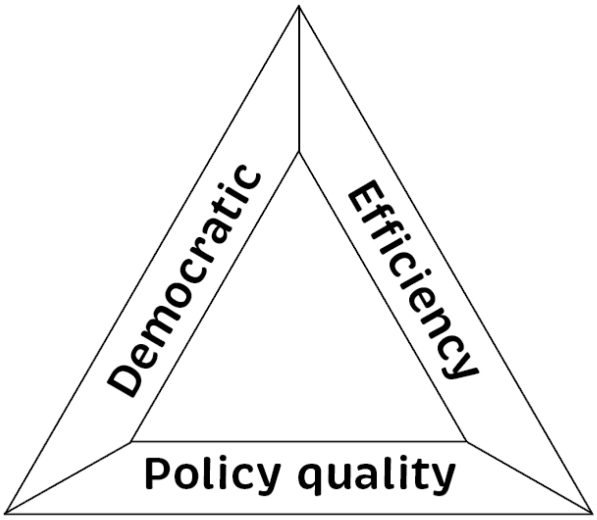


Figure 4. Triad of types of benefits from AI for policy

Policy quality benefits, including improved policy development processes and decision-making, better policy and the resulting benefits to society.

Democratic benefits include improvements to democratic values like openness, fairness and equity. And procedural improvements like increased public participation. And contributions to the rule of law.

Efficiency benefits include improved labour productivity, reduction of administrative tasks and resulting financial savings.

2.2.4 Risks of AI for Policy

The use of AI for policy can propagate a variety of risks. Some of these are inherent to artificial intelligence, think of risk related to the technical functionality model, which exists independently of the application domain or environment. Other risks are unique to AI for policy, related to fundamental changes to the institutions and processes of policy development .To allow for responsible AI for policy, these risks have to be understood so they can be counteracted for AI to be responsible. At the same time, the underutilisation of AI for policy poses its own risks, such as the societal costs of missed opportunities and unrealised gains in policy quality and other benefits (Floridi et al., 2018).

The risks involved in AI for policy are identified in various domains and conceptual levels. Chen et al. (2023) describe three types of challenges: societal governance challenges, like authoritarian abuses, the replacement of jobs and the disappearance of human discretion, and increased power asymmetry; data quality, processing and outcome challenges; and public value challenges. Similarly, specified for Generative AI Sætra (2023) describes three levels. On the micro level, it affects individuals and relationships through manipulation and cognitive effects. On the meso level, it influences organisational dynamics, bias, and power relations. On the macro level, it impacts democratic processes, institutions, and societal structures. However, no typologies specifically suited to risks of AI for policy were found.

Continuing, the risks of AI for policy identified in the literature are discussed according to typology presented by me in the conclusion of this subsection (see **Error! Reference source not found.**): *Model and user risks*, *Ethical and societal risk*s, *Liability risks*, and *System risks.* Described through their associated underlying types of risks (see ***Error! Reference source not found.).***

2.x.x.x Model and User Risks

Technical and data risks

Technical risks include a set of risks related to the AI models themselves and associated data. . Bias, etc. Robustness . Bias . Importantly, compared to the private sector, the impact of these risks can be aggravated by the fast power of the state .

The use of AI often implies the use of data, which can result in data security and privacy risks (Valle-Cruz et al., 2020). Data can be severely sensitive to the privacy and security of individuals and organisations. These risks exist when models and data are not stored securely. Or when external or contracted applications are and there is no clear indication whether the data uploaded into commonly used AI systems is stored securely. For many systems, this cannot be guaranteed .

Human–AI interaction risks

Other risks arise from the interaction between users and the AI system. Users may not fully understand the capabilities of a system, or may use it inappropriately. users are often unaware of system limitations, while AI systems often lack insight into their own limitations and blind spots and provide limited information on them . The outcomes produced by AI models are predictions with an associated confidence level; if these are not presented to users, they are likely to misinterpret them. Especially when outputs are presented with a definitive of confidence. Exemplary of this is generative AI, which can produce hallucinations, generating human-sounding and seemingly plausible but incorrect information . AI models also often do not work fully as intended or expected , creating risks when compounded with overreliance, resulting from the human tendency to trust machines .

Another risk in the Human-AI interaction is that of losing expertise and skills, as reliance on AI can lead to their underutilisation and gradual erosion (Sætra, 2023). Furthermore, when AI for policy does not lead to tangible harm, it can have questionable effects, by shaping users frame a problem, for instance, encouraging a categorically instead of a holistic view, as that is how algorithmic systems structure data (Mittelstadt et al., 2016). Compounded, this can have transformative effects and create systemic risks.

2.x.x.x Ethical and Societal Risks

Ethical risk

In discussions on the risk of AI in government, much focus is on the ethical dimension . Ethical risks include risks to privacy, discrimination and fairness in general.

Ethical risks are often related to other risks. They can result from lower-order risks, like technical risks, such as data or model bias, or human AI interaction risks. And often translate into outcome risks to data or decision subjects, or for people using the systems. The latter is the case for ethical risks for which the responsibility is codified in a legal sense, creating legal liability risks. The perception of moral responsibility can also result in political consequences, meaning the ethical risk turns into a political risk.

Societal risks

Societal risks are risks to society resulting from AI for policy. These include risks of bad policy, such as discriminatory policies, or policies that are difficult to implement, or policies that have negative effects on society in general or for certain groups. In both the literature and practice, there is much focus on the socio-economic risks of adopting AI, such as job loss, changing roles, and the need for retraining, as well as challenges linked to automation, including reduced transparency in how systems work and ensuring that implementation respects people's dignity and rights (Madan & Ashok, 2023; Valle-Cruz et al., 2020). Other societal risks that are discussed are related to the large amount of resources like energy used by AI systems and the potential consequences this has on the environment .

2.x.x.x Liability Risks

Liability is an important motivator for servants. In a governmental organisation, key concerns in all actions relate to the political risk (Hood, 2011). Liability risks refer to the risk of facing consequences when an accountability forum passes judgment (Bovens, 2007). This can occur within the organisation itself, for example, when a manager reprimands a policy official. Or external, when a judge orders the government to take action or compensate. When the accountability for other risks is codified, they can translate into some form of legal risk. In government, liability risks extend beyond the legal, as political leadership can be held to account by the media and ultimately by parliament, which can judge them liable and remove them from office . Internal liability risks thus can result from the distributions of liability risks before external accountability forums, such as a judge (legal) or media and parliament (political).

Legal risks

The use of AI for policy comes with several legal risks. These risks can result from the outcome of using AI for policy or from the process in which it is used. Relating to the outcome, AI systems themselves cannot be held accountable for their conclusions or findings, which means the responsibility ultimately falls on the user, unless it is delegated to another party . Legal risks in the process when AI is used include breaches of privacy and data security. Juridical challenges around the use of data are mostly absorbed by the pre-existing legal infrastructure, like the GDPR, in the case of (Kulk & van Deursen, 2020). These risks are therefore similar to earlier data-use challenges associated with other technologies. The use of AI can also produce discriminatory outcomes, for example, based on gender, ethnicity, or background. Such bias may put the responsible public authority at risk of legal action if it results in unfair treatment or disadvantage .

Especially with generative AI, there is also a risk of copyright infringement, since the models are trained on large amounts of text and imagery. Creating risks of unauthorised use of copyrighted material, both by using the model and by reproducing the training data in results created by the model . Furthermore, when governments intend to correct mistakes or unfair treatment by means of compensation, they risk setting a legal precedent for future cases .

Political risk

In the literature, risks of AI for policy are mostly viewed from the perspective of harm to public value or ethical risks. However, in the context of AI adoption, they additionally need to be understood as political risks or risks to political leadership. Political risks arise from the use of AI for policy when it is perceived to have done harm, and political agents are held liable. Individual actors in political leadership can face political consequences, like removal from office , either directly by parliament, or indirectly through a decline of public popularity and support in elections. A type of risk related to political risk is that of political liability for the organisation, for instance, when parliament requires the compensation of citizens who have been harmed beyond what is legally required .

Conversely, democratic risks can arise from attempts to avoid political risks through the obfuscation of political responsibility by the delegation of decision-making responsibilities to AI systems (Chen et al., 2023).

2.x.x.x System Risks

Organisational and institutional risks

The interaction of users with an AI model creates risks of losing expertise and skill. When this is compounded at the organisational level, this creates a risk of losing organisational knowledge and capability . Other organisational issues related to the development and implementation risks of AI for policy include things like cost overruns .

Democratic risks

In the literature, there are various discourses on the risks of AI for policy to the democratic system of governance. These have identified risks both to the democratic ethos (normative) and to the functioning of the democratic polity (procedural and affective). The former relates to the ideal , and the latter to the ways this is achieved in practice, in the dimensions of institutions and practices , as well as the emotional dimensions that are equally important in its functioning .

Part of the discussion on democratic risks of caries over from literature on the larger shift towards evidence-based policy-making, in which decisions are supported by scientific research and data. Contributing to a trend of management by measurement, where complex social issues, such as social cohesion, are addressed through quantitative indicators (Gray & Mcdonald, 2006). This development is often linked to the criticised idea that numbers are more objective than other types of knowledge (Porter, 1995). Leading to the relevant idea that data and AI systems are "objective", which may discourage critical questions or the expression of opposing views (Newman et al., 2022; Porter, 1995). Altogether, creating risks of depoliticisation of the policy process, leading to increasingly technocratic policy-making (Kitchin, 2016; Newman et al., 2022). Although depoliticisation is thought to be a great evolution in the policy development process by some, mainly governmental institutions (Flinders & Wood, 2014). Many others, especially in academia, point out that policymaking is a political process in the end, in which relying heavily on evidence can mask inherent political trade-offs and value judgements (Kitchin, 2016; Newman et al., 2022). Which is not diminished by depoliticisation, but only denied (Flinders & Wood, 2014). And some question if increasingly collecting and centralising evidence can lead to better policy at all (Cairney, 2022).

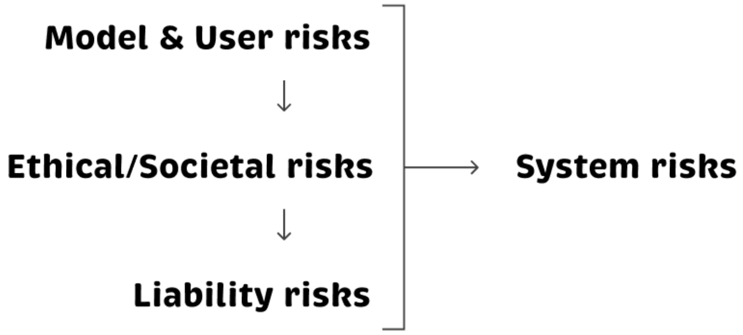
The use of AI for policy brings with it a shift of powers, as the use of AI in the policy process implies a change in the procedural policy tools (Bali et al., 2021) and a renegotiation of institutions in the policy chain. For one, the adoption of AI-based systems changes the allocation of discretion in government (van Noordt, 2023), reducing the ability of policy officers to apply complex and context-dependent professional judgements, as decision-making authorities shift, at least in part, from public servants to the AI system (Bullock, 2019; Liu & Dijk, 2022). Even if people make the final call, their decision will be shaped by the advice or output of the system or its influence on the process, for example, by changing how a subject is framed.

AI systems are not neutral technical systems that deal in facts based on raw data, for they are human-made (Lampo et al., 2018; Winner, 1980). Even data is never truly neutral (Desouza & Jacob, 2017). Someone has decided to collect certain data, and to do so, that person has made decisions on what to include and what not, or, for example, because a certain group has more access to the technology used for data collection and gets overrepresented (Rampton, 2014, as cited in Desouza & Jacob, 2017). Meaning data can be distorted and prejudiced. Some argue that the selection of evidence is never fully neutral and that data can be used selectively to steer policy in a particular direction. In the development of AI models, decisions reflect the aims and assumptions of those who engineer them, decisions about he (training) data that is used, the structure of a model, its fine-tuning, and so on. This risks embedding and codifying the interests of those in power within the system, reducing the potential for renegotiation of these institutions through democratic debate . This is aggrieved by the government's limited ability to alter or replace systems, leading to path dependence, risking further depoliticisation .

The shift of powers extends to the balance between the separated powers of government. Creating risks to the rule of law. At the same time, the use of AI in policy formulation introduces questions about transparency, shared responsibility, and the risk of presenting political choices as neutral outputs. Over-reliance on automated suggestions may limit space for political judgment (Valle-Cruz et al., 2020). Another risk to the power dynamics of democratic policy-making is the potential for increasing reliance of governments on a small pool of large tech companies capable of developing advanced AI models (Gray Widder et al., 2023), which risks leaving less room for democratic governance over said large tech companies.

Importantly, the risks of AI for policy to the functioning of the democratic system go beyond these more theoretical risks related to the democratic ethos and procedural legitimacy. Compounded, the risks described before can interfere with the affective dimension of democracy, which is crucial to the functioning of the democratic polity and society. Potentially compromising the feeling of democratic legitimacy and leading to a decline in trust in government (Levi & Stoker, 2000), resulting from other AI-related risks or from the perception of such risks (Brown et al., 2019).

In conclusion, the use of AI for policy brings with it a variety of risks. To understand the relationship between the different types of risks of AI for policy, this thesis proposes the *Model of the layers of risks in AI for policy*, see Figure 7. The model highlights how the different types of risks build on other risks, specific to the context of AI for policy. Highlighting how higher-order risks are related to, include or build on lower-order risks. The layers are not too dissimilar to an approach using micro, meso and macro levels, like Sætra (2023) model of AI dangers. But separates the ethical and societal risks from liability risks in what is effectively a meso-level. The model includes the following types of risks:



The risks involved in AI for policy can be categorised in four types: *Model and user risks*, *Ethical and societal risk*s, *Liability risks*, and *System risks.* Their relationship in modelled in figure x. With *Ethical and Societal risk*s at the centre, upstream *Model and user risks,* and downstream *Liability risks.* And *System risk* as resulting meta risk.

- *‘Technical and data risks’* result from the technical makeup of AI models and the data that is used. Including, for example, risks like bias, lack of robustness, and breaches of data security.
- *‘Human-AI interaction risks’* arise from the interaction of people with AI systems. For example, users may not fully understand a system’s capabilities or limitations and may rely on it too much.
- *‘Ethical risks’* are risks to ethical values. For example, privacy violations, unfair treatment, and discrimination. Misuse and malicious manipulation add another layer of concern with a large ethical dimension.
- *‘Societal risks’* are risks of the negative impact of AI for policy and the resulting policy on society. For instance, from an unsuitable or discriminatory policy. And environmental risks, as a result of AI systems' significant energy and resource uptake.
- *‘Legal risks’* arise as AI systems cannot be held accountable; legal responsibility remains with their users or the institutions that deploy them. For example, risks of legal liability for breaches of privacy, biased or discriminatory outcomes, and copyright issues.
- *‘Political risks’* are risks of political consequences for political leadership or institutions when they are held liable for harm or the perception thereof, resulting from AI for policy.
- *‘Organisational and institutional risks’* are risks to the functioning of the organisation and institutions, for instance, the gradual erosion of skills and knowledge due to reliance on AI. Or risks like cost overruns of development projects.
- *‘Democratic risks’* are risks to the ideals and functioning of the democratic polity. Risking depoliticising policy development. And resulting, risks to trust in government and the affective dimension of democratic legitimacy.

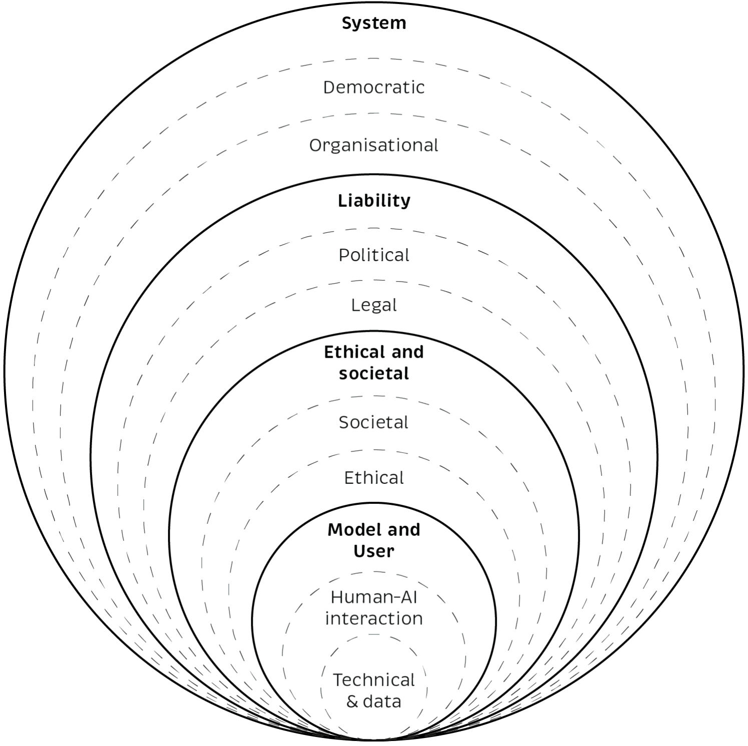


Figure 5. Model of the layers of risks in AI for policy

Risks relevant to policy officers

Looking at the consequential risks resulting from AI for policy, relevant to actors in the policy chain, three distinct types of risk can be identified. Typology of consequential risk specific to AI for policy:

- *Risks to policy quality and society* are risks of delivering poor quality policy, and the resulting negative effects on society or individuals. Also, including risks related to AI for policy application in development and deployment, such as risks to privacy when personal data is used. Resulting from risks related to the model, data and AI human interaction.
- *Risk to government* - Risk of liability, of agents and institution(s), for (perceived) delivery of bad policy. This includes legal and political risk, reduced trust in government (the institution)

- *Risk to democracy and rule of law* – Higher order risk, to democratic ethos and rule of law, through a deterioration of the policy creation system, including the policy chain, national governance system and democratic polity at large.

Of these, *risk to policy quality and society* and *risk to government* are directly and especially relevant to policy officers. As they connect to key considerations of public servants and policy officers . The *Risk to democracy and the rule of law* is an important consideration; however, it is conceptually further removed from the experiential world of policy officers and their practices.

The risks of AI for policy can occur as multiple types of risks at the same time. For example, a biased system is a technological risk, can be a legal risk when it causes harm under the law, and a political risk if it can lead to public dissatisfaction. Other risks only emerge when they combine, for instance, when a technical flaw is reinforced by human overreliance, resulting in the adoption of an incorrect AI outcome.

2.2.5 Value Tensions and Dilemmas

A focus on risks of AI might give the impression that the responsible thing would be to just design in de benefits and remove or neutralise the risks. However, this can only be done to a degree. The benefits and risks of using AI are often connected as balancing forces, resulting in value tensions. In the development and implementation of AI, dilemmas have to be resolved by means of trade-offs. The importance of these tensions and trade-offs is underscored by the development of frameworks and tools facilitating their identification and evaluation (Saxena et al., 2021; Yurrita et al., 2022).

Among these, the circular model of competing values of Yurrita et al. (see **Error! Reference source not found.**) is a useful conceptualisation of tensions between values. And the reality that you can't optimise for all values at the same time. The model places the values on a circular continuum, with similar values, like *Human agency* and *Human control,* next to each other, with competing values like *Fairness* and *Respect for public interest* on the opposing side of the circle.

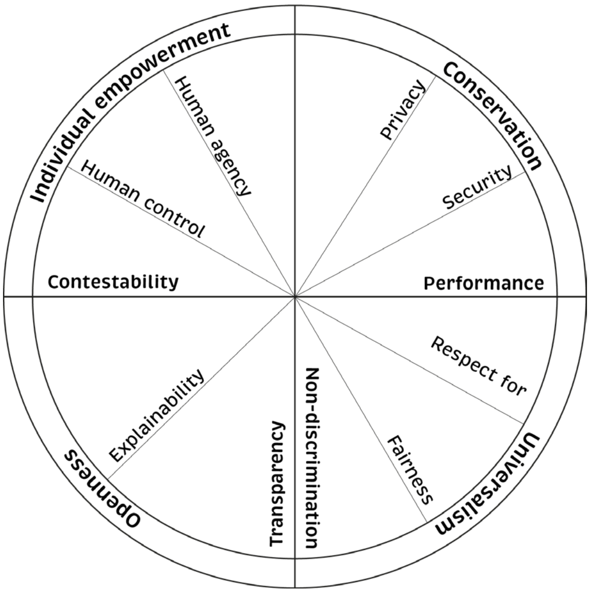


Figure 8. Circular value-based assessment framework. Adapted from Yurrita et al., (2022)

Madan & Ashok (2023) present a categorisation of key conceptual tensions and governance themes that emerge in the deployment of AI systems. Framing the societal and ethical concerns relevant to the design and regulation of AI:

- *Automation vs. Augmentation,* the tension between the concerns around job displacement and the potential to enhance human decision-making;

- *Nudging vs. Autonomy*, the dilemma between using AI for behavioural steering by the state and respecting individual freedom;

- *Data Accessibility vs. Security and Privacy* - Balancing the potential of open data use with privacy risks and informed consent;

- *Predictive Accuracy vs. Discrimination, Biases, and Citizen Rights* - Tensions between optimising AI model performance and protecting citizens from built-in bias and digital inequality;

- *Predictive Accuracy vs. Transparency and Accountability vs gaming the system*, the challenge of being transparent about how systems work, without enabling misuse or undermining their function.

Concluding, benefits and risks are often interconnected, and at times even represent two sides of the same phenomenon. In some cases, risks can be mitigated without compromising the corresponding benefits. More often, trade-offs must be made, sometimes between ethical values and practical considerations, and between competing ethical values, as well as among competing practical considerations. Awareness and acknowledgement of these tensions is a crucial step toward making well-considered and responsible trade-off decisions.

2.2.6 Barriers of AI for Policy

The reluctant adoption of AI for policy and AI in government in general fits with a recurring pattern in the adoption of technology in government observation in literature, that the availability of novel technologies does not directly translate to their adoption in governmental organisations (Madan & Ashok, 2023; Neumann et al., 2022; Selten & Klievink, 2024). Slow innovation in government creates the risk of opportunity costs, which has been raised as a concern in relation to AI's public value potential (Floridi et al., 2018).Zuiderwijk et al. (2021) categorise these adoption challenges into data challenges, organisational and managerial challenges, skills challenges, interpretation challenges, ethical and legitimacy challenges, political, legal, and policy challenges, social and societal challenges, and economic challenges, in part relating to the risks involved in AI for policy. Some barriers are relatively consistent across settings, while others are more technology- or context-specific (Neumann et al., 2022).

A recurring finding is that adoption is closely tied to AI capability. Van Noordt & Tangi (2023) stress the lack of AI capabilities as a major factor limiting AI adoption in government. Here, AI capabilities refer to both technical and non-technical capacities needed to initiate and implement AI projects that create public value. Barriers such as lack of literacy, skills, organisational alignment, or AI-specific infrastructure directly impact whether AI systems can be adopted, and more importantly, whether they can be adopted responsibly. Madan & Ashok (2023) similarly, emphasise the importance of internal demand ("pull") over external technological push in driving early-stage adoption. When organisations do not perceive or frame the relevance of AI for their own operational context, adoption stalls. Overcoming adoption challenges and building responsible AI capacity are not separate efforts, but part of the same (design) challenge. An important factor in the AI capacity of a governmental organisation is the AI literacy and skills of the staff. This AI-educated workforce is often not present (Sienkiewicz-Małyjurek, 2023; Wirtz et al., 2019). Although this is impart due to a general lack of AI-trained personnel in the labour market. The need for capacity is also empirically supported by (Selten & Klievink, 2024). The concepts of AI capabilities, AI capacity and AI literacy are discussed in more detail in the next section on the background on responsible AI.

Others point to the recognition of AI's potential is itself a key step in the innovation process. The willingness to innovate with AI depends on both individual initiative and whether the organisational culture creates space for it (Kamal, 2006). Recent work has looked at the role attitudes, framing, and institutional sensemaking play in AI adoption. Madan & Ashok (2024) explore how public administrators interpret the mix of positive and negative signals they receive about AI. Suggesting that attitudes developed through institutional sensemaking have a significant impact on decisions on AI use. For instance, through internalised bounds on what is perceived as possible and appropriate within a governmental organisation. These boundaries can limit actors in their ability to reframe AI in ways relevant to their work. This helps explain why initiatives like AI strategies or pilot programmes often struggle to take root. Additionally, van Noordt (2023) notes that many government AI strategies focus too narrowly on data-related barriers. Neglecting the broader organisational, institutional, and human factors that influence adoption.

In conclusion, the barriers to implementing AI for policy go beyond the risks involved. They include cultural, organisational, and institutional constraints that shape the very conditions under which adoption becomes possible. A design outcome that aims to support the responsible adoption of AI for policy will therefore need to directly engage with these constraints. This means focusing on organisational and non-technical human elements of AI capacity, including AI literacy, institutional awareness, and the ability to recognise and act on relevant opportunities in the policy development process.

As risk avoidance is one of the main drivers of adoption hesitance, the limiting of risk, through insight into the risks and a practical translation of said risks into actionable. This both reduces the actual risks involved in the innovation and adoption of the novel technologies and, just as importantly, gives innovators, managers and political leadership a feeling of control (over the risks). Making the existence of guidelines and practical tools that support this is an important requirement for the organisational willingness to innovate.

2.2.7 Conclusion: A Balanced Consideration of AI for Policy

AI for policy, defined as "the use of AI in the development of policy", offers a variety of use cases in the phases of the policy cycle, based on complex, generative, and narrow AI technologies. These can provide policy quality, democratic and efficiency benefits. But also comes with a variety of risks, including risks to policy quality and society, risks to government and risks to democracy and rule of law, that are closely related to the experiential world of policy officers. These risks, together with some additional barriers to adoption, like the investment required for AI for policy, result in hesitant adoption of AI for policy. Creating risks of opportunity cost, as underutilisation of AI for policy can result in missing out on potential benefits. In some cases, risks can be mitigated to enjoy only the benefits. Still, more often trade-offs must be made, sometimes between ethical values and practical considerations, and sometimes between competing ethical values themselves. Highlighting that responsible AI depends on a balanced consideration of benefits, risks and costs.

The cost factors of responsible AI
To make a balanced consideration of AI for policy, the factors involved need to be translated into an isomorphic dimension. Although cost calculations are not perfect for public value accounting (Moore, 2014). Costs are a useful conceptual isomorphic dimension to express the various factors involved. The resulting cost factors of responsible AI for policy can be found in Figure 7.

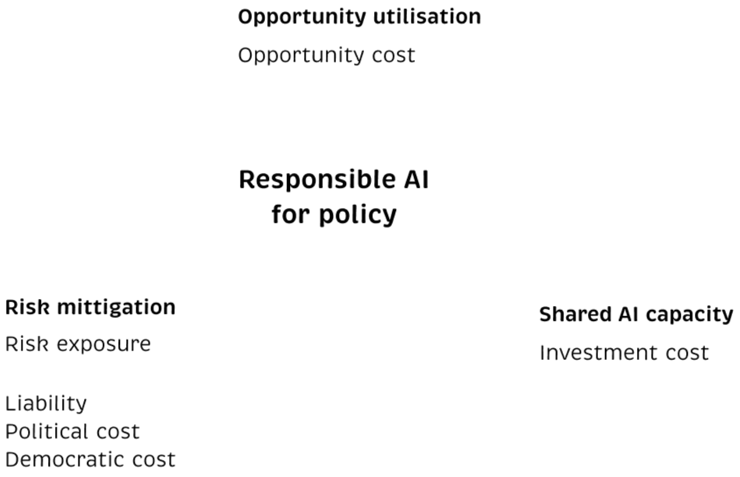


Figure 6. The cost factors of responsible AI

Decision Rule Model

To understand the consideration of responsible AI for policy more structurally, it can be constructed as a decision rule based on cost factors, see Figure 8. Now, of course, this does not make it easy to decide if the use of AI is responsible. It leaves many factors that need to be quantified. This is quite doable for costs like financial investment in resources or FTE. But valuing risks is more subjective. Furthermore, some risk that they are. On the other side, negating existential threats to the nation could be valued to infinity as well.

The benefit of modelling responsibility in this way is that it gives a balanced and relatively complete overview of the meta-level considerations that should go into a decision on the use of AI in government, to come to responsible use of AI.

Although the model is expressed in terms of cost as a metric, the aim of this model is not to quantify the productive performance of an AI application in a neoliberal liberal sense, following Moore (2014). Rather, attempting to capture the dynamics involved in decision-making about the use of AI in AI for policy descriptively. And to allow for the explication of these sometimes intangible costs in decision-making about AI. Helping to fairly weigh the values of affirming or denying the implementation of an AI application.

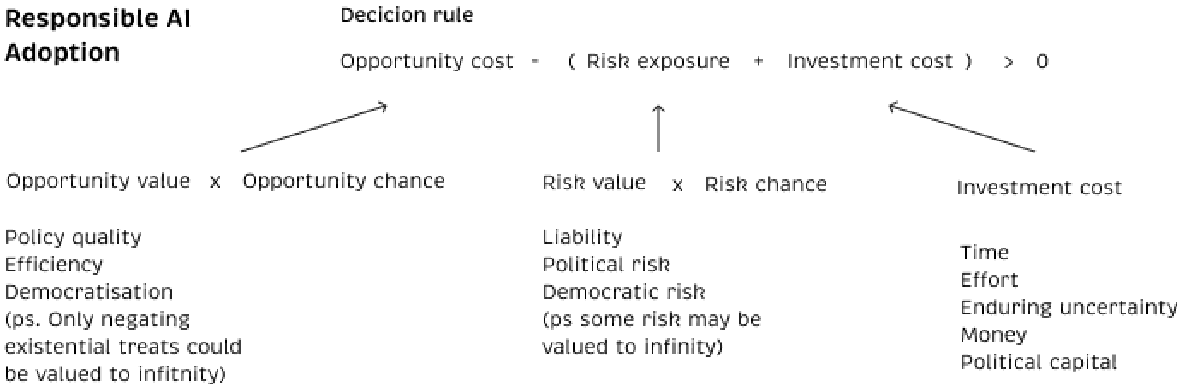


Figure 7. Decision rule for responsible AI: expressed in cost factors

2.3 Responsible AI

The term *responsible* has a threefold meaning, namely: *duty*, "having the duty of taking care of something"; *good judgment*, "having good judgment and the ability to act correctly and make decisions on your own"; and *blame*, "being the cause of a particular action or situation, esp. a harmful or unpleasant one" (Cambridge University Press, n.d.b).
The *duty* layer of responsible AI has gotten much attention through a focus of literature on ethical AI, or the ethical principles that should guide AI design, implementation and use. The methods in these discourses mostly involve technical solutions (Birhane, 2021), ensuring that the models have *good judgment*. Others explore responsible AI through the lens of public value creation, which also focuses on the duty layer. The *blame* layer of responsible AI is explored in discussions on governance of AI, accountability and liability. Another aspect of responsible AI is the attribution (moral) responsibility, focused on who can be and is made to be responsible, and how they can be held responsible (Sattlegger et al., 2022).

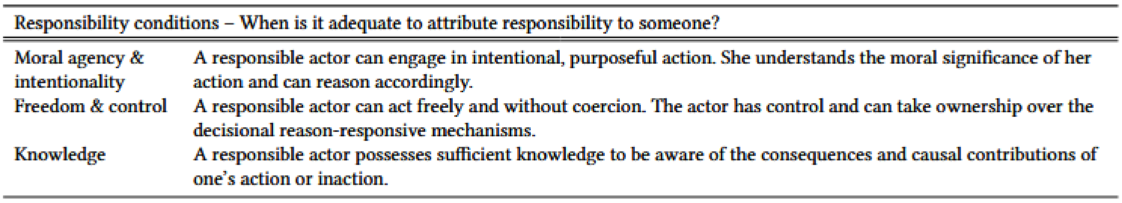
Responsible AI is, importantly, a means to trustworthy governance with AI, to counteract the risks of a decline of trust in government . Therefore, trustworthiness provides a strong basis for the requirements of responsible AI for policy. Trustworthiness is made up of two elements. "The first involves a commitment to act in the interests of the truster because of moral values that emphasise promise keeping, caring about the truster, incentive compatibility, or some combination of all three. When we call someone trustworthy, we often mean only this commitment, but there is, in fact, a second dimension, namely competence in the domain over which trust is being given. The trustworthy will not betray the trust as a consequence of either bad faith or ineptitude." (Levi & Stoker, 2000, p. 476). The first of these dimensions is covered in the discourses mentioned before. The second dimension of *competence*, which is related to part of the good judgement layer of responsibility, is reflected in the discourses on critical approaches centred around problematisation and contestation. And in discussion on AI Capabilities and AI Capacity, including AI literacy, as organisational prerequisites for responsible AI.

This section proceeds as follows: First, the relation between responsibility, accountability and liability is explored more deeply at a theoretical level, to frame the discourses of responsible AI approaches and methods. After, a variety of dominant and critical approaches to responsible AI are discussed. Ending with a discussion of organisational AI capabilities, including AI literacy, that support responsible AI.

2.3.1 Responsibility, Accountability and Liability

Moral Responsibility
The duty layer of responsibility can be viewed from the perspective of attribution, looking at who is responsible. The attribution of moral responsibility forms the basis for effective or operational distribution of responsibility through mechanisms like accountability and liability. Literature on moral responsibility is concerned with the degree to which an individual or organisation is responsible. Moral responsibility conditions (see figure x), provide a basis for understanding if an agent is in a position to take on the responsibility of developing or using an AI system (Sattlegger et al., 2022).

Looking at task responsibility (Sattlegger et al., 2022) a framework for designing for responsibility. This approach applies to the whole lifecycle of an AI System. Focusing not only on the system itself but also on its implementation as a social-technical system. Situated in the context of political institutions.



Responsibility conditions – When is it adequate to attribute responsibility to someone?	
Moral agency & intentionality	A responsible actor can engage in intentional, purposeful action. She understands the moral significance of her action and can reason accordingly.
Freedom & control	A responsible actor can act freely and without coercion. The actor has control and can take ownership over the decisional reason-responsive mechanisms.
Knowledge	A responsible actor possesses sufficient knowledge to be aware of the consequences and causal contributions of one's action or inaction.

Responsibility condition (Sattlegger et al., 2022, p. 218)

Mechanism of Attributed Responsibility: Accountability & Liability

Effectuating responsible AI by means of the attribution of responsibility requires moving from responsibility (normative), through accountability (normative/institutional), to liability (legal). In which accountability is related to the good judgment layer of responsibility, and liability with the blame layer, when extending the layers of responsibility outside the normative dimension.

Accountability has a double meaning; it can be conceptualised as a virtue and as a mechanism (Bovens, 2010). Accountability as a *virtue (Normative),* meaning to be inherently open to questioning. It is mainly used as an adjective with a positive valency. Accountability as a *mechanism (Institutional)* consists of an institution between an actor and a forum, usually hierarchical, in which the forum can question the actor. Making that accountability can support transparency and procedural legitimacy. Furthermore, accountability mechanisms ensure adherence to the virtuous of accountability (Bovens, 2010), making agents more likely to act responsibly. In the case of elected officials, political accountability.

Liability plays an important role in accountability as a mechanism, resulting from the judgment of an accountability forum when "the actor may face consequences" (Bovens, 2007, p. 450). Although this could be seen as an integral part of accountability as a mechanism (Bovens, 2010), it is a distinct concept. The emotional valency with liability in common parlance is negative, as the term is usually used in the context of an actor being liable for damages or fines (Merriam-Webster, n.d.). However, the conceptual mechanism of liability is itself neutral. If the judgment of an accountability forum is positive, the consequences faced by the liable actor may be as well. This nuance is important so as not to load accountability and liability overly negatively. And avoid undue resistance from the actors involved, like model developers, policy officers, managers and leadership .

An example of liability in the democratic political context is political actors like politicians and political parties, who can face democratic consequences in elections. And the political system that can be judged as a whole, being held accountable for its trustworthiness through a trust judgment by the public (Levi & Stoker, 2000), potentially enduring consequences of reduced cooperative participation , or populist disruption of the democratic system (Inglehart & Norris, 2016; Schmidt, 2017), and ultimately revolution (Arendt, 1963/1990).

2.3.2 Dominant Approaches to Responsible AI

The developments in AI for policy and its adoption are an important junction in the evolution of the policy development process, one that requires careful composition and continuous, iterative correction (Coeckelbergh & Sætra, 2023). The theoretical discussions on responsible AI form an important conceptual basis for responsible AI practice in government. However, they often remain far removed from the practical and functional approaches needed in real-world settings (Hagendorff, 2020). Approaches and methods to responsible AI aim to bridge this gap. The approaches to responsible AI described in the literature differ in their level of abstraction and in the perspective they take on what it means to act responsibly. This subsection gives an overview of key concepts and approaches to responsible AI that are relevant to the adoption and use of AI for policy.

Approaches Based on Technical Solutions

Technical solutions focus on the improvement of AI systems and their outcomes through better system design and high-quality training data. The proper technical functioning of a model is, in a real sense, at the centre of responsible AI. Much attention in the scholarship on responsible AI has been given to responsibility at the algorithm or AI model level (at the level of the model), of the application level (the use of the model). The technical approaches to responsible AI are most often guided by ethical principles. Conversely, literature generally focuses on ethical issues that could be solved by technical means (Prem, 2023), as technical solutions would be a one-stop shop approach to responsible AI. One approach, the FEAS framework by Toreini et al. (2020), links trustworthiness of AI systems to the qualities of fairness, explainability, auditability, and safety, exploring what technologies can improve these characteristics. However, these technical solutions have not always proven to be effective (Gonen & Goldberg, 2019), highlighting that there are limits to the responsibility that can be achieved through technical solutions. Supporting the notion that the responsibility and trustworthiness of AI go beyond technical performance, and should include considerations on human context (König & Wenzelburger, 2020).

Approaches Based on Ethical Principles

A large fraction of the discussions on responsible AI are focused on ethical values and principles. Many of these principles are derived from the ethics discourses in the medical field, which has a long history of developing ethical principles, known as 'principlism'(Prem, 2023). A key example of an ethical principles approach is the AI4People framework developed by Floridi et al. (2018), which builds on the principles of: beneficence, non-maleficence, autonomy, justice and explicability, derived from bioethics. This framework has strongly influenced the European Commission's guidelines on Trustworthy AI High Level Expert Group on Artificial Intelligence (HLEGAI, 2019), and later the OECD's guidelines (2024), according to Floridi & Cowls (2019). Resulting in these ethical values becoming a focus central to many approaches to responsible AI (Prem, 2023).

Some authors reject the idea that ethical principles should function as the foundation for responsible AI. Arguing that ethical discussions mostly remain in academia and that the values discussed are inherently conflicting or too high-level and abstract to inform practical, technical, or organisational actions (Munn, 2023), which is supported by the fact that technical communities tend to see them as peripheral additions to technical consideration (Hagendorff, 2020). Arguing for problematisation that is less smooth, with more inherent focus on the complexity and situated nature of responsibility in AI, or AI justice (Birhane, 2021).

Approaches to Trade-offs

Dealing with value tensions and making trade-offs is often unavoidable in the development, implementation and use of AI for policy. Some have proposed responsible AI approaches centred on making these explicit. The ADMAPS framework of public sector algorithmic decision-making by Saxena et al. (2021) puts a focus on dealing with the interdependencies and trade-offs between algorithmic decision-making and bureaucratic processes, and human discretion. Arguing for the use of algorithmic systems as part of a holistic assessment as a means to improve decision-making. Similarly, Yurrita et al. (2022) proposes an approach for assessing algorithmic systems with explicit attention to the competing nature of (ethical) values. Developed for the development of the model application or assessment before deployment. Looking to problematize competing values and negotiate balanced trade-offs together with a wide, pluriform network of stakeholders.

Responsibility Structures

In answering the questions around responsibility in the sense of attribution of responsibility, some have proposed public AI governance frameworks (Wirtz et al., 2020). These approaches to responsible AI focus on the attribution of responsibility around AI systems. These responsibility structures aim to make explicitly clear who is accountable for certain parts of the process, from development to use, and to make sure this accountability is supported by real institutional mechanisms instead of remaining an abstract ethical idea. Building on the notion that ethical guidelines are not enough, risking becoming an empty “checkbox” if they are not backed by actual enforcement (Hagendorff, 2020). And, Mittelstadt et al. (2016) stress that traceability, the ability to follow who made which decision during the design and use of AI, is key for ensuring that people can be held accountable.

A relevant example of this type of approach is the work by Sattlegger & Bharosa (2024), who propose an ethical risk responsibility model based on the three lines of defence approach. Attributing responsibility across the levels: strategic oversight, operational compliance, and independent reflection, making the task responsibility explicit. Arguing that unclear divisions of responsibility, such as who carries out or reviews algorithmic impact assessments, can weaken both moral and political accountability.

2.3.3 Critical Approaches to Responsible AI

Agonistic Problematisation

To reduce the risks of over-reliance on AI systems and to move beyond responsible AI through technical solutionism (Birhane, 2021), there is a need for formal procedures and informal conditions that encourage critical reflection and reintroduce political debate into the policy-making process . Building on Mouffe's (1993)conception of pluriform politicisation as a means to counter the autocratic tendencies of increasing technocracy, one way to do this is by introducing the problematisation of AI for policy throughout its lifecycle.

Problematisation, as conceptualised by Foucault (1997), involves more than pointing out issues; it focuses on creating the conditions for critical responses to emerge by defining the frames and boundaries that shape how problems are interpreted and what kinds of responses can be developed. It engages with the problematic, the underlying conditions and assumptions that determine how issues are understood and acted upon. In doing so, problematisation opens up space for disagreement, debate, and alternative perspectives.

Problematisation can occur at various levels: individually, through reflection (Schön, 1983), between internal stakeholders such as colleagues and internal experts, and externally through engagement with supervisory authorities, oversight bodies, civic organisations, or public participation.

Agonistic democratic mechanisms can be used to create this kind of space for critical engagement. The concept of agonistic democracy is proposed by Mouffe (1999) as an alternative to deliberative models, focusing on pluralism and sustained disagreement as a healthy aspect of democracy. Prioritising support for the development and the expression of competing views over consensus. Mouffe distinguishes between antagonism, where opponents are treated as enemies, and agonism, where they are recognised as legitimate adversaries. The aim is to transform antagonism into productive disagreement that sustains democratic engagement.
Research on agonistic design has shown how institutions can be shaped around principles such as contestation, interdependence, and openness to uncertainty (Lowndes & Paxton, 2018). Including collaborations with civic groups to uncover the political assumptions embedded in data-driven urban systems (Bunders & Varró, 2019). Contestability is central to this approach; it is not merely about allowing disagreement but actively designing for it. These designs are provisional, recognising that no solution is final, and treating ongoing disagreement and change as vital elements of democratic vitality (Lowndes & Paxton, 2018).

Contestation

Contestation by decision subjects is an important element in reducing the risks that come with AI-based decision-making. Research on the depoliticisation of democracies highlights the need to embed direct forms of contestation within representative systems to address the shortcomings of depoliticised governance (Pettit, 2004). The goal of contestation is to strengthen democratic control and support the common good by creating ways for citizens to take part in decisions beyond casting a vote. This helps ensure that a wider range of interests and perspectives are heard and considered.

In the context of AI in government, contestation mainly refers to giving people the opportunity to question or challenge decisions made by algorithmic systems (Alfrink et al., 2022, 2023). While democratic control is a defining characteristic of democracy in theory, exercised by citizens over the state, in practice, this control is limited. Direct forms of democracy are difficult to apply at scale, which leads to a reliance on representative models. These come with their own problems, such as political leaders prioritising re-election or party interests over broader public concerns (Pettit, 2004).

Metaphor of agonistic arena (Alfrink et al., 2024)

2.3.4 Organisational Capacity for Responsible AI

In addition to approaches and methods to responsible AI organisations need to create the necessary conditions to follow up on these, to adapt, extend to their specifics and contest. Building the *competence* dimension that is the basis of the *good judgment* layer of responsibility. These competences include AI capability and AI capacity, and AI literacy.

AI Capability and AI Capacity

Part of the risk of AI implementation comes from poor AI systems, technical and data facilities, and organisational incapability . One of the key factors in the successful and responsible adoption of AI is the presence of sufficient capabilities . Making AI capabilities are needed to enable adoption, but still, more is needed to adopt AI responsibly. Meaning that, in addition to the presence of a specific skill or resource, its scale and the ability to utilise it matter. Moral responsibility and trustworthiness both require the ability to act. This means proper capabilities and capacity have to be present for responsible AI. AI capacities can be separated into: hard capabilities, which consist of the technical capabilities like data, ICT infrastructure and AI models. And soft capabilities, the human and organisational capabilities required to develop, implement and use AI responsibly and effectively, such as organisational processes, human skills, like leadership, collaboration, and AI literacy.

A term that has been used in a similar context is AI capability . Mikalef & Gupta (2021) define AI capability as "the ability of a firm to select, orchestrate, and leverage its AI-specific resources" (p. 4). They propose a categorisation of organisational resources that make up artificial intelligence capability. Tangible resources consist of data, technology, and basic resources. Human resources refers to technical and business skills. Intangible resources include inter-departmental coordination, organisational change capacity, and risk proclivity. This is consistent with van Noordt (2023), who finds a distinction between the capabilities needed to develop AI systems and the capabilities needed for the implementation, while both are needed for the effective implementation of the technologies and delivering value. Mikalef et al. (2022) further explore the importance of the human and intangible AI capability factors, noting that innovation culture within the organisation positively correlates with AI capacity.

However, for the organisational ability to adopt AI, resulting from the combination of hard and soft capabilities, the term AI capacity is better fitted. The term capacity is used for a similar concept in the TOE framework , which Madan & Ashok base their framework for AI adoption. "4.2.1.4. Absorptive capacity. A global theme of absorptive capacity emerged across all the TOE contexts. In the context of AI adoption, absorptive capacity is manifested through a strong path dependency on existing infrastructure developed through previous e-government innovations, collaborations between organisations, and a network of external technical specialists…" (Madan & Ashok, 2023, p. 7).

AI Literacy

A concept that is related to AI capabilities is AI literacy, the ability to understand the technology and contextualise information about an AI system (Ng et al., 2021). This is important for one's ability to use AI responsibly. Ng et al. (2021) describe AI literacy as made up of four connected domains. The first is knowing and understanding AI, which includes basic functions and how to use AI applications. The second is using and applying AI, where the focus is on working with AI knowledge and concepts in different situations. The third domain is evaluating and creating AI, which focuses on higher-order thinking skills such as evaluating, predicting, or designing with AI. Finally, AI ethics addresses human-centred considerations, including fairness, accountability, transparency, and safety. Together, these domains offer a comprehensive framework for understanding what it means to be AI literate in practice.

AI literacy is a requirement for conversations about the responsible use of AI . General AI literacy also helps to facilitate the communicative processes between diverse stakeholders required for negotiating clashing values (Yurrita et al., 2022) as increased AI literacy opens up more possible means for communicating about an AI system. When AI literacy is widespread, the use of AI for policy and the (ethical) questions that arise from it can be discussed between more people and with people closer in the organisation. There are loads of concerns policy officers would not be willing to step towards an expert with, but you would very much be willing to spar about with a colleague at the coffee machine. Lowering the barrier and likely increasing the ability to deal with arising challenges. The spread of a base level of AI literacy, combined with guidance, would allow minor or local challenges to be dealt with throughout the organisation. This makes AI literacy an important form of AI capacity, especially when it comes to responsibly incorporating AI.

2.3.5 Conclusion: Towards collective soft AI capacity for constructive problematisation

A variety of approaches to responsible AI are described in the literature. Including approaches centred on technical solutions, ethical frameworks, responsibility structures and critical, reflective practices like problematisation and contestation.

In designing responsible AI for policy, there is an inherent tension between building responsibility top-down, through formal rules, controls, and paternalistic safeguards, or bottom-up, through internal reflection, problematisation, and open contestation.
Ultimately, the ability to realise responsible AI depends on AI capacity in the organisation. If the needed capacities are not in place, organisations cannot take on responsibility in the sense of Sattlegger et al. (2022). The development of *soft AI capabilities*, such as *AI literacy* and reflexive competence, is crucial for creating the conditions in which responsibility can be enacted. These capabilities enable stakeholders to communicate across disciplines, question assumptions, and challenge the design, implementation, and use of AI systems. The policy quality governance system has only limited influence on technical or “hard” capabilities, but it can play a key role in the development and utilisation of soft capacities, through guidance, support and education. The question of how to achieve widespread AI literacy remains open; this thesis and its design proposal suggest AI curiosity as a possible mechanism.

Building on this synthesis, the proposed *Model of Responsible AI Use* (see **Error! Reference source not found.**) is a sketch of a framework for a balanced critical approach to responsible AI for policy, placing awareness and problematisation at its centre. Relating the opportunities and the risks in responsible AI for policy, the need to mitigate or limit risk as much as possible while making trade-offs, by means of the requirements and mechanisms of awareness and problematisation centre. It recognises that addressing the dilemmas surrounding AI for policy is about creating the structures and cultures that allow for informed, ethical, and adaptive responses.

Model of Responsible AI Centred on Awareness and Problematisation

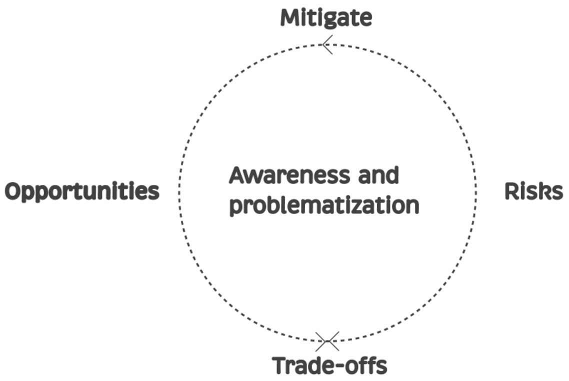


Figure 8. Model of responsible AI use with awareness and problematisation at it centre

2.4 Curiosity

Curiosity is the eager wish to learn about something (Cambridge University Press, n.d.a). While external circumstantial forces might motivate people to seek information that might help them deal with the challenges at hand. Pressure stands in the way of true curiosity, because as Kashdan & Silvia (2009) state: "When we are curious, we are doing things for their own sake, and we are not being controlled by internal or external pressures concerning what we should or should not do."(p. 368). Similarity, Golman & Loewenstein (2018) argue that anxiety is an antagonist of curiosity in the drive for information, leading to information avoidance rather than acquisition.

Empirical research by Kang et al. (2009) shows the relation between the level of knowledge about a subject and curiosity is related (Figure 13). Curiosity increases when people get to know a bit about a subject, before decreasing when they get to know a lot about the topic. Who conclude, "The fact that curiosity increases with uncertainty (up to a point) suggests that a small amount of knowledge can pique curiosity and prime the hunger for knowledge"(p. 972).

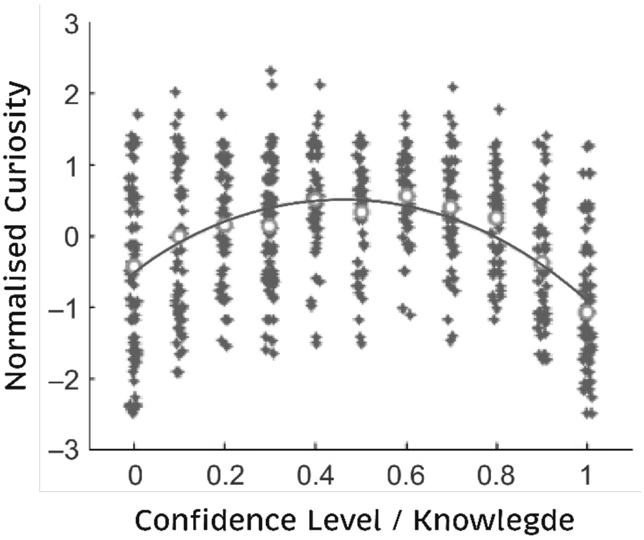


Figure 9. Distribution of curiosity ratings as a function of confidence. Adapted from (Kang et al., 2009).

They also demonstrate that curiosity has an impact on neurological activation in a variety of distinct brain regions, as well as resulting in a physiological effect in pupil dilation. This supports that curiosity impacts neurological circuits related to motivation involved in valuing and anticipating primary rewards, like those from food and sex, which are activated by curiosity for information, showing that novel information is valued similarly (Oudeyer et al., 2016). Pointing to underlying neural mechanisms that explain why the level of curiosity corresponds with increased memory and learning, as demonstrated by Kang et al. (2009).

In conclusion, curiosity has the potential the make people proactively seek information and gather knowledge. A small amount of knowledge can engender further exploration, seeking more. With that, it can potentially act as a mechanism to enable the development of AI literacy and soft AI capacities. However, pressure or anxiety hinders the development of curiosity.

3. Design context

This chapter discusses the context of the design project. Including the Dutch democratic system, the governmental organisation and the wider related field. With extra attention for the direct context of the design challenge, consisting of the policy quality system and the Policy Compass. Furthermore, the context of me as a designer within the organisation is explained.
The description of the design context builds on the conceptual framing from the background and is based on desk research and the insights from early design activities, like interviews and observations. The method and design approach are further discussed in Chapter 4.

3.1 Policy Development in Dutch Democracy

The system that creates policy is important in the context of responsible AI for policy. This system is explored with a broad introduction to the Dutch democratic polity. After, the chapter zooms in towards the ministerial sectors responsible for policy quality, the direct context of the design project.

3.1.1 From Vote to Policy and Law

In a democracy, the process of policy development starts with a vote by the people[2]. In the Netherlands, politicians for the House of Representatives [*Tweede Kamer*] are elected through a system of direct national proportional representation, using open party lists (Jacobs, 2018). The Senate [*Eerste Kamer*] is elected indirectly through the provincial assembly elections. The government is granted democratic legitimacy by the parliament [Staten-General], which scrutinises the government through the right to information, co-legislation, voting on laws, and can ultimately send home the government by means of a motion of no-confidence.

3.1.1.1 The policy cycle

The main organisational system in the design environment is the ministerial policy cycle (van der Staaij & Sneller, 2023). The policy cycle, as used in the Dutch national government, consists of the following phases:

1. Agenda setting is the phase where the political (and departmental) priorities for policy are defined.

2. Policy preparation is the phase where policy officers explore the problem, the intended goal, the possible policy options, including the type of policy instrument, and their consequences are explored. The Policy Compass is one of the tools used in this phase to support well-reasoned choices.

3. Decision-making is the phase where a decision is made about the developed policy, either explicitly or implicitly (by not continuing progression). In the case of Laws, this includes parliamentary approval, otherwise through implementation of the policy by the government, as well as informing parliament.

4. Implementation and execution, if an affirmative decision is made, the policy enters the phase where it is put into practice. The novel policy is also monitored here.

[2] If not practically, in spirit and by democratic legitimation.

5. Evaluation is the phase where the implemented policy is analysed, looking at the real effect and impact of the policy.

6. Reorientation, based on the evaluation, a decision is made to continue, adjust, or stop the policy altogether.

3.1.2 Policy Governance System

The policy governance system is responsible for setting rules and procedures for policy development, ensuring that policy is legally sound, consistent, and well-aligned.

3.1.3 Policy Quality

Policy officers are tasked with preparing a high-quality policy proposal through a rigorous process. However, in practice, policy development often takes place under pressure resulting from limited time and capacity, as well as political expectations on the content and form of policy. Making it difficult to carry out a careful and considerate process, reducing the quality of the process and policy proposal.

To support the quality of the policy development process, the use of the Policy Compass has been made a requirement (Ministerie van Financiën, 2021). This tool helps structure the process and ensures that core tenets of good policy-making are used in the development of each policy proposal. On the one hand, it helps standardise the process from the top down. On the other hand, it can be used by policy officers to push back against pressure when there isn't enough time or space for proper preparation . In this way, it can support policy quality from the bottom up. The role of the policy compass is further discussed in the following subsection. The Policy Compass will be explored further in the following section.

Additionally, internal quality checks are an important part of ensuring policy quality. The legal quality review [Toets op wetgevingskwaliteit] is used to test whether proposals meet priority requirements (Dekker, 2021). These include: Human scale [menselijke maat], does the policy leave room to take individual situations into account? Doability [doenvermogen], is the policy realistic for people to follow or understand? Feasibility [uitvoerbaarheid], has the proposal been developed together with the implementing organisations, and can they carry it out? These checks help make sure that proposals are not only legally sound but workable and socially aware as well.

3.1.4 The Policy Compass

The Policy Compass is the tool that supports policy officers officer structure their policy development process, as well as helping to reflect on key steps to ensure an improvement of the policy quality. Promoting good practices across the government and making policy development more consistent. The design proposed in the design project of this thesis is intended to explore how the topics of AI can find a place in, or be related to, the Policy Compass. That means the design needs to align with how the Policy Compass works, how it is used, and how it is positioned within the policy quality system. Making it a key reference artefact in the direct context of the design challenge.

The Policy Compass is developed as a replacement for the older Integrated Framework for Policy and Regulation Assessment (IAK) (Haag, 2010). Following renewed attention for the IAK and policy quality after the child benefits scandal, and OECD recommendations to make better use of the IAK, especially to check if policies are actually feasible and executable in practice . With the Policy Compass, the government aims to improve the usability and uptake of the IAK, by updating the content, reorganising the structure, and offering a more user-friendly website and support tool (Kamerstuk 35925-VI-124). Hereto, the Policy Compass is continually updated and improved by several groups that work together. See the formal organisational structure of the Policy Compass in Figure 16.

The Policy Compass is available in different formats. There's an interactive website version, see Figure 13, and a form version, which can be used depending on how and when it is used in the policy process.

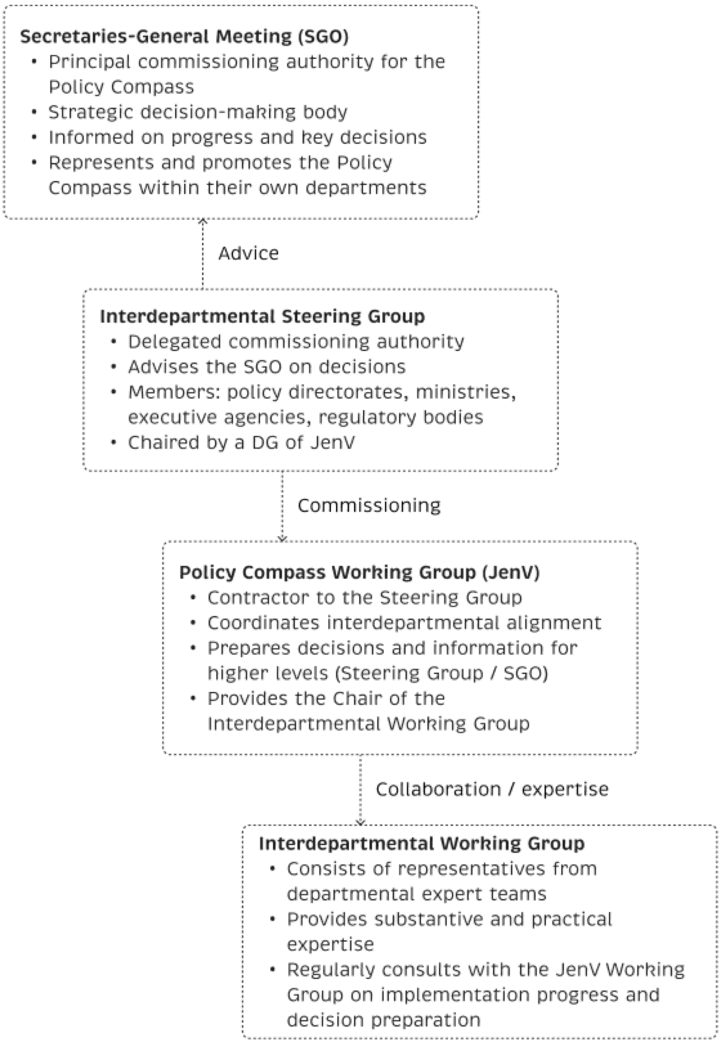

Secretaries-General Meeting (SGO)
• Principal commissioning authority for the Policy Compass
• Strategic decision-making body
• Informed on progress and key decisions
• Represents and promotes the Policy Compass within their own departments
Advice
Interdepartmental Steering Group
• Delegated commissioning authority
• Advises the SGO on decisions
• Members: policy directorates, ministries, executive agencies, regulatory bodies
• Chaired by a DG of JenV
Commissioning
Policy Compass Working Group (JenV)
• Contractor to the Steering Group
• Coordinates interdepartmental alignment
• Prepares decisions and information for higher levels (Steering Group / SGO)
• Provides the Chair of the Interdepartmental Working Group
Collaboration / expertise
Interdepartmental Working Group
• Consists of representatives from departmental expert teams
• Provides substantive and practical expertise
• Regularly consults with the JenV Working Group on implementation progress and decision preparation

Figure 10. Formal organisational structure of the policy Compass, based on Koppenjan et al. (2024, p. 22)

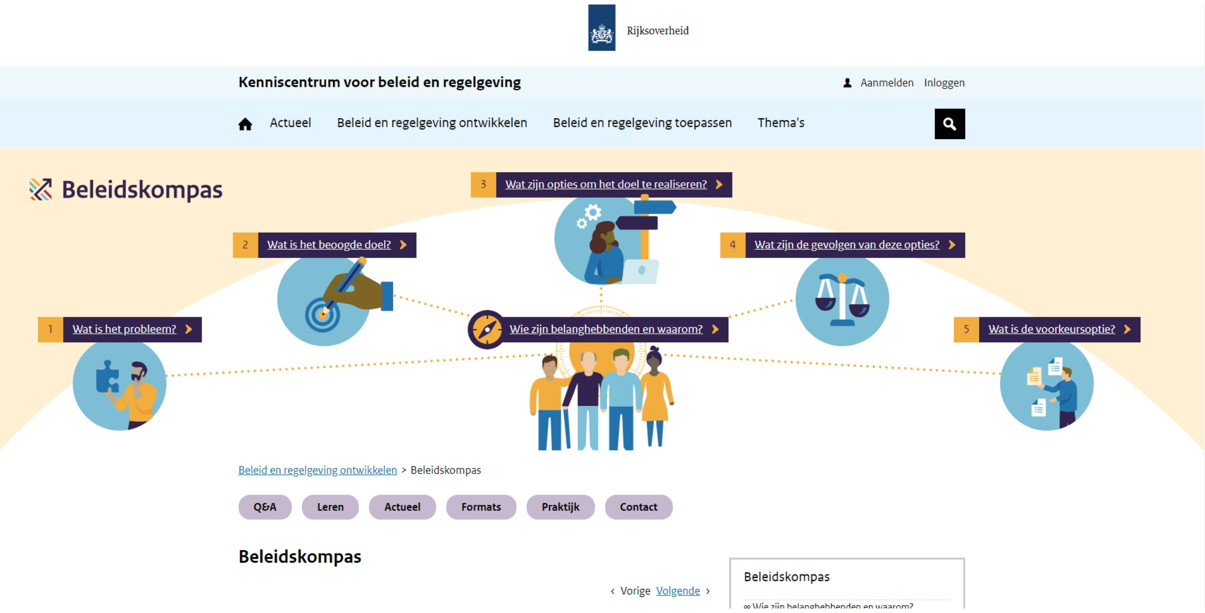
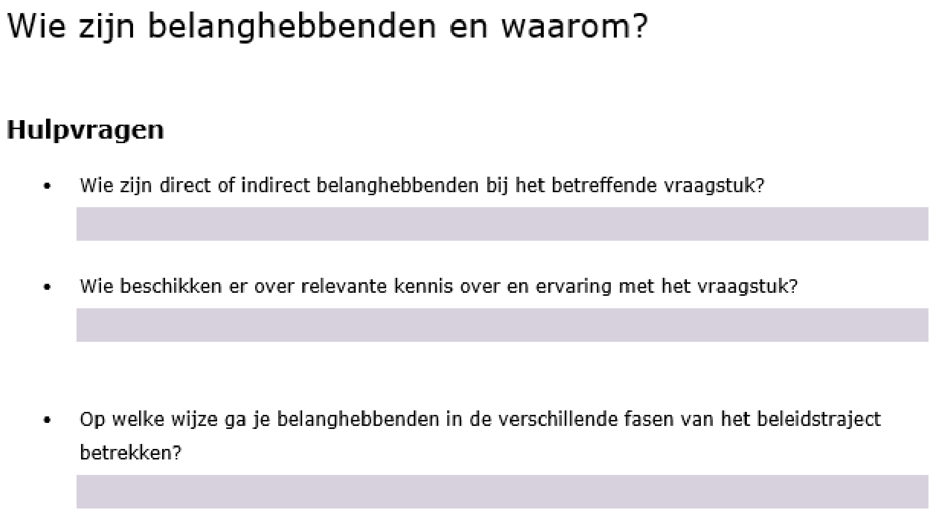

Rijksoverheid
Kenniscentrum voor beleid en regelgeving
Aanmelden
Inloggen
Actueel
Beleid en regelgeving ontwikkelen
Beleid en regelgeving toepassen
Thema's
Beleidskompas
3 Wat zijn opties om het doel te realiseren?
2 Wat is het beoogde doel?
4 Wat zijn de gevolgen van deze opties?
1 Wat is het probleem?
Wie zijn belanghebbenden en waarom?
5 Wat is de voorkeursoptie?
Beleid en regelgeving ontwikkelen > Beleidskompas
Q&A
Leren
Actueel
Formats
Praktijk
Contact
Beleidskompas
‹ Vorige Volgende ›
Beleidskompas
Wie zijn belanghebbenden en waarom?

Figure 11. The Policy Compass (Beleidskompas | Kenniscentrum voor Beleid en Regelgeving, n.d.)

The main mechanism the Policy Compass design uses is *guiding questions*, supporting policy officers to reflect on the steps they need to take in their policy development process. Figure 14 shows a section of the Dutch *Policy Compass*, containing guiding questions designed to identify and engage stakeholders during the policy process.



Wie zijn belanghebbenden en waarom?

Hulpvragen

- Wie zijn direct of indirect belanghebbenden bij het betreffende vraagstuk?
- Wie beschikken er over relevante kennis over en ervaring met het vraagstuk?
- Op welke wijze ga je belanghebbenden in de verschillende fasen van het beleidstraject betrekken?

Figure 12. Section of the form version of the Policy Compass (Formats | Kenniscentrum voor Beleid en Regelgeving, n.d.)

Conclusion: Preserve Procedural Legitimacy and Enhance Policy Quality

The policy process is supported and supervised by a network of governance structures that safeguard legal and policy quality.

Policy quality can come under pressure as the development often happens under pressure, for instance, from political expectations. The Policy Compass is devised to give guidance to policy officers through reflection, structure, and quality checks. Supporting good policy development practices.
Preserving procedural legitimacy, therefore, depends on maintaining the integrity of these institutional checks and reflective tools while allowing enough space for professional judgment. The Policy Compass embodies this balance, linking formal quality requirements with reflective practice to uphold the democratic ethos and improve the substantive and procedural quality of policy development.

The promotion of AI as a standalone goal does not fit the role of the Policy Compass nor DWJZ, whose role is to support the development to increase policy quality. It should rather be aligned with responsible use for the benefit of policy quality.

3.2 Dutch Government and AI

This section outlines the Dutch government's current approach to AI, with a focus on the context relevant to AI for policy. It includes recommendations made to the government, the government's stated intentions and commitments, the degree to which AI is currently being explored and used, the related government capabilities, and actors involved in the government AI ecosystem.

A proposed design has to align with government policy and its trajectory. Policies and formal commitments (for example, to parliament) are the basis for a justification of the development and implementation of a new approach. At the same time, the proposed design and thesis may recommend something that stretches- or goes against standing policy, which comes with solid argumentation. For instance, by building on prior recommendations made to the government or intentions expressed by the government.

3.2.1 Recommendations, Intentions and Commitments

Important advice given to the government regarding AI it the 2021 report by the Netherlands Scientific Council for Government Policy (*WRR*). In it, the WRR argues that the government needs to stop treating AI only as something that happens "out there" in society and make it an explicit goal in its own work, to learning how to apply AI. "The transition we advocate does not mean that the government should start proclaiming the truth about AI to society. However, it will have to work on learning about AI in its own actions, and therefore make evaluating AI and reflecting on the intentions with AI an integral part of its functioning across the board." (WRR, 2021, p. 405) In its official response to the report (*Kamerbrief met kabinetsreactie op WRR-rapport 'Opgave AI'*), the cabinet acknowledged this recommendation. But the tone was somewhat cautious, and the actions described were limited. While the WRR called for a structural shift in how government relates to AI, the response mostly confirmed support without really showing much urgency.

3.2.2 Government Policy and Applicable Law

Governments' internal AI policy and the use of AI for policy endure far greater scrutiny than the public sector (Desouza et al., 2020). This is reflected in the Dutch context. Policy officers working with AI are bound by the law and government policy. The law sets out the minimum requirements by defining what is legally allowed or required. Whereas policy tends to be more cautious, setting rules to avoid both legal and political risks. For the development of this policy, the government collects input from advisory bodies and organisations in civil society, such as the Dutch Data Protection Authority (AP).

Key legal frameworks in the context include: the General Data Protection Regulation (GDPR), which relates to privacy and data protection; and the EU AI Act, which introduces requirements for AI applications based on the level of risk posed by AI systems. Sometimes existing frameworks in other legal areas apply as well, such as intellectual property, sector-specific regulation, or fundamental rights, under the Dutch Constitution and the European Convention on Human Rights (ECHR).

More detailed guidance on the limits of AI use in the government is provided government's policy. Especially in areas where the interpretation of how the existing legal frameworks apply to novel technology is still developing. For example, regarding the use of generative AI (Rijksoverheid, 2024).

3.2.3 AI for Policy in the Government

Currently, the development and use of AI for policy in the Dutch government is mostly limited to experiments and pilots. In relatively small-scale projects, focus on supporting internal processes, rather than automating steps in policy development. Some ministries experiment with the development and use of generative AI, like internal information retrieval. Additionally, some commercial solutions with limited capabilities are being used on a small scale. These applications are still in early phases and often limited to exploratory environments run by innovation labs or internal working groups.

3.2.4 Barriers to adoption in the Government

In the Dutch context, insights from the interviews and other fieldwork activities in the design project point to barriers like fragmentation of efforts and capabilities within and between ministries, lack of insight into AI opportunities, perceived uncertainty and effort, and challenges in AI literacy and capacity. These capability barriers are in part linked to historical divestments in technical infrastructure or procurement challenges within government. There is also a lack of empirical validation, ethical dilemmas that go beyond legality (IBDS), and an absence of processual or policy cycle perspectives. Mirroring barriers described in the literature. The following are key insights on the barriers to responsible adoption of AI for policy of the government.

Political risk - The additional risk is political risk beyond legal liability. And the risk of setting a legal precedent by being gratuitous to one party once. Even when things are within the legal limits, it can still be a political scandal. This can be seen in the case of the duo fraud, where the political reaction of a full admission of fault does not match the concerns raised by the algorithm audit about the algorithm assumptions not being statistically significant.

Disjointed efforts - In government, there are many groups and projects concerned with the experimentation and regulation of AI (for policy). However, these groups do not work together well, if they even know of each other's existence. And as is common in government, they all have their own priorities and goals. This makes it incredibly difficult to find the right resources and expertise for anyone.

Lack of AI capacities - Although the conditions and capacities are varied between ministries, the lack of capacities needed for the responsible adoption of AI is lacking or underdeveloped. In modern history, the government has outsourced a lot of the technical capacities needed for AI implementation.

3.2.5 Tools for Responsible Data, Algorithms and AI

In and around the Dutch government, several tools and frameworks have been developed that can support the responsible development and implementation of AI, algorithms and data. These tools available to government employees and teams are the following:

- *Impact Assessment for Human Rights and Algorithms (IAMA)*, created for the Ministry of the Interior, this tool helps teams think through whether and how to use algorithms in a way that connects ethical concerns to legal and policy frameworks.

- *Fundamental Rights and Algorithms Impact Assessment (FRAIA)*, created for the Ministry of the Interior, is an impact assessment tool that is similar to the IAMA, but with a greater focus on risks to fundamental rights.

- *AI Impact Assessment (AIIA)*, developed by the Ministry of Infrastructure and Water Management, the tool supports reflection on the use of AI in public projects. It supports transparency and accountability by recording decisions made during development in AI projects.

- *Data Protection Impact Assessment (DPIA),* developed by the Ministry of the Interior, this tool helps assess privacy risks and the sufficiency of (existing) safeguards. Its use is a requirement when personal data is involved.

- *Toolbox for Ethical Innovation*, from the Ministry of the Interior, gives civil servants and public organisations a starting point for using new technologies in a way that reflects public values. Hereto, the toolbox builds on the *Code for Good Digital Public Governance* [*Code Goed Digitaal Openbaar Bestuur*] (CODIO).

- *Code for Good Digital Public Governance (CODIO)* was developed in 2021 for the Ministry of the Interior in collaboration with Utrecht University. CODIO is based on an earlier code for good public governance from 2009 and introduces a value-based approach to digitalisation in public administration. The code and its principles are broad enough to apply to data and algorithmic systems, including AI. It is built around three main principles of: democracy, the rule of law and administrative power. Under *democracy*, the code highlights participation through citizen involvement, inclusivity, transparency, and collaboration. And societal value, which includes broader societal values such as sustainability, harm prevention, and collective interest. *The rule of law* relates to procedural fairness and human rights. This includes values like non-discrimination, explainability, user-friendliness, and the ability to contest decisions, as well as privacy, autonomy, and human dignity. *Administrative power* focuses on governance quality and responsibility, including adaptability, risk awareness, integrity, accountability, and human oversight. In total, the code includes around thirty values, grouped into seven categories. Each value comes with a short description and a practical example.

Concluding, although the available tools for responsible AI provide useful entry points for working with algorithms and data are large and complex. They are developed as comprehensive approaches that require a lot of time and attention to work through. Aimed at larger development and implementation projects, with the manpower and time to match. Making them unsuited for use by less demanding users, like individual policy officers requiring guidance on the responsible use of existing systems.

3.2.6 Conclusion: Tools for the Few

The Dutch government's approach to AI is still developing. On one side, the WRR has advised the government to start learning how AI can be used in its own work, to develop the required capacities. While others, like the state advocate and the AP, have advised acting with more caution and restraint. This leads to relatively restrictive governmental AI policy.

The use of AI in the Dutch government is mostly limited to small-scale pilots and experiments that support internal processes rather than the development of policy.
Insights from the field work, including the interviews, highlight barriers such as fragmentation of AI efforts within and between ministries, limited insight into the opportunities of AI for policy, limited AI literacy and capacity, and political risks.

A range of tools for responsible AI development and implementation exists, but these are large and complex. This leads to a high barrier to entry for policy professionals engaging with responsible AI. Echoing the concerns in literature, ethical principle-driven approaches are too high-level and abstract to inform practical, technical, or organisational decisions and actions (Hagendorff, 2020; Munn, 2023). Especially individual policy officers or teams. Meaning that a tool has to have a low barrier of entry to be inclusive of the needs of a broader audience, and to attract wide use.

3.3 People and organisations in the Design Context

This subsection first describes the actors (people and organisations) in the design challenge context. First, sketching the contextual terrain and the types of groups of actors, including a context map. Followed by an overview of the actors with a brief description of their position and function in the context.

3.3.1 Stakeholder Mapping of the Field

The contextual focus of this project is on policy officers as the main user group. To design something that works, it needs to fit with how they work and with the wider environment they operate in. That means understanding internal dynamics, political leadership, legal structures, and expectations from the public and civil society. To this end, the stakeholders in the wider field of the design problem are mapped. The map is based on the fields *Inside national government, National government adjacent,* and *Outside of (national) government*. Another axis to differentiate actors in the context is the distance to the design challenge, based on how involved they are. Both within and outside the government, different actors vary in their distance (in influence, interaction, etc.) from the design challenge of responsible AI in policy preparation. For the resulting context map, see Figure 15. Continuing, the mapped actors are discussed per field in the following sub-sections.

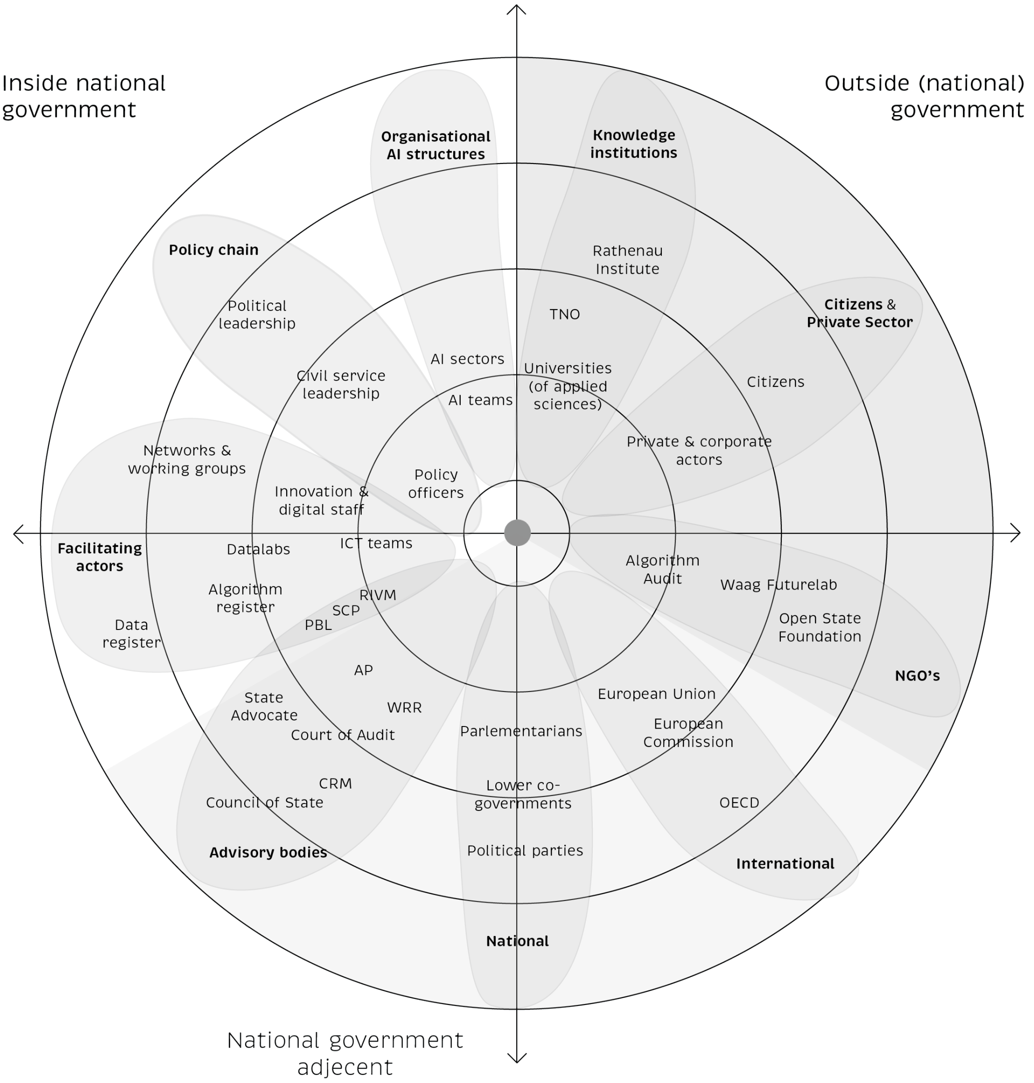


Figure 13. Stakeholder map of actors in the design context

3.3.2 Actors in National Government

The main people and institutions in the application domain are those involved in the policy preparation process. Among them are policy officers and their leadership. They are responsible for delivering policy in a way that is legally sound, politically viable, and publicly legitimate. Their main interest is developing policy that works, that builds trust, and that can be implemented . Secondary interests may include political positioning, visibility, and maintaining or gaining influence (Claessen et al., 2021). Political leadership is also part of this policy development chain and has the final say in government policy, especially regarding internal policy. Of course, dedicated organisational AI structures such as AI sectors and AI team are important stakeholders in the development or governance of AI Within the national government there an additional group of facilitating actors, these are actors who aren't part of the policy chain, but are involved in the data systems or required capabilities within the organisation, Think for instance of human resources- and IT departments.

National government
Policy development chain

- *Policy officers* - are the direct users of the design. They carry out policy preparation work, often under pressure, within complex systems of rules, expectations, and time constraints.

- *Civil service leadership* - plays a key role in whether the design gets used. They put things on the agenda, approve internal processes, and make sure the necessary structures are in place.

- *Political leadership* - including ministers and state secretaries, have the formal say over the policy direction of the ministry, within the limits set by parliament and the courts.

Note: More insights about these direct stakeholders of AI for policy and the future design, collected during the field work of the design project, are described in the 'Emergent themes' subsection in the proposed design chapter.

Organisational AI structures

- *AI* sectors *and teams* – multiple ministries have sectors or teams in their organisation that are dedicated to working on AI and related subjects. They have a lot of relevant knowledge and expertise.

Facilitating actors

- *Technical and implementation actors* - including ICT teams, datalabs, and service organisations.

- *Innovation and digital staff* - across ministries and implementing organisations, there are directorates and staff dedicated to innovation, which can be a supportive force in enabling innovation and new ways of working around responsible AI for policy.

- *Government-wide networks and working groups* - interdepartmental networks and working groups within the national government.

- *Netherlands Bureau for Economic Policy Analysis (CPB), Netherlands Institute for Social Research (SCP), Netherlands Environmental Assessment Agency (PBL), Research and Documentation Centre (WODC), National Institute for Public Health and the Environment (RIVM)* - are independent advisory bodies that are organisationally part of ministries. These institutions advise the government with advice in a variety of fields (e.g. ECP, 2018). Furthermore, they have much experience with data and modelling that inform the policy process.

- *Algorithm register* - led by the Ministry of the Interior, the register allows government organisations to publish which algorithms they use and how they work. Reflecting a broader move toward accountability around data, algorithms and AI within the government.

National government and adjacent actors

The government works together with actors like co-legislators and other governing bodies. And Governmental organisations for international collaboration.

National

- *Parliamentarians and political parties* have a direct say in what gets prioritised. In the Dutch context, most political leaders within ministries come from political parties and remain tied to their agendas (Otjes & Louwerse, 2018). In addition to approving laws proposed by the government, parliamentarians are co-legislators, with the constitutional right to propose legislation[3]. Furthermore, their work is informed by the internal policy preparation within ministries (van der Staaij & Sneller, 2023)

- *Lower co-governments* - like provinces and municipalities that are organised in the IPO and VNG. Use the tools and insights created at the national level. Reverse, they often have more freedom to experiment with innovation. Lessons learned on the local level can be valuable at the national level.

International

- *European Union (EU) and European Commission (EC)*, responsible for the development of European regulation, European Union-wide AI, regulations with the EU AI Act, and other applicable legal frameworks like the General Data Protection Regulation (GDPR).

- *Organisation for Economic Co-operation and Development (OECD)* is an international collaboration organisation, hosting a variety of expert groups on subjects related to AI. As well as providing recommendations (OECD, 2024).

3.3.3 Independent Advisory Bodies and State Advocate

Legal and other advice and expertise are, in part, organised internally. But for complex or independent advice, requested and unrequested, the government relies on independent advisory bodies and knowledge institutions. For complex legal questions, external legal expertise is contracted at the State Advocate [landsadvocaat].

- *The Advisory Division of the Council of State (Afdeling advisering van de Raad van State)* is a constitutional, independent council to the government, that reviews major policy and legislative proposals to provide an opinion before they are submitted to parliament. When proposals involve new technology, they are mainly concerned with questions about legal certainty, enforceability, and whether the policy respects fundamental rights (Raad van State, 2021).

[3] Grondwet, Art. 82.3

- The *Scientific Council for Government Policy* (WRR) gives strategic advice on long-term policy issues. Their report, *Mission AI [Opgave AI],* adviced the government to consider learning about AI a key priority, and warns against over-reliance on technical systems and stresses the importance of human dignity in public AI use (WRR, 2021).

- The *Dutch Data Protection Authority* (AP) monitors how personal data is handled and has previously warned about the risks of using algorithms that aren't transparent or fair. In the case of *SyRI*, the AP was one of the organisations that raised concerns about the lack of explainability and legal safeguards. The system was eventually ruled to be unlawful (Autoriteit Persoonsgegevens, 2024).

- *Court of Audit [Algemene Rekenkamer]* are tasked with analysing the functioning of public systems, in addition to monitoring if money is being spent responsibly. They have published several reports warning of risks in digitalisation, pointing to how systems sometimes do not deliver what was promised..

- *Netherlands Institute for Human Rights (CRM)* focuses on equality and fairness in automated systems. Have raised concerns about algorithmic discrimination and emphasised the need for transparency and the right to an explanation.

- *The State Advocate* is a legal firm contracted by the state. They represent the state in court cases, and can be requested to provide advice on complex legal questions. For example, to provide analysis and recommendations for the government's policy, like advice on the legal risks involved in the use of generative AI.

Knowledge institutions

- *Netherlands Organisation for Applied Scientific Research (TNO*) works with ministries on the technical and ethical evaluation of AI systems. They support questions around explainability, robustness, and legal compliance.

- *Rathenau Institute* researches how technology affects democracy and public values. Calls for stronger democratic control over digital systems,

- *Universities and universities of applied sciences* like the TU Delft and the University of Utrecht have great academic expertise in relevant fields. They can provide or be contracted for academic research, and can be invited to share expert knowledge and reflection. Furthermore, technical universities develop leading-edge technical expertise.

3.3.4 Citizens, Civic Society and the Private Sector

Citizens and the private sector are not necessarily directly implicated in the design challenge; they play a key role in shaping what is seen as legitimate, fair, and acceptable in the public domain. Companies in the private sector can be important partners in the development and implementation of technical systems.

Non-governmental organisations

- *Algorithm Audit* is an independent non-profit that offers algorithm audits and advice. They work mostly with public organisations and help build knowledge on responsible AI use. An example is their audit of DUO's risk profiling system, which found an overrepresentation of certain groups and led to concrete recommendations for fairer design and implementation.

- *Interest groups and civil society organisations* represent social and political viewpoints that shape how policy is received. Representing public values or offering critical perspectives. Think of organisations like *Waag Futurelab*, and the *Open State Foundation,* which advocate for ethical and participatory approaches to public technology. While these groups operate outside of the central government, they are involved in shaping public debate and collaborating.

Citizens and the Private Sector

- *Citizens* are the people affected by the policy. They play a vital role in the democratic polity by engaging and complying with the government. Their trust in government is shaped by whether the policy works, whether they feel heard, and whether decisions are seen as fair (Schakel, 2021).

- *Private and corporate actors* are needed to build AI capabilities and applications. Furthermore, they have a significant influence on the political and policy agenda (Schakel, 2021), especially in the context of the Dutch *polder model* of consensus-building between government and industry (Schreuder, 2001) .

3.3.5. Conclusion: Missing Pathways to a Fragmented AI Landscape

In and around government, there is a wide and diverse network of actors involved in developing AI capabilities and AI capacity at large. This includes colleagues with experience in AI or data, innovation advisors, or staff working in digital teams or datalabs within executive agencies. There are several organisations outside of the ministerial structure that play a role in shaping how technologies like AI and data systems are used in the Dutch government. These organisations do not make policy, but their advice, audits, and tools influence what is seen as legally sound, socially responsible, and politically acceptable. Together, they form what's often called a quadruple helix, a collaboration between government, research, the private sector, and society (Carayannis & Campbell, 2012; Bharosa & Janssen, n.d.). Each of these groups can play a different role in the responsible adoption of AI for policy. Some focus on legal checks, others on ethics, technical support, or policy reflection. All needed for success.
Together, these organisations, formal and informal, institutional and civic, help set the boundaries for how AI is used in the public sector. They influence not only what is legally possible, but also what is seen as legitimate and responsible. These actors influence what is (seen as) possible, necessary, or risky. Their involvement may not be visible in day-to-day policy work, but they help define the broader conditions under which AI enters public administration.

The challenge is that this network is fragmented. Knowledge, tools, and people are spread out across different teams and organisations, making it difficult to know where to go or who to involve. Especially to policy officers unfamiliar with the field. In practice, most policy officers start by asking people within their own organisation.

4. Proposed design

A design process is inherently nonlinear (Roozenburg & Cross, 1991), and includes many micro iterations and activities. Still, this section aims to describe the process, focusing on the main activities and insights. The chapter sets out with the design approach of the project. Followed by the results derived from the fieldwork and design activities, namely the principles, which form the basis for the proposed design. Finally, the proposed design is discussed with a description of the prototype and rationale.

4.1 Methods: Constructive Design Research

The thesis aims use a design project to help answering the research question. This approach is called research through design (Frayling, 1993; Zimmerman & Forlizzi, 2008). However, this term is somewhat underdetermined, as it is used to refer to differing concepts by its varying proponents. More specifically, the form of research through design used in this thesis is based on the practices of constructive design research as described by Koskinen (2011).

Constructive design research is a method that uses the construction of a design (in most cases, a prototype) as a means of creating knowledge (Koskinen, 2011). The fieldwork of constructive design research doesn't need to produce generalisable knowledge. It helps answer the secondary research question (*RQ 1: What does responsible use of AI entail in policy preparation?* And, *RQ 2: How should a tool for policy preparation professionals be designed to effectively incorporate advice?*) that take on a role as design question in the design project, to inform the design of a prototype. The prototype is a materialisation of a theoretical hypothesis. The prototype design is evaluated through evaluation sessions with participants from the design context. Forming the basis for answering the main research question of this thesis (RQ 0: *How can we design practical tools to guide the use of AI in the governmental policy development process?*). Adding to knowledge in this field by focusing on the unique case at hand (Koskinen, 2011). The method of the evaluation session is discussed in Chapter 5. Evaluation.

4.2 Design approach

There is an ever-growing variety of design approaches, tailored to different types of design challenges. This subsection first discusses the considerations that should be taken into account in the selection of an approach. Next, the type of design problem of the current design project is discussed. Followed by a description of the selected approaches, including rationale and a description of how they come together as one approach for the project. Ending with an overview of the design process, including its phases and methods.

4.2.1 Considerations in Design Approach Selection

Multiple factors have to be considered in the selection of the design approach for a project. First, the nature of the design challenge is an important consideration (Dorst, 2004; Roozenburg & Cross, 1991). For example, designing the ergonomics of an office chair requires a different approach than designing the interaction between a user and a smart voice assistant helping vulnerable citizens file their taxes. Additionally, the designer's qualities and experience with approaches are important factors that need to be taken into account (Roozenburg & Cross, 1991). The character of the design challenge of this design project is described in the following subsection. How this, and the designer's qualities and experience of me, are taken into account in the selection of the methods is described in the following subsections about the project's design approach and methods.

4.2.2 Nature of the Design Problem

The design challenge in this thesis project, enabling responsible use of AI for policy, is a “wicked” problem (Buchanan, 1992; Churchman, 1967; Rittel & Webber, 1973). As it is situated around rapidly developing technology in an open, dynamic and complex context (Howlett & Ramesh, 2023) of government, which has a great societal impact. This context includes a large number of stakeholders, with contradicting interests, making the design problem especially “open, complex, and dynamic” (Dorst, 2015b, p. 15), centres around paradoxes and value tensions. A dynamic and opaque problem and context, like the one faced here, makes ex ante scripting of process and result impracticable. Necessitate that problem, approach, and solution co-evolve throughout the project.

4.2.3 Selected Approach(es)

To allow for the co-evolution of design challenge, design approach and solution, the design approach for the project is based on three theoretical design approaches: Frame Innovation (Dorst, 2015b), Vision in Product Design (ViP) (Hekkert & van Dijk, 2011) and Value Sensitive Design (VSD) (Friedman & Hendry, 2019). These approaches together cover the wider field of challenges that were to be expected throughout the project. In the selection of the approaches, I opted for two approaches with which I have prior design project experience (Frame Innovation and ViP). Furthermore, both Frame Innovation and ViP utilise verbal reasoning qualities, which is a strength of me as a designer. Adding Value Sensitive Design due to the large role values play in the design challenge at hand. The individual approaches are further explored subsections below.

Frame Innovation

The Frame Innovation approach, developed by Dorst (2015a, 2015b), is a method for wicked, open-ended, socio-technical challenges. The approach centres around the assertion that design challenges with clashing values (a paradox) can often be resolved by viewing the situation from a different perspective (reframing). The approach embraces the complexity of the design challenge and context. And being open to novel solutions by starting with a desired value, leaving the means of achieving it up to the final phases of the design process. The method is an anticipation of designers moving out of the traditional design field to apply themselves to design challenges in society at large, where they encounter wicked problems in more complex and networked contexts. For the steps of the approach, see Figure 16.

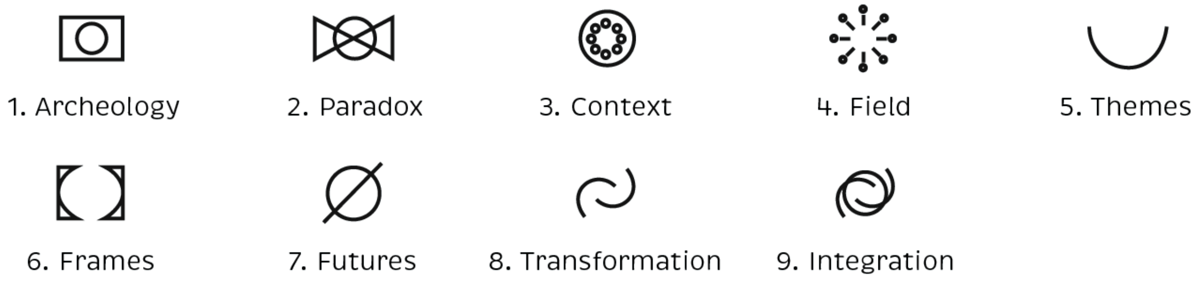


Figure 14. Frame Innovation steps. Adapted from Dorst (2015b).

Dorst's design approach is especially fit for "technical communicators" to use their linguistic, conceptual cognitive abilities in solving wicked problems according to Weedon (2019). The method of Dorst centres around the use of "rhetorical skills" (p.2) to frame wicked problems in a new way. (Weedon, 2019). This aligns with my abilities as a designer, which are centred around linguistic, conceptual cognitive abilities. and the problem that is the subject of the project. Furthermore, I have used the approach before in a project for the same organisation.

Vision in Product Design (VIP)

Vision in Product design (ViP), developed by Hekkert & Dijk (2011) is a future-focused design approach centred on finding the raison d'être of a design., The Vision in product approach is aimed at taking a position by creating a vision "Responsible and authentic that will steer the conceptualisation.". Clarifying what the designer wishes to enable in the future for the people in the design environment. Creating a vision as a precursor to defining how the design can do this.

The VIP approach consists of a preparation and a design phase. The preparation phase does not apply to the design challenge and approach used in the project; only the design phase is adapted. The design phase initially focuses on the future context. For the steps of this phase, see Figure 18. Considering what is interesting and relevant to the design challenge. Integrating supporting facts and allotted personal motives and intuition, as well as the aims or desires of the client or market. A core selection of these forms the basis for the worldview. Statement, the position or for the deep should the offer, do you people to experience. can be transformed into a desired product

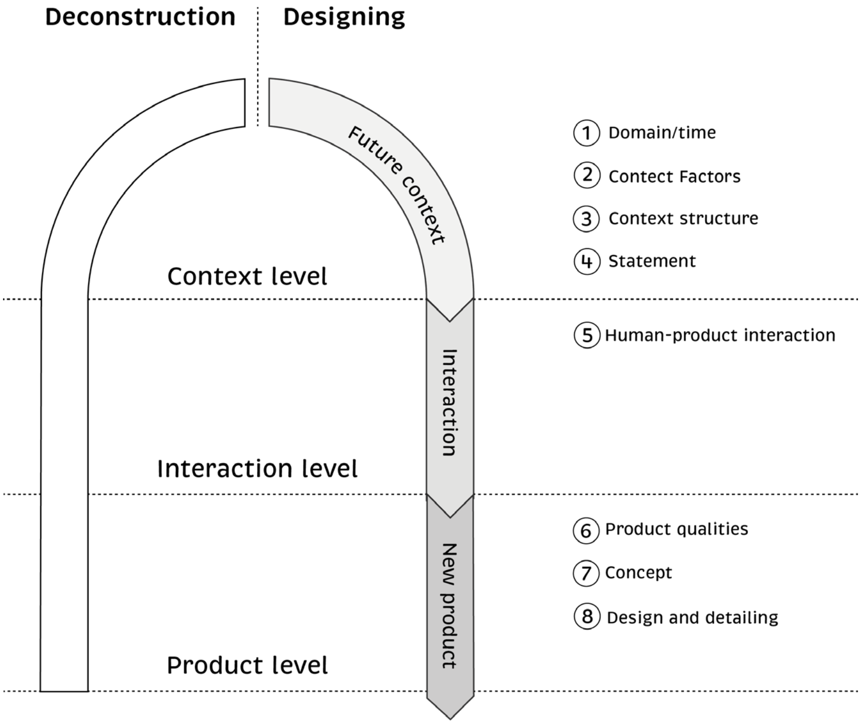


Figure 15. 8 steps of the process embedded in the ViP model. Adapted from Hekkert & Dijk (2011, p. 133).

interaction. After which, the product characteristics can be refined.

The approach is less concerned with user involvement early in the design process. *"What we take issue with is not end-user involvement, but that the insights thus obtained are often rooted in the situation the user is in at that moment. Users act in certain ways because of the designed environment they are in." (p.184)*

The method has a strong focus on developing an understanding of the impact and the desired interaction (a vision) before starting to design the product itself. Leaving less room for iteration of the design itself. The method helps to find a position, but relies on the designer to take on the position and defend it consistently and convincingly. The approach relies strongly on the designer(s) conceptual and abstract thinking, which is a strength of me as a designer.

The approach used in the project is partly based on the approach described in the book and on the approach taught by Paul Hekkert in the ViP elective at the Faculty of Industrial Design Engineering of the TU Delft (Hekkert, 2023).

Value Sensitive Design

Human values are always reflected in technologies in some form, and in turn, technologies influence human values. The Value Sensitive Design (VSD) approach (Friedman & Hendry, 2019) foregrounds active engagement with human values in the design process. Creating "creative opportunities for technical innovation as well as for improving the human condition." (p.1). "Specifically, it provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process." (p.34). Defining human values as "what is important to people in their lives, with a focus on ethics and morality." (p.4)

The engagement with human values is especially relevant to the intersection of & government and AI. The approach is equally of interest to the design process of this project as to those of its subject, the development and use of IT in the policy development process. The definition of human values is open to interpretation as to what is important to the lives of people, and what ethical and moral principles the designer should be sensitive to. Here, the VIP method is useful as it helps set a vision. Allowing me to take a position based on historical-societal values, institutional values, the values of current stakeholders, as well as my of me as a designer.
Tripartite methodology: conceptual, empirical, and Technical investigations

Approach Cohesion

To make the three theoretical approaches work together as one, cohesive design approach, they are chopped up and moulded together to the requirements of the design process. As design methods are mostly descriptive, they cannot fully account for the nonlinear nature of a design process (Roozenburg & Cross, 1991) they function as a reference that gets a project. Such adaptation of the steps and methods of the approach, to fit the project and the designer's needs, is explicitly embraced by the selected method (Dorst, 2015b, p. 99; Friedman & Hendry, 2019, p. 102; Hekkert & van Dijk, 2011, p. 132).

Frame Innovation is the main methodological grounding of the design approach of this project. Additionally, the design problem poses philosophical, political, and experiential questions. Questions that have no definitive answers, but rather require a vision. To this end, the vision in product design method (ViP) was invoked, in conjunction with methods and attention for themes derived from VSD.

The phases of the Frame Innovation and ViP approach have different focal points, but can be massaged to loosely line up. This allows the approaches to be used simultaneously, and leaning more heavily on one or the other as the design process requires (see Figure 23). The VSD approach is less concerned with the structure of the design process. It is used as a lens through which the design process as a whole is coloured. By focusing on human values throughout the process, and by appropriating the methods of design activities in the steps of the other approaches.

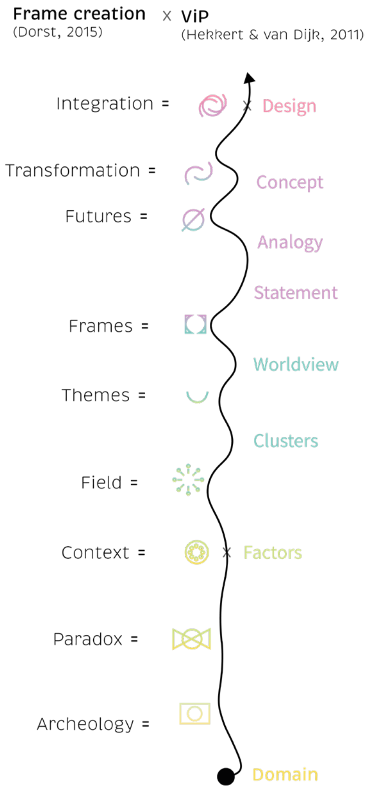


Figure 16. Illustration of the use of the Frame Innovation and Vision in Product design methods throughout the project

Additional design methods

As the approaches are focused on a higher-order conceptual level, they are combined with practicable methods of design activities. Additional sources that informed these methods are, good old product design structure and methods by Roozenburg & Eekels (2016), the Delft design guide (van Boeijen et al., 2013). These methods are greatly influenced by a tradition of Human Centred Design (HCD) (Giacomin, 2014). As this thesis project is undertaken in the context of an MSc in Strategic Product Design, the project is naturally also influenced by approaches and perspectives from that field. By taking a systemic view, and the consideration of systemic pressures in the development of the design and strategic implementation recommendations how the approaches and methods come together in the project's design approach is visualised in Figure 24.

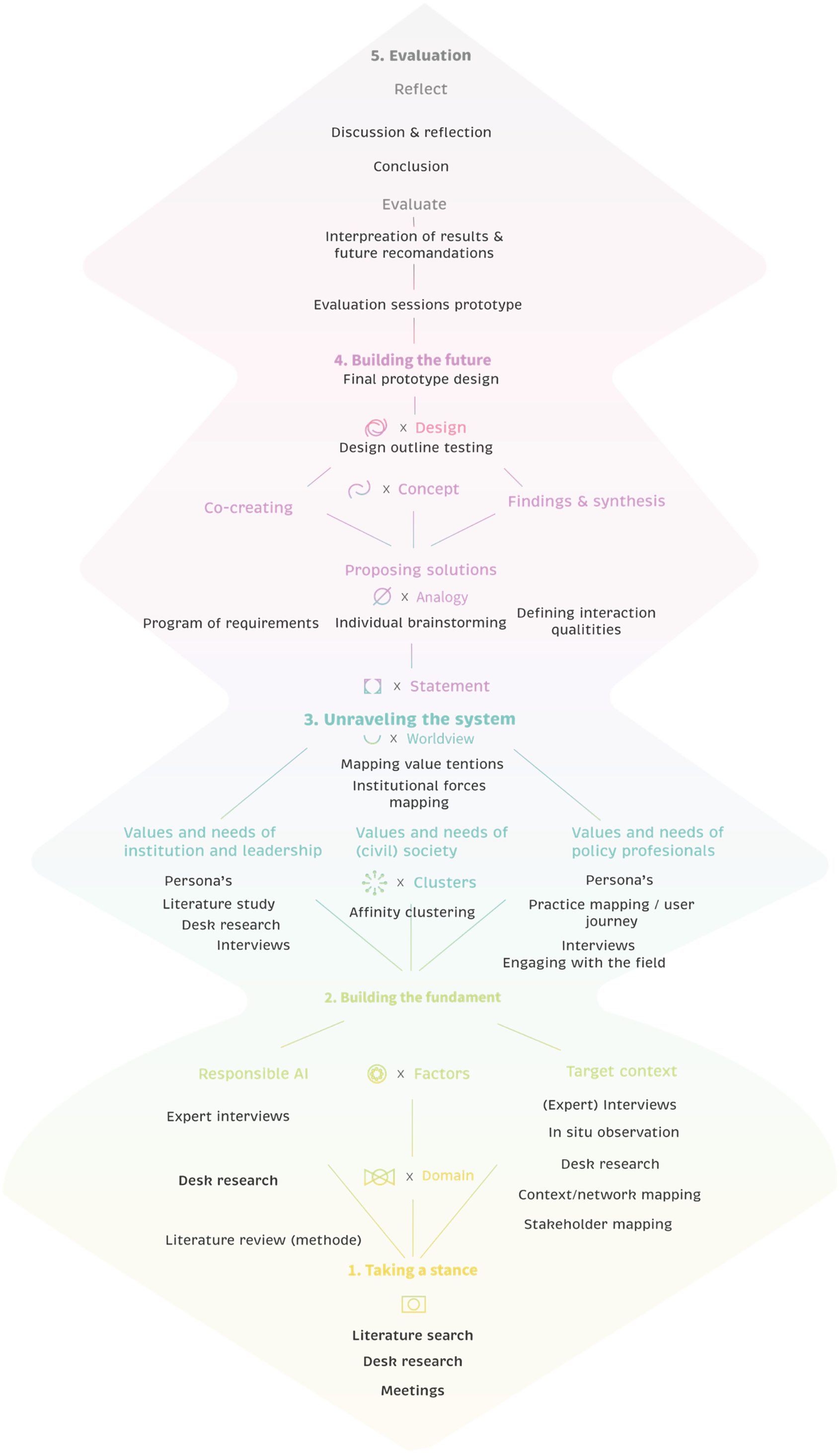


Figure 17. Overview of the design process

4.2.4 Design Process Overview

The design process consists of multiple stages (see Figure 24, starting from the bottom):

1. Taking a stance, the initial stage of defining the domain of the design problem. Here, the design challenge is outlined, and an initial scope and framing of what is relevant to the project are defined through an initial framing.

2. Building the fundament - the second phase is concerned with understanding what is happening and who is involved. Creating an overview and basic understanding of the context, on which the later stages of the project can build.

3. Unravelling the system – the third phase dives deeper into processes and values in the context and the wider field, and structures the insights to understand the system as a whole.

4. Building the future – the fourth phase moves from understanding and framing the problem to reframing and ideating concepts for the design. Iterating towards the design prototype.

5. Evaluation – the fifth and final phase is where the prototype is evaluated and new insights are used to create recommendations.

4.1 Fieldwork

The fieldwork performed in the project consisted of participant observation, by engaging in a diverse range of meetings and activities in the governmental organisation. And semi-structured interviews with experts and stakeholders. As well as attending events and meetings of organisations concerned with topics related to algorithms, AI, and digital technology in government to engage with the wider field. Making up an important part of the first 3 phases of the design process. Adding to the literature and desk research by providing deep situated insights.

The aim of the fieldwork was twofold. Firstly, ethnographically informed inquiry (Friedman & Hendry, 2019) surfacing meaning, values, behaviours, and informal norms and practices. Secondly, gathering contextual information on formal structures, capabilities, systems, infrastructure, procedures, and roles. Together, providing a holistic comprehension of the design challenge situated in its context. The insights from the fieldwork contributed to the description of the design context in the previous chapter, and the design work described in the following sections.

Role of fieldwork in the design approach
The approaches of Frame innovation and sensitive design, field work takes a significant role. In Frame Innovation fieldwork is aimed at understanding the context and the discourses shaping the design challenge. SVD puts fieldwork at its core as a way to uncover and engage with the values of the stakeholders. In the ViP method, fieldwork plays a less important role in the initial phases of the design process as it aims to detach more from the current constraints. Yet with this in mind, the fieldwork is a great way of collecting factors.

Attribution of insights from the fieldwork
While some insights can be attributed to specific organisations and people, not all can. The interviews were conducted under the promise that insights could not be traced back to the interviewees to allow them to speak freely. Other insights have been collected through off-the-record conversations and in non-public meetings and can therefore not be attributed directly.

Continuing, the following subsections discuss the fieldwork activities, including a description of the methods that were used. It starts with the participant observations, followed by a description of the interviews conducted. The section concludes with the synthesis of the fieldwork findings in the emergent themes.

4.1.1 Participant Observations

Participant observation (Aktinson & Hammersley, 1998), observation by participating in the organisation. During the project, I did internship activities like attending meetings on topics not directly related to the topics of this work that have helped gain an understanding of the dynamics of policy-making, the function of the Policy Compass as well as its functioning.

The focus is limited to AI for policy (in the development of policy and laws), as opposed to the related subjects of AI in policy (execution), and AI regulations. Mainly aimed at the preparative stages of policy development, although evaluation and broader uses of AI within the organisation related to policy development could be part of the scope.
Through the graduation internship at the governmental organisation, I was able to attend meetings. I attended events from organisations in and around the government to gain an understanding of the broader field. An overview of the meetings participated in during the project, including key insights, can be found in Appendix 1. Furthermore, throughout the project,

I held countless meetings with company mentors to better understand the needs of the organisation, and to gather information based on their expertise on various subjects related to the subject.

Additionally being in the office, listening in to all sorts of conversations has provided countless minor insights that add a lot of depth to the understanding of the dynamics within the governmental organisation. Examples of these conversations I overheard are colleagues asking others for their perspective on a specific detail in their policy proposal, and strategic discussions on how to bring a subject into the council of ministers to get it passed without hiccups.

Visiting events in the field
To gain an understanding of the concerns and developments of organisations in the network of the governmental organisation, I attended events organised by organisations in the field (illustrated in Figure 25). An overview of the events attended, including key insights, can be found in Appendix 2.

4.1.2 Interviews

Semi-structured interviews were conducted with the use of an interview guide. Mixing traditional informational questions with a value-oriented interview approach (Friedman & Hendry, 2019). Inquiring interviewees about information and ethnographically informed insights. Interviews were conducted with two experts on the developments around AI at the governmental organisation, an expert on the development and implementation of the Policy Compass, a policy assessor with expertise on privacy and AI law and regulations, and an expert on neuropsychology. To get deep information and visions from experts and stakeholders. A list of the interviews conducted can be found in Appendix X.

In addition to these formal interviews, many informal, non-structured interviews were conducted throughout the project. The insights from these conversations and the other fieldwork are clustered and integrated into emergent themes. These are discussed in the next subsection.

4.2. Worldview of Responsible AI for Policy in the Government

All previous steps of the design project have collected and structured insights, including the literature in the background chapter, the design context, including the wider field, in the design context chapter, fieldwork and emergent themes sections. These have created some order in the context. Following playing around to find interrelatedness to develop a worldview, which is a prominent step in the approach of (Hekkert, 2023), and is part of the third step of the design phase in the ViP approach (Hekkert & van Dijk, 2011). Bringing all insights together, converging towards the frames (Frame Innovation) and statement (ViP) that follow this stage. Resulting in the following worldview of responsible AI for policy in the government:

Responsible AI requires '*a balanced consideration of AI for policy.* ' To enable a balanced consideration throughout the life cycle of an AI system, at all levels of the organisation, the government needs to move *towards collective soft AI capacity for constructive problematisation'.* including AI literacy. Creating the conditions for critical responses and contestation, and allowing for the renegotiation of the institutions of the policy cycle, to '*preserve procedural* *legitimacy and enhance policy quality'.* By going from '*tools for the few'*, to one tool for the many that fills the *'missing pathways to a fragmented AI landscape'.* Using a guidance approach based on '*balancing guardrails and trust in professional judgment'.*

4.3 New Frames and a Design Vision

In the ViP approach, the vision consists of the statement, an interaction, and product qualities. These steps can be taken together with Frame Innovation's frames, futures and transformation stages. Both the statement and frame define a new way of seeing the world/the design challenge. The interaction qualities and futures are oriented towards the implications of behaviour and experience. And the product qualities and transformation move to translate the conceptual into tangible design attributes.

4.x.x New Frames for the Design Challenge

The creation of frames is the pivotal step in the Frame Innovation approach. These 'frames' are new ways to view the design situation, in a way that reduces the tension between the values of stakeholders. Allowing for the development of new solutions. Frames can be expressed in the form: "If the problem situation is approached as if it is ... , then ... " (Dorst, 2015b, p. 78).

The initial framing problem of the design problem in this project was: we need to give people guidance on responsible use of AI for policy, to make sure people act responsibly. Here, the problem situation is approached as a regulatory challenge, seeking solutions based on strict rules and control. This creates a paradox, as it reduces the engagement and investment of the very people who must follow the guidance to realise responsible use of AI for policy in practice.

Throughout the project, many ways of seeing the problem situation have been explored, for example: If responsible AI for policy is approached as a soft capacity challenge, then all stakeholders have a valuable role in the solution; and, if we see responsible AI for policy as a conversation, responsibility becomes something created together through dialogue, not dictated through rules. Leading to the final frames:

If we see responsible AI for policy as a shared garden, then all stakeholders have a role in tending, nurturing, and sustaining it.

AND

If we approach guidance for responsible AI like medical protocols, then users are given enough guidance to avoid catastrophic mistakes, while being encouraged to keep thinking for themselves, leading to trustworthy professional judgment.

4.3.1 Design Vision: AI curiosity

The future vision materialises the design goals into a concrete desired outcome in the design context. Bringing to life an optimal future. It includes a statement with an associated analogy and the product qualities derived from the statement.

4.4.x.x Statement

The statement is the step in the ViP approach where the designer takes a position (Hekkert & van Dijk, 2011, p. 156). As such, the statement is an expression of what the design should bring about in the world. The following is the statement formulated in the project (illustrated with Figure 26):

Figure 18. Girl playing curiously. Adapted from (Pixabay, 2014).

Creating a tool that inspires AI curiosity by lowering the barrier of entry to AI knowledge to develop organisation-wide AI literacy.

As a means to develop the capacity to responsibly adopt AI for policy, through a self-learning organisation with a broad base for internal and external problematisation and contestation.

4.4 Construction of the Design

The previous steps of the design process have focused on gaining an understanding got the effect the design should offer, and what sort of experience is an effective means. In the construction of a design, these conceptualisations need to be transformed into a concrete prototype design with a specified content, structure and form

4.4.1 Design Qualities, Mechanisms and Features

Brainstorming

Working backwards from the desired effect, established in the worldview, frames, and statement, the experience, product qualities, and design mechanisms were ideated through iterative brainstorming. This process created the basis for the elements that make up the design: its form, content, and structure. The relation between the elements of the design, the experience, and the effect is visualised in Figure 27.

In the ViP approach, product qualities capture both the product's personality and its behavioural qualities .Discribing what should provoked either by design or evoked through interaction (Hekkert & van Dijk, 2011). These qualities are mostly derived from the statement, where they are implied in the analogy.

Additionally, design mechanisms are needed, which set out what the design should do and formulate a strategy to do so. These were ideated based on the statement and the frames developed earlier, as their metaphors already imply certain mechanisms, as well as on themes and factors derived from the literature study and the fieldwork.

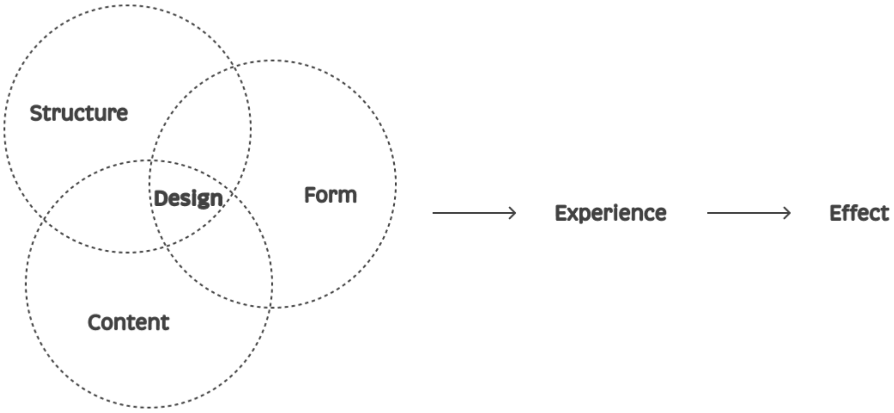


Figure 19. The elements of the design, and their relation with its experience and effect

The statement and associated analogy inform the design and its interaction qualities. However, AI curiosity needs to be embedded in a broader design. To explore this, additional metaphors were used to develop the functional product qualities.

Functional metaphors

To further explore the functional interaction the design should fulfil, metaphors were used, see Figure 23. Should using AI through the design be like food delivery, “ready to eat”? Or should people who want to use AI in the organisation pass an inspection, like at a border crossing? These metaphors reflect varying degrees of paternalism versus freedom, and conservatism versus ease of use.

The functional metaphor that best fits the frames and the statement is that of a foundation, combined with a coatrack [kapstok], to which all relevant resources can be attached, serving a coatrack function [kapstokfunctie].

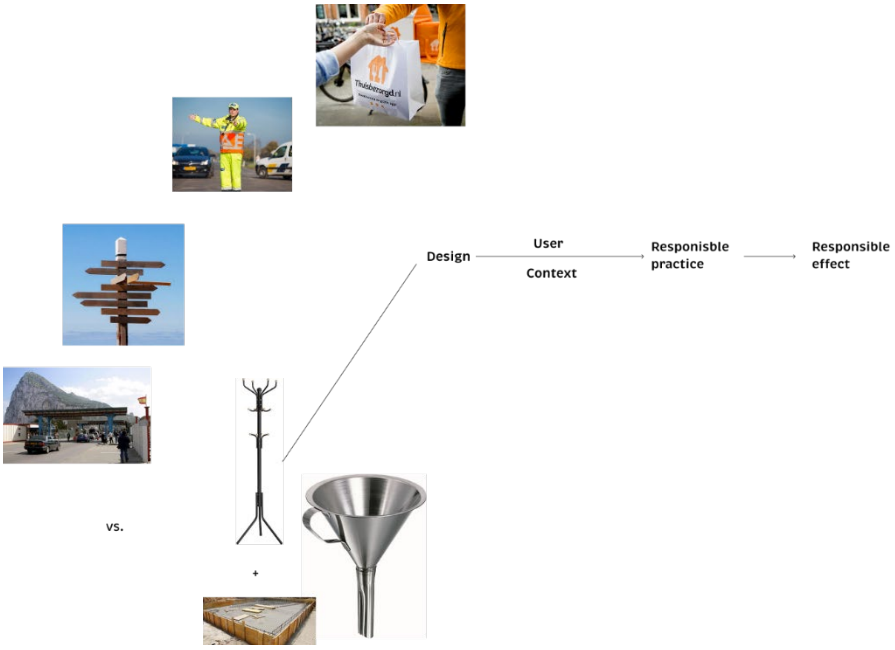


Figure 20. Metaphors used in the design process

Design Mechanisms of Curiosity, Trust, and Collective Capacity

The iterative brainstorming, based on the results of the preceding design steps and the ideation of novel functional metaphors, resulted in the following design mechanisms. These mechanisms integrate the design qualities and translate the conceptual vision into functional strategies that guide the design.

Collective AI capacity

The smaller a risk mitigation issue or dilemma is, the closer someone would want to find an answer before overstepping it entirely and hoping for the best. Part of responsible AI, so you can talk about minor things with a colleague at the coffee machine. If a degree of AI literacy is present throughout the organisation. The central role of the Policy Compass in the policy development process (at least in theory) makes it a great place to lay a collective foundation for collective AI capacity and, with that, the responsible adoption of AI in the policy development process.

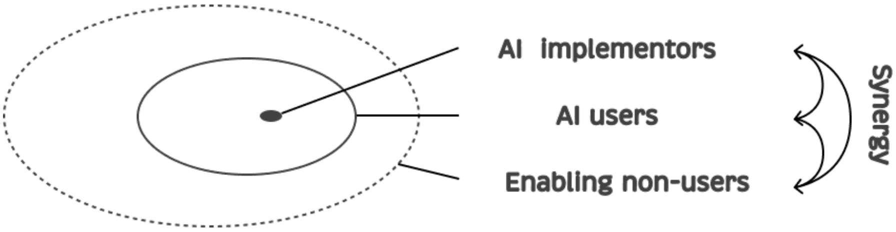


Figure 21. Groups needed for synergetic AI capacity

Inciting AI curiosity

AI curiosity - a mechanism the design intends to utilise is the mental pathway of curiosity, catalysing learning, and moving from passive understanding to acting. This mechanism also informs the design qualities of the design. Curiosity is an eager wish to know or learn more about something (Cambridge University Press, n.d.a). The design is to inspire a desire to learn more about (responsible) AI for policy. To create an artefact that supports the development of AI capacity, the ability to adopt AI in the organisation. To facilitate adoption, the capabilities should be present throughout the organisation. Engagement with AI, resulting from AI curiosity, can be a way to ensure critical thinking about AI use, distribute responsibility, and create resilience, limiting the risk of failures. A design element that is already part of the Policy Compass design is *support questions*, questions help users think about the topic at hand concerning their situation.

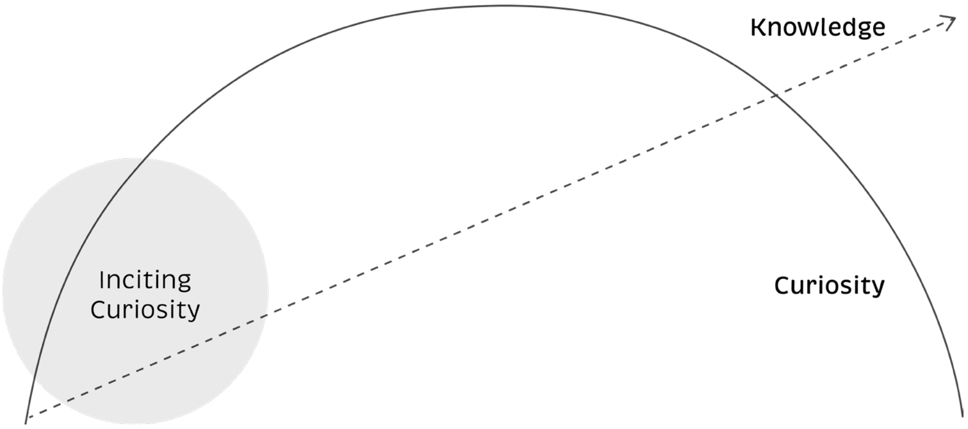


Figure 22. Simple model of the relationship between knowledge and curiosity.

Trust and responsibility

In line with the need to create enthusiastic AI curiosity while avoiding harm from the use of AI in the policy development process, a balance is needed that ensures (enough) risk encapsulation while not crushing the inherently curious spirit of policy officers. Based on competence and trust. A popular example, offering a compelling analogy, is the story of the management book "Turn this ship around" (Marquet, 2019) describing how U.S. Navy Captain David Marquet transformed one of the worst-performing submarines into one of the best by shifting from a command-and-control model to one based on trust and distributed competence. Empowered his crew to make decisions within clear boundaries of responsibility and safety, fostering engagement, accountability, and professional pride.

Uphold standards

The rules should be made clear to the users. If it is not worth it to use AI responsibly, including the additional effort that is required for this, it simply isn't worth it to use it at all.

Attention for dilemmas

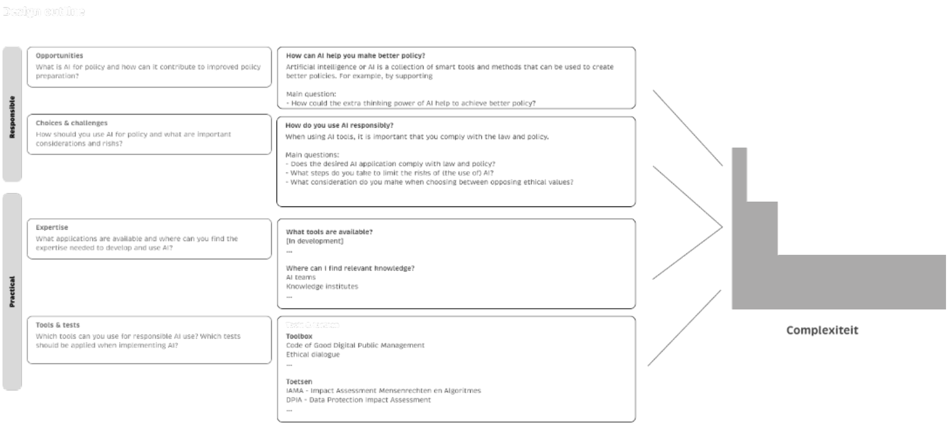
Not all situations can be captured with a responsible AI policy. And even so, if someone is scraping the guardrails put in place to avoid disasters, they sure are not practising responsible AI. As the experiences of IBDS highlight, many of their cases ended up being ethical dilemmas without codified answers in law or policy. Responsible AI is a result and a process stakeholders have to actively part take in. The design has to make users aware that active exploration of dilemmas in the use of AI is needed to use AI responsibly, as well as providing tools they can use to this end.

To allow for the adoption, ministerial staff have to be AI-literate. For example, managers and human resources officers who have a better understanding of what AI implementation requires of them are better able to recruit and develop a workforce that is literate and skilled in AI (Wirtz et al., 2019).

Anchor function
There is a rapid increase in different organisations creating recommendations, guidelines and tools. Furthermore, more AI applications can touch on a lot of areas like data and privacy, for which even more registers, tools, and assessments exist. This disjointed nature of the efforts, resources, and expertise relevant to the responsible implementation of AI in the policy process leads to stacking uncertainty and complexity. It is important to bring together and make available all this expertise and support to policy developers open to responsibly implementing AI, to make it easier for users to navigate. The design can take on an anchor function [kapstok function], redirecting users to the information, tools, and expertise they require. The Policy Compass is a natural place for this, as it is central to the policy development process.

Low barrier to entry
Because of the diversity of the user groups of the tool, think of the different needs of managers and policy officers. Add to that the differences think of the differences within these groups in their exposure and understanding of AI and technology in general, and all the other factors that influence their needs for this tool. It is the role of the Policy Compass to speak to all.

Layering
To provide user groups with different information and support needs with the right information, the design has to be layered in multiple ways. Firstly, by presenting lighter, more positive or motivating content at the top, followed by increasingly more complex content and resources (illustrated in Figure 26). Secondly, while one of the principal design choices is to make all users engage with all fundamental aspects of responsible AI for policy, not all users need to read all the information in the different sections. To accommodate the different needs, some information can be made available in the collapsible section. Or through linking to external resources.



Opportunities
What is AI for policy and how can it contribute to improved policy preparation?

How can AI help you make better policy?
Artificial intelligence or AI is a collection of smart tools and methods that can be used to create better policies. For example, by supporting

Main question:
- How could the extra thinking power of AI help to achieve better policy?

Choices & challenges
How should you use AI for policy and what are important considerations and risks?

How do you use AI responsibly?
When using AI tools, it is important that you comply with the law and policy.

Main questions:
- Does the desired AI application comply with law and policy?
- What steps do you take to limit the risks of (the use of) AI?
- What consideration do you make when choosing between opposing ethical values?

Expertise
What applications are available and where can you find the expertise needed to develop and use AI?

What tools are available?
[In development]
...

Where can I find relevant knowledge?
AI teams
Knowledge institutes
...

Tools & tests
Which tools can you use for responsible AI use? Which tests should be applied when implementing AI?

Tools & toetsen
Toolbox
Code of Good Digital Public Management
Ethical dialogue
...

Toetsen
IAMA - Impact Assessment Mensenrechten en Algoritmes
DPIA - Data Protection Impact Assessment
...

Figure 23. Illustration of the layering of complexity

4.4.3 Design Sketches

Throughout the project, a virtual design concept and some partial sketches were developed. In the transformation (Frame innovation) and concept (ViP) these initial ideas and the design vision are developed further to materialise a concrete design concept. These design sketches were made using the common design methods of individual brainstorming and HKJ's. Ideating the structure, content, and form of the design (see Figure 28). The resulting prototype design is discussed in the coming subsection.

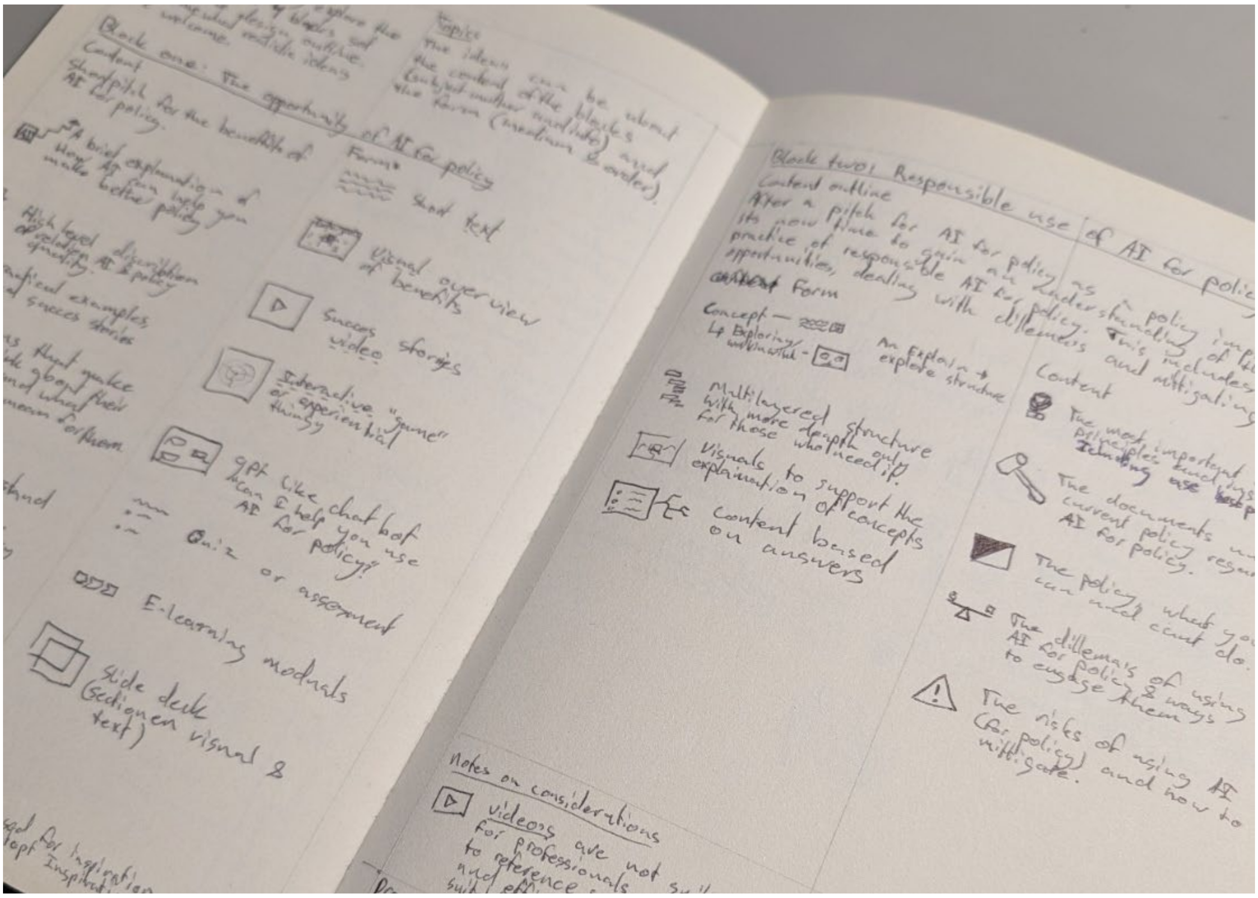

Block one: The opportunity of AI for policy
Content
Shortpitch for the benefits of AI for policy.
A brief explanation of how AI can help you make better policy
High level description of relation AI & policy
Practical examples and succes stories
Topics
The ideas can be about the content of the blocks (subject matter and info) and the form (medium & order).
Form
Short text
Visual overview of benefits
Succes video stories
Interactive "game" or experiential thingy
gpt like chat bot "can I help you use AI for policy?"
Quiz or assesment
E-learning moduals
Slide deck (Sectionen visual & text)
Notes on considerations
videos are not suit for professionals to reference
Block two: Responsible use of AI for policy
Content outline
After a pitch for AI for policy as a policy im... its now time to gain an understanding of the practice of responsible AI for policy. This includes opportunities, dealing with dillemmas and mitigating
Form
Concept — Exploring
An Explain + explore structure
Multilayered structure with more depth only for those who need it.
Visuals to support the explaination of concepts
Content based on answers
Content
The most important principles
The documents current policy regarding AI for policy.
The policy what you can and cant do.
The dillemma's of using AI for policy & ways to engage them
The risks of using AI (for policy) and how to mittigate.

Figure 24. Design ideation brainstorm

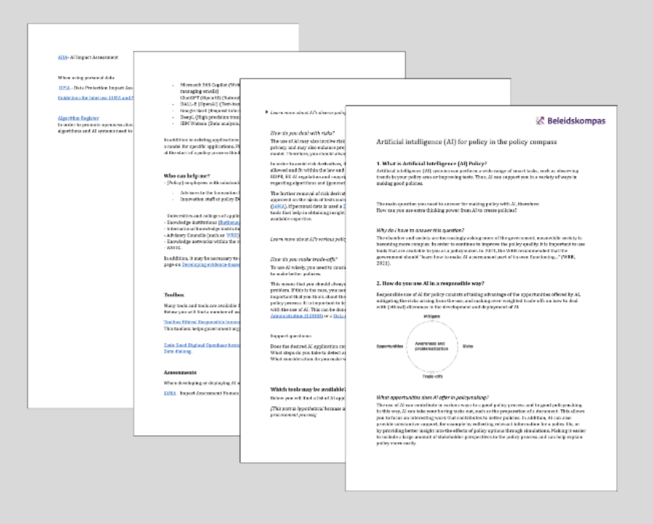
4.5 Design prototype

This section first sets out how the prototype’s structure, content and form were built up. Before diving into the prototype (see Figure 34) itself in the next subsection.

4.5.1 Design of the Prototype

4.5.1.1 Structure of the prototype

The main goal of the content order is to cater to the different user groups and their needs. A leading consideration is that the design should make those not as familiar with AI (for policy) more curious. This necessitates a low barrier of entry, with an introduction that highlights the positive sides of using AI (the opportunities). Following this, the design builds towards increasing complexity and completeness, including the practical implementation of AI, the dilemmas in the use of AI for policy, and the risks and robust risk mitigation tools.



Beleidskompas

Artificial intelligence (AI) for policy in the policy compass

1. What is Artificial Intelligence (AI) Policy?

Artificial intelligence (AI) systems can perform a wide range of smart tasks, such as observing trends in your policy area or improving texts. Thus, AI can support you in a variety of ways in making good policies.

The main question you need to answer for making policy with AI, therefore:
How can you use extra thinking power from AI to create policies?

Why do I have to answer this question?

The chamber and society are increasingly asking more of the government, meanwhile society is becoming more complex. In order to continue to improve the policy quality, it is important to use tools that are available to you as a policymaker. In 2021, the WRR recommended that the government should "learn how to make AI a permanent part of its own functioning..." (WRR, 2021).

2. How do you use AI in a responsible way?

Responsible use of AI for policy consists of taking advantage of the opportunities offered by AI, mitigating the risks arising from the use, and making well-weighted trade-offs on how to deal with (ethical) dilemmas in the development and deployment of AI.

What opportunities does AI offer in policymaking?

The use of AI can contribute in various ways to a good policy process and to good policymaking. In this way, AI can take your boring tasks out, such as the preparation of a document. This allows you to focus on interesting work that contributes to better policies. In addition, AI can also provide substantive support, for example by collecting relevant information for a policy file, or by providing better insight into the effects of policy options through simulations. Making it easier to include a large amount of stakeholder perspectives in the policy process and can help explain policy more easily.

Figure 25. The design prototype

4.5.1.2 Content of the prototype

The prototype design consists of four core blocks, with distinct functions in the users' journey. From developing initial curiosity to bringing together the resources and support needed for larger AI development and implementation projects, catering to a wide group of potential users. The four blocks are the following:

The first function: AI for policy is an initial pitch for AI for policy as a policy improvement aid. Relating AI for policy to policy quality and craftsmanship. Bringing the benefits of AI for policy to the forefront, with a brief explanation of how AI can help the user in making better policy. To engage users' interests and foster curiosity, support questions are included that make people think about what AI could mean for them in their practice. The section is aimed at capturing the attention of all users. It is made clear and easy to understand, positive, and applicable to almost all employees with a function in the policy development cycle.

The second function: Responsible use of AI for policy aims to help users gain an understanding of the practice. Responsible AI for policy. This includes making use of opportunities, dealing with dilemmas, and mitigating risks. This block contains the most important. Universal principles and insights. Including best practices for opportunities. The dilemmas of using AI for policy, and ways to engage them. And the risks of using AI for policy, and how to mitigate them. This includes the policy on what you can and cannot do within the Ministries, underpinned by the documents related to these policies.

Function three: Available applications and expertise. This part is allocated to practical steps needed for the use of AI in a user's project workflow. Directing users towards available tools and expertise, and knowledge needed for successful adoption. Including the allowed AI applications.

Who can help you with the introduction? Of AI applications. And external partners for expertise and knowledge, like universities.

Function four: Tools and tests - the last part of the prototype design includes tools and tests that are useful or obligatory. Or the use of AI applications for policy. This part includes tools for smaller use cases and big or risky projects. Obligated tests have to be performed before use. As well as tools for supporting the exploration of (ethical) dilemmas.

Selection of sources
The prototype includes multiple references, both sources of the content and for further information. For the user, it is important to trust a reference; this can be based on familiarity and authority. The included references are from and to familiar institutions like governmental agencies, WRR, AP, OECD, EU, etc. These are preferred even when similar or slightly higher-quality materials on a topic are available elsewhere.

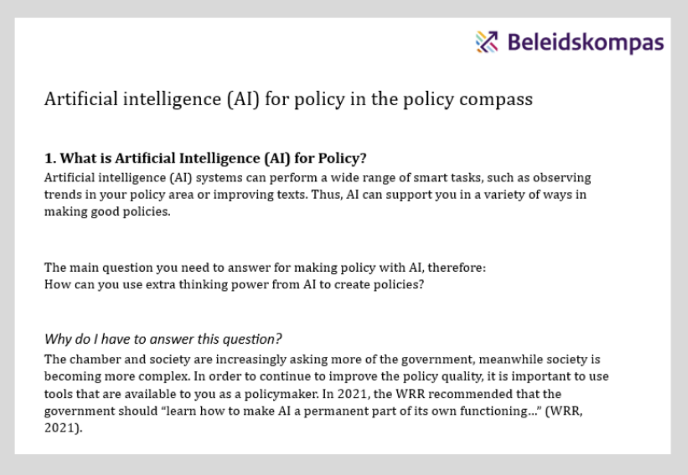
4.5.1.3 Form of the prototype

Due to time constraints and given the evaluation's main focus on the content and its order, the prototype mock-up is made in Word. The prototype includes two collapsible sections in which more examples of risks and opportunities can be found, denoted by "*Learn more about…*". And in-text clickable links to resources.

4.5.2 The Design Prototype

The prototype is to be imagined as a sub-page in the Policy Compass. The prototype description below is an English version of the original Dutch prototype used during the evaluation sessions. The Dutch text is translated to English locally using the MarianMT model (Helsinki-NLP, 2021). The original Dutch prototype can be found in Appendix 4.

The prototype starts with an introduction to AI for policy. Explaining the relationship between AI for policy and policy quality. The support question is meant to make users connect their practice and needs.



Beleidskompas

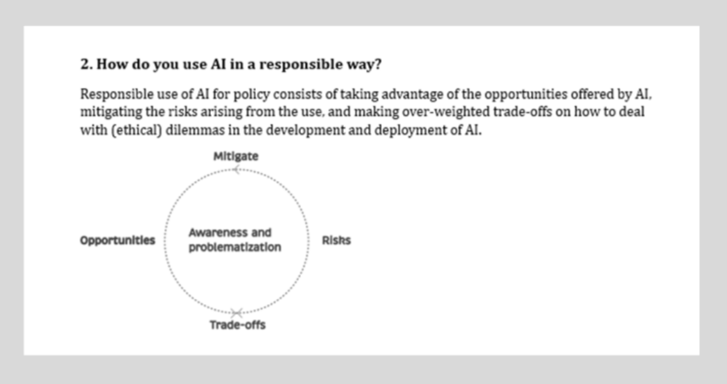
Artificial intelligence (AI) for policy in the policy compass

1. What is Artificial Intelligence (AI) for Policy?
Artificial intelligence (AI) systems can perform a wide range of smart tasks, such as observing trends in your policy area or improving texts. Thus, AI can support you in a variety of ways in making good policies.

The main question you need to answer for making policy with AI, therefore:
How can you use extra thinking power from AI to create policies?

Why do I have to answer this question?
The chamber and society are increasingly asking more of the government, meanwhile society is becoming more complex. In order to continue to improve the policy quality, it is important to use tools that are available to you as a policymaker. In 2021, the WRR recommended that the government should "learn how to make AI a permanent part of its own functioning..." (WRR, 2021).

Figure 26. Prototype section 1. What is Artificial Intelligence (AI) for policy?



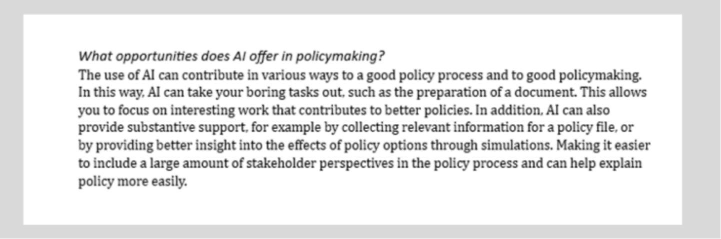
2. How do you use AI in a responsible way?

Responsible use of AI for policy consists of taking advantage of the opportunities offered by AI, mitigating the risks arising from the use, and making over-weighted trade-offs on how to deal with (ethical) dilemmas in the development and deployment of AI.

Figure 27. Prototype section 2. How do you use AI in a responsible way?

The introduction to AI for policy is followed by an introduction to the fundamentals of responsible AI.

A summary of the opportunities that AI could bring to the policy development process.

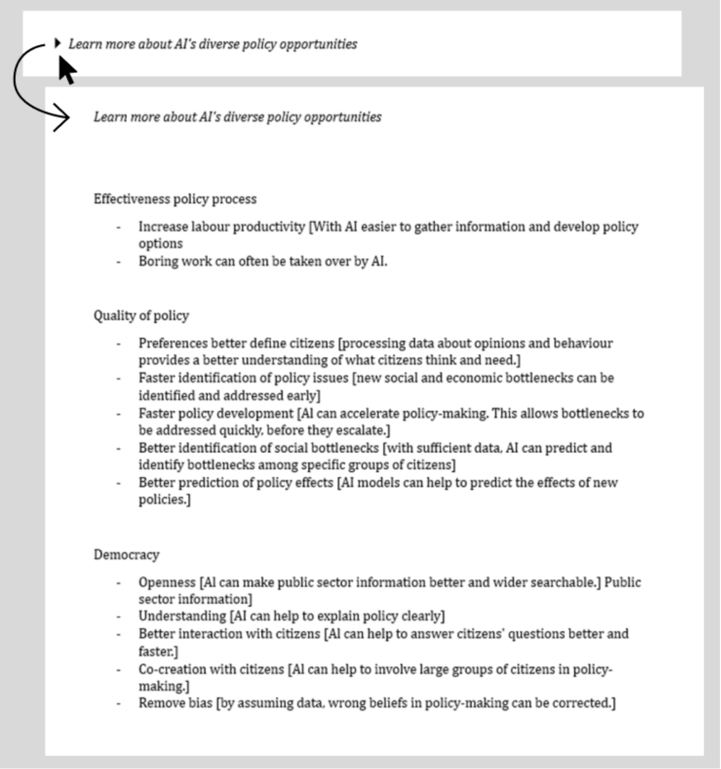


What opportunities does AI offer in policymaking?

The use of AI can contribute in various ways to a good policy process and to good policymaking. In this way, AI can take your boring tasks out, such as the preparation of a document. This allows you to focus on interesting work that contributes to better policies. In addition, AI can also provide substantive support, for example by collecting relevant information for a policy file, or by providing better insight into the effects of policy options through simulations. Making it easier to include a large amount of stakeholder perspectives in the policy process and can help explain policy more easily.

Figure 28. The opportunities of AI for policy

By clicking on "Learn more about AI for policy's diverse opportunities", examples of the various benefits AI can bring to the policy development process and the resulting policy become visible.



▸ *Learn more about AI's diverse policy opportunities*

Learn more about AI's diverse policy opportunities

Effectiveness policy process

- Increase labour productivity [With AI easier to gather information and develop policy options
- Boring work can often be taken over by AI.

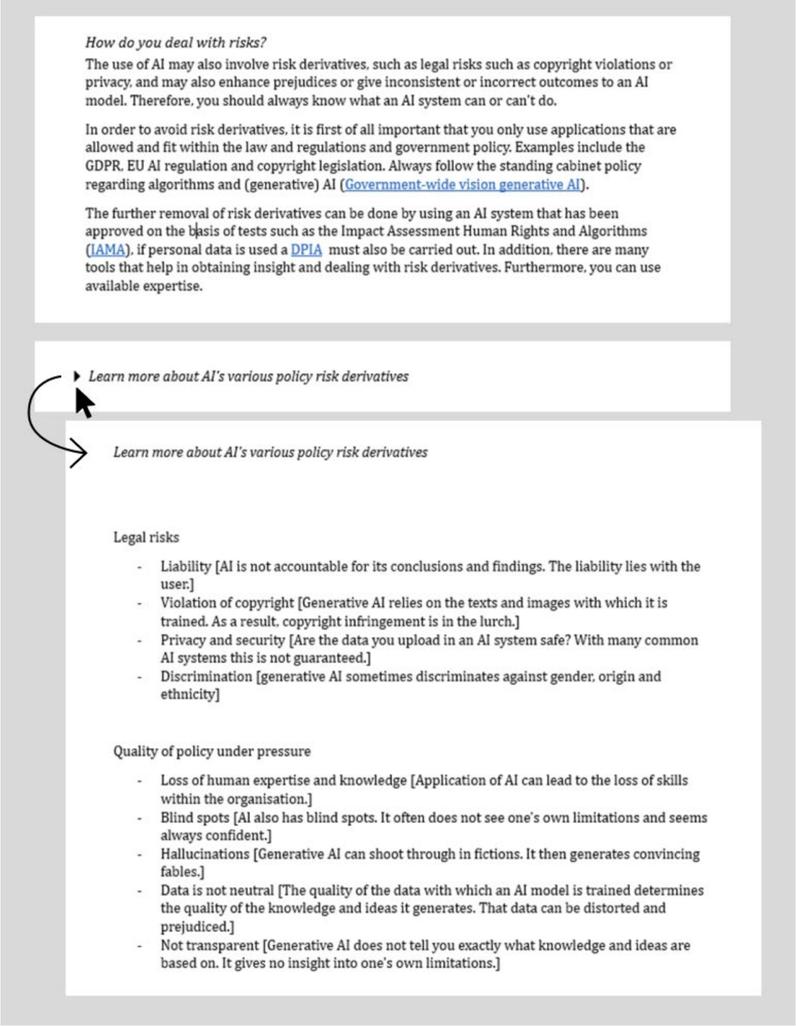
Quality of policy

- Preferences better define citizens [processing data about opinions and behaviour provides a better understanding of what citizens think and need.]
- Faster identification of policy issues [new social and economic bottlenecks can be identified and addressed early]
- Faster policy development [AI can accelerate policy-making. This allows bottlenecks to be addressed quickly, before they escalate.]
- Better identification of social bottlenecks [with sufficient data, AI can predict and identify bottlenecks among specific groups of citizens]
- Better prediction of policy effects [AI models can help to predict the effects of new policies.]

Democracy

- Openness [AI can make public sector information better and wider searchable.] Public sector information]
- Understanding [AI can help to explain policy clearly]
- Better interaction with citizens [AI can help to answer citizens' questions better and faster.]
- Co-creation with citizens [AI can help to involve large groups of citizens in policy-making.]
- Remove bias [by assuming data, wrong beliefs in policy-making can be corrected.]

Figure 29. Learn more: About opportunities



How do you deal with risks?

The use of AI may also involve risk derivatives, such as legal risks such as copyright violations or privacy, and may also enhance prejudices or give inconsistent or incorrect outcomes to an AI model. Therefore, you should always know what an AI system can or can't do.

In order to avoid risk derivatives, it is first of all important that you only use applications that are allowed and fit within the law and regulations and government policy. Examples include the GDPR, EU AI regulation and copyright legislation. Always follow the standing cabinet policy regarding algorithms and (generative) AI (Government-wide vision generative AI).

The further removal of risk derivatives can be done by using an AI system that has been approved on the basis of tests such as the Impact Assessment Human Rights and Algorithms (IAMA), if personal data is used a DPIA must also be carried out. In addition, there are many tools that help in obtaining insight and dealing with risk derivatives. Furthermore, you can use available expertise.

▸ *Learn more about AI's various policy risk derivatives*

Learn more about AI's various policy risk derivatives

Legal risks

- Liability [AI is not accountable for its conclusions and findings. The liability lies with the user.]
- Violation of copyright [Generative AI relies on the texts and images with which it is trained. As a result, copyright infringement is in the lurch.]
- Privacy and security [Are the data you upload in an AI system safe? With many common AI systems this is not guaranteed.]
- Discrimination [generative AI sometimes discriminates against gender, origin and ethnicity]

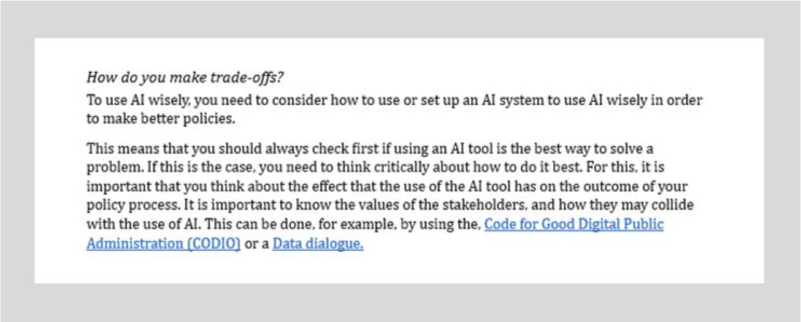
Quality of policy under pressure

- Loss of human expertise and knowledge [Application of AI can lead to the loss of skills within the organisation.]
- Blind spots [AI also has blind spots. It often does not see one's own limitations and seems always confident.]
- Hallucinations [Generative AI can shoot through in fictions. It then generates convincing fables.]
- Data is not neutral [The quality of the data with which an AI model is trained determines the quality of the knowledge and ideas it generates. That data can be distorted and prejudiced.]
- Not transparent [Generative AI does not tell you exactly what knowledge and ideas are based on. It gives no insight into one's own limitations.]

A summary of the risks related to using AI in the policy process.

By clicking on “Learn more about AI's various risks”, examples of the various risks of using AI in the policy process become visible.

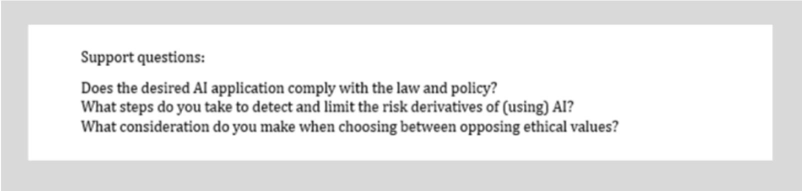
A section on making trade-offs when faced with (ethical) dilemmas. Including hyperlinks to resources.



How do you make trade-offs?

To use AI wisely, you need to consider how to use or set up an AI system to use AI wisely in order to make better policies.

This means that you should always check first if using an AI tool is the best way to solve a problem. If this is the case, you need to think critically about how to do it best. For this, it is important that you think about the effect that the use of the AI tool has on the outcome of your policy process. It is important to know the values of the stakeholders, and how they may collide with the use of AI. This can be done, for example, by using the, Code for Good Digital Public Administration (CODIO) or a Data dialogue.

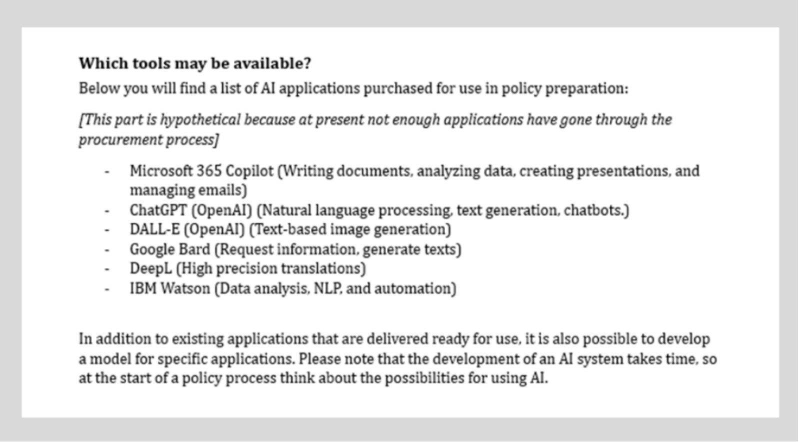


Support questions:

Does the desired AI application comply with the law and policy?
What steps do you take to detect and limit the risk derivatives of (using) AI?
What consideration do you make when choosing between opposing ethical values?

The second function ends with three support questions to help users explore their use case for an AI application in a critical way.

A (hypothetical) list of procured AI applications, including a description of what they are used for.



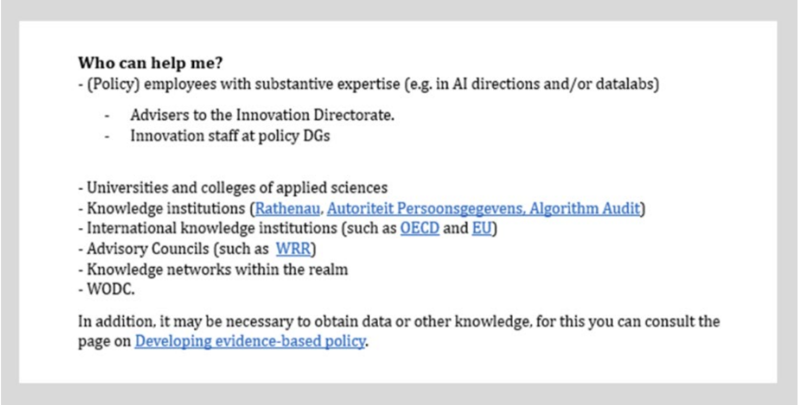
Which tools may be available?

Below you will find a list of AI applications purchased for use in policy preparation:

[This part is hypothetical because at present not enough applications have gone through the procurement process]

- Microsoft 365 Copilot (Writing documents, analyzing data, creating presentations, and managing emails)
- ChatGPT (OpenAI) (Natural language processing, text generation, chatbots.)
- DALL-E (OpenAI) (Text-based image generation)
- Google Bard (Request information, generate texts)
- DeepL (High precision translations)
- IBM Watson (Data analysis, NLP, and automation)

In addition to existing applications that are delivered ready for use, it is also possible to develop a model for specific applications. Please note that the development of an AI system takes time, so at the start of a policy process think about the possibilities for using AI.



Who can help me?

- (Policy) employees with substantive expertise (e.g. in AI directions and/or datalabs)

  - Advisers to the Innovation Directorate.
  - Innovation staff at policy DGs

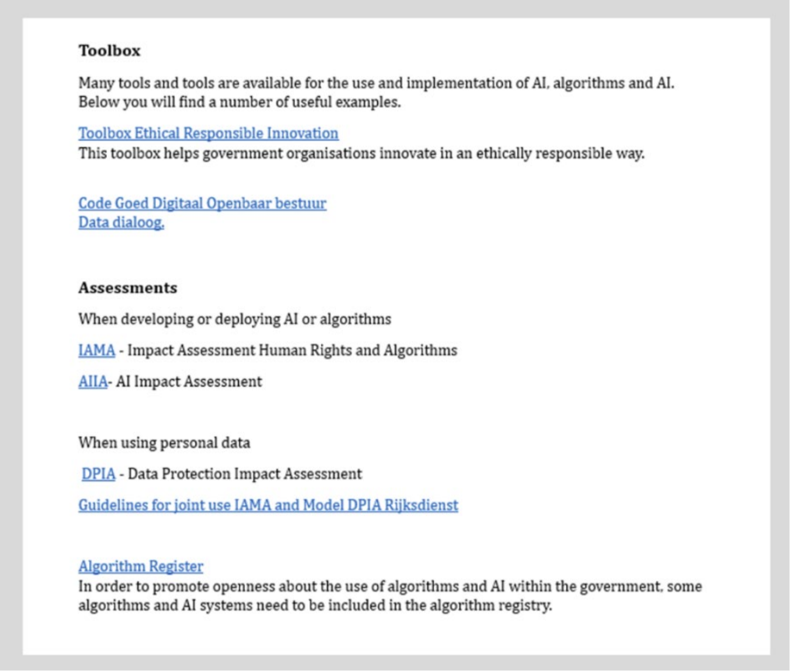
- Universities and colleges of applied sciences
- Knowledge institutions (Rathenau, Autoriteit Persoonsgegevens, Algorithm Audit)
- International knowledge institutions (such as OECD and EU)
- Advisory Councils (such as WRR)
- Knowledge networks within the realm
- WODC.

In addition, it may be necessary to obtain data or other knowledge, for this you can consult the page on Developing evidence-based policy.

A list of internal and external parties that can offer support and expertise. Including hyperlinks to the relevant web pages and materials.

A toolbox with hyperlinks to tools and reference materials for responsible AI in government.

And the assessment and registers that might be required for the implementation of AI in the policy process.



Toolbox

Many tools and tools are available for the use and implementation of AI, algorithms and AI. Below you will find a number of useful examples.

Toolbox Ethical Responsible Innovation
This toolbox helps government organisations innovate in an ethically responsible way.

Code Goed Digitaal Openbaar bestuur
Data dialoog.

Assessments

When developing or deploying AI or algorithms

IAMA - Impact Assessment Human Rights and Algorithms

AIIA- AI Impact Assessment

When using personal data

DPIA - Data Protection Impact Assessment

Guidelines for joint use IAMA and Model DPIA Rijksdienst

Algorithm Register
In order to promote openness about the use of algorithms and AI within the government, some algorithms and AI systems need to be included in the algorithm registry.

5. Evaluation

5.1 Evaluation goal

The purpose of the evaluation is to investigate how effective the prototype design is at achieving the design goals and observe whether the design mechanisms work as intended. In addition, the evaluation was set up to better understand user preferences and elicit recommendations for how the design could be improved.

Evaluation of whether the prototype strikes the right balance between risk avoidance and mitigation through rule and guideline enforcement, and warning of the risks. And optimise the benefits of using AI in policy development and the development of AI capacity in the organisation through exciting users, and inducing AI curiosity. The balance struck in the prototype is used as a starting point against which to react. Both as a vision for the design and the effect the prototype is expected to have on the different users and stakeholders, and their interaction with AI in their practices.

The first design goal evaluated is providing *comprehensive guidance*, creating guardrails by upholding existing rules, regulations and frameworks, as well as providing available tools (completeness, correctness, clarity & actionable). The second design goal is to create *AI curiosity* to facilitate the development of soft AI capacity.

To answer the

RQ 1: What does responsible use of AI entail in policy preparation?

RQ 2: How should a tool for policy preparation professionals be designed to effectively incorporate advice?

In oder to anwer the main research question of this thesis:

RQ 0: *How can we design practical tools to guide the use of AI in the governmental policy development process?*

5.2 Evaluation Session Setup

The evaluation sessions were conducted in person with individual participants, in a meeting room at the governmental organisation. During the session, the participants were first asked some questions to gauge their familiarity with (responsible) AI and their understanding of how it could potentially be of value to their work. The participants were provided with the prototype on a laptop. To make all participants engage with the prototype from the perspective of an end user, participants were given a brief case description of a policy officer, a task a hypothetical user of the design would try to perform using the tool. Together with the instruction, they should vocalise their thought process. Other than that, the participants were given the freedom to explore and comment on the prototype in whatever way they preferred. The second half of the session consisted of a semi-structured interview about their experience with the prototype, the fit of the prototype with the governmental organisation, and its user groups.

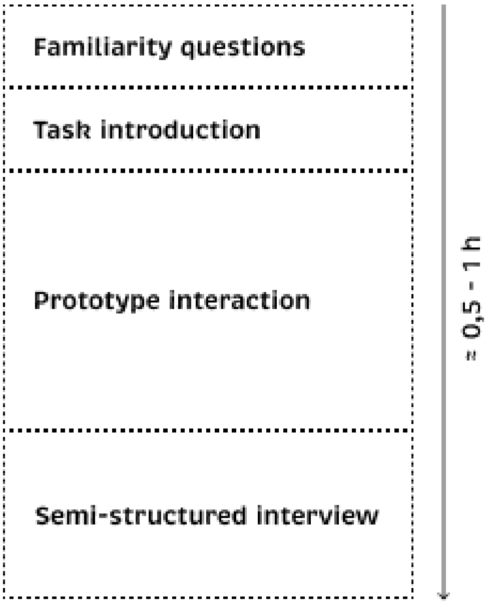


Figure 30. Evaluation session structure

5.3 Main data collected

Perceived strengths and recommendations

Throughout the evaluation session, the participants were encouraged to express features they appreciated, as well as things they did not like and ideas and recommendations for the improvement of the design. A few comments were made on the strengths of the prototype design. A large number of comments were elicited relating to improvements to the prototype, to make it more effective and fitting with the needs and wants of users. Including comments about changes the participants would like to see in the prototype, indicated by wanting elements to be removed or changed or by expressing ideas for the inclusion of things or changes to the form of the design.

Effectiveness

In the semi-structured interview, participants were asked what effect they would expect the tool to have on themselves and various other user groups, based on experiences in their current and previous roles within ministries.

To gain insight into the prototype's ability to induce AI curiosity, signs of curiosity, like seeking additional information by asking questions about the subjects discussed in the prototype, were observed during the sessions.

5.3 Participants

The evaluation sessions were conducted with 9 participants from thegovernmental organisation. Participants were recruited through an invitation after the presentation at the organisation and through personal requests via contacts. Participants additionally included people involved in the support of the project.

Most of the participants were active in the section in which the project was done in collaboration, and were therefore not actively involved in the development of policy. The background of participants in prior roles was diverse and included many policy-development positions. Participants of a committee on the Policy Compass from other ministries were also invited, but this did not result in participation.

5.4 Analysis Methods

Coding of the transcripts

The recordings of the evaluation sessions are transcribed[4]. Before thematic analysis is used to interpret the evaluation sessions. In the first round of coding, descriptive codes were applied to utterances that seemed potentially relevant to evaluating the design goals, answering both the design question, the evaluation of the existing prototype, and what can be learned more broadly by the scientific and professional field. These descriptive codes are grouped by the type of information they include and what question or evaluation topic they are related to. Within these groupings, code categories were developed based on closely related codes. For the recommendations, these code groups were developed further into themes that more succinctly capture the trends in the underlying codes.

[4] Using Wispher, run the model locally, with responsible AI practices and data safety in mind.

6. Results

This chapter describes the results from the evaluation sessions of the proposed design prototype described in previous chapters. Firstly, an overview of the changes participants would like to see in the design [form and function]. This gives more granular insight into the diverging needs and wants of the participants and users. The results are presented on how well the prototype achieves its design goals of inducing AI curiosity and guiding users to consider the responsible use of AI in policy development, and whether this aligns with the needs and desires of participants.

6.1 Perceived Strengths and Recommendations

6.1.1 Perceived strengths

Most participants expressed things they appreciated about the prototype, while some did not have anything they liked about the prototype, even when asked directly. Something appreciated by a few participants is the prototype support questions, helping users actively think and reflect. One participant expressed that she valued that the prototype linked responsible AI use to improved policy quality, as policy quality is a shared value among policy developers.

6.1.2 Recommendation Themes

The identified improvement themes are accessible, inspiring, and meticulous. Recommending how the design should evolve:

- *Accessible*: having a low barrier to entry and high ease of use;
- *Inspiring*: being appealing to the user and motivating the user to take action;
- *Meticulous*: being comprehensive and precise, and requiring the user to be thorough.

These themes are related to categories and code groups. see the code tree of the evaluation session results in Figure 35. Exemplary comments and descriptions of the codes can be found in Appendix 3. In the following sub-sections, the themes and the categories they are built on will be explored further.

6.2.2.1 Accessible

The theme *accessible* builds on a group of recommendations for making the tool and responsible AI practices by means of the tool, easy to approach, understand, and use. This theme is connected to the categories *supporting* and *actionable*.

Supporting | includes recommendations relating to support by the tool itself or experts (via the tool). Taking the user by the hand and providing tangible aid and assistance to users. Helping them get through all the steps of the process. To reduce barriers in the exploration and implementation of responsible AI (practices), and utilise and propagate the available expertise in the organisation.

Actionable | Allowing or enabling the user to take action in exploring or implementing responsible AI (practices). For instance, by including concrete steps users should take and less theoretical information that is easy to adapt to the needs of users. Or avoiding users having to read information that is not relevant to their individual situation. And avoiding users having to do further research themselves.

6.2.2.2 Inspiring

The theme *inspiring* bundles recommendations relating to functions and qualities that make the tool uplifting, energising activating. This theme is connected to the categories *appealing, motivating* and *actionable*.

Appealing | includes recommendations related to making the form of the tool appealing as well as making its content appeal to users' own interests and positive emotions. *(also linked to accessibility).* The form of the tool should be attractive to the user, and the content should appeal to their needs and situation, for instance, through relatable examples and success stories.

Motivating | includes recommendations related to making the tool drive people to take action, for instance, by making the opportunities and benefits more prominent. Giving users positive energy in relation to the topic of responsible AI and its implementation in their practice. Through examples and clarification of what is in it for them. Appealing to the user's experiences through relatable examples, appealing to the user's values and interests, and appealing to look at and use.

Actionable | see under accessible.

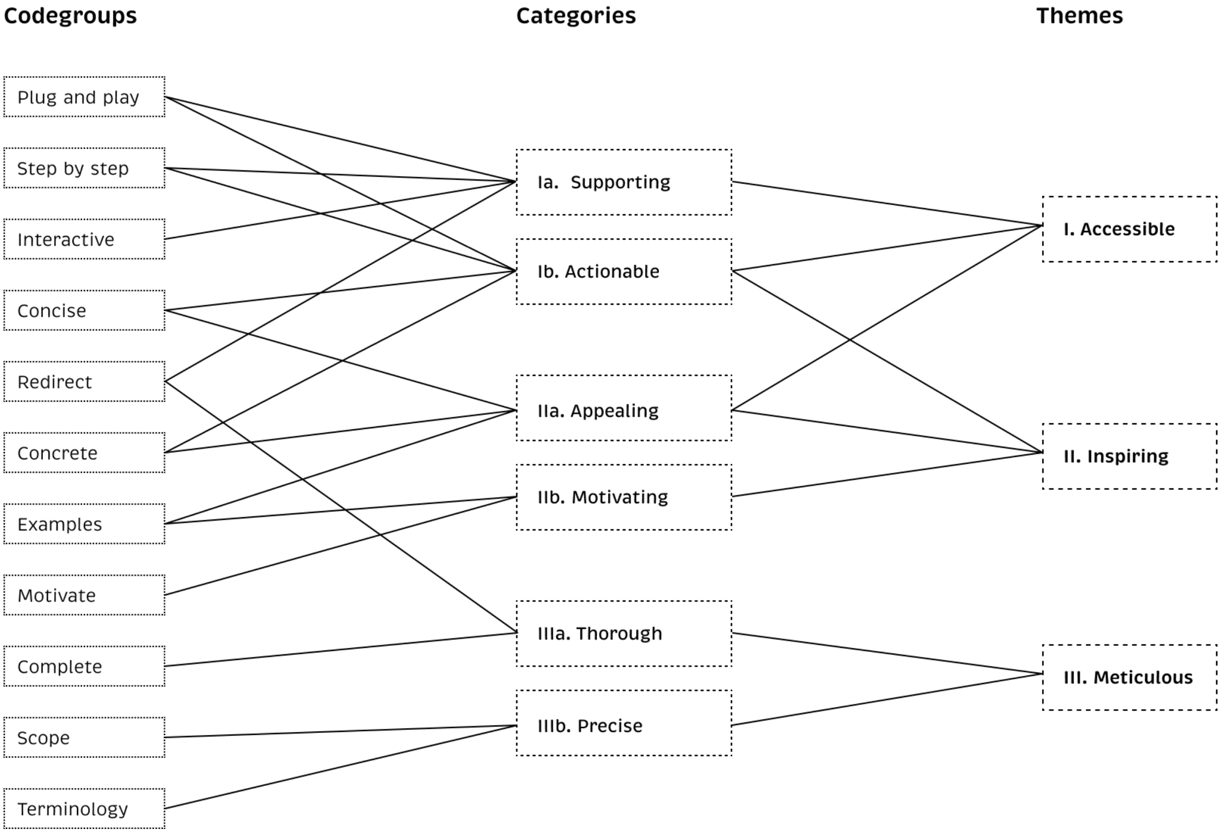


Figure 31. Code tree of the evaluation session results

6.2.2.3 Meticulous

This theme brings together recommendations about the inclusion Recommendations to expect of users that they need to understand and do more themselves, or to include more detailed information on a broader range of related subjects. As well as critical feedback on the terminology used in the prototype, and caution for the importance of using the right language and framing.

Thorough | brings together recommendations that call for a broader inclusion-related content and comprehensive explanations, as well as explaining to the user what is and what isn't included and why. Favouring completeness over conciseness.

Precise | captures codes relating to the scope (what should be included, and precise communication about what is included in the scope of the text) and terminology (what terms should be used, both regarding accuracy and the tone of the text).

6.2 Effectiveness

6.2.1 Effect Expectation

The expectations are grouped into five categories, ranging from avoiding AI altogether, being tempered, neutral, and curious, to being motivated to adopt AI. The majority of participants expect that users of the design, as presented in the prototype, would take away a cautious attitude towards AI for policy, due to due to the prominent attention to the risks involved in the use of AI in the prototype. Some participants expressed that the prototype would make them, or policy officers, wary about the risks. Making them only interested in using AI when all risks are removed by experts and leadership. On the other side of the spectrum, some participants expressed that they, or policy officers, could get motivated and be open to taking some calculated risks in trying to use AI responsibly, even if not fully within the official guidelines. Regarding the effect the tool would have on the attitudes of managers noted that the perceived focus on risk could make them hesitant, but that the inclusion of the design in the Policy Compass could spark curiosity, leading to encouragement of exploration by junior team members, with the effect of the design expected to depend greatly on their personality and familiarity with technology.

6.2.2 Inducing Curiosity

Four of the participants spontaneously offered new to them ways AI could be used in their practice during their interaction with the prototype. Notably, these ideas were about their own practice, and not related to the task in the hypothetical scenario.

In these participants, the prototype is effective at inciting curiosity toward AI for policy. Mainly through the motivating effect of including application examples and benefits. Another sign of emerging curiosity was participants seeking further information on the topic. Most participants asked questions about responsible AI for policy, beyond clarification of the contents of the prototype.

7. Discussion

This chapter discusses what has been learned from the design and the evaluation of the prototype design, and how this contributes to the understanding of designs for guidance on AI for policy.

7.1 Discussion of Results

7.1.1 Discussion of prototype evaluation results

What Makes a Responsible AI tool Work
The key findings from the evaluation session are the recommendations themes that express how the participants would like to see the design evolve:

Accessibility related a low barrier to entry. Participants wanted support, either through the tool or from experts via the tool, and suggested making sure users didn't have to figure things out on their own or do further research elsewhere.

Inspiring is about making the tool feel energising. That included making the form more appealing and using examples that feel relevant to the user's own situation. Participants wanted the benefits of responsible AI to be more visible and framed in a way that motivates action.

Meticulous points to something else. It includes recommendations that ask for more detail, more explanation, and more precision, in both content and terminology. Participants wanted to know what's in and what's out, and why. Some comments were about being cautious with language and framing, especially in a policy context where tone matters.

The overlap and duality of the themes and categories and even code groups highlight that there are broad tendencies that could feed into meta-narratives on responsible AI for policy and how this can be facilitate trough a design for guidance. Conversely, the themes accessible and inspiring clash with the theme meticulous, highlight tensions. These tensions are more explicit at the level of categories where, for example, *thoroughness* and *accessibility* are directly at odds with each other.

Effectiveness of the prototype: Curiosity and Cautiousness
The results of the prototype evaluation show that the prototype can spark curiosity in some users. Four participants spontaneously offered new ways to them ways AI could be used in their practice, not tied to the task they were given. Showing participants curiously applying insights from the prototype to their own work. In these participants, the prototype appears to motivate a form of constructive curiosity, supported by the inclusion of a few relatable AI use cases and their potential benefits. This is supported further by the fact that most participants asked questions about responsible AI for policy, inquiring beyond the clarification of things in the prototype. However, the majority of expectations were that users would take away a cautious attitude. Some participants noted that the design could make them or other policy officers more hesitant about AI because the risks are presented clearly, and the responsibility for implementation is not yet settled.

In that sense, the design does draw attention to responsible use, but in a way that might reinforce a "wait and see" attitude, at least in more risk-averse functions or teams. Others said it could have the opposite effect, making them or other policy officers more open to taking calculated risks or trying things out, even if not fully within the official rules.

Feedback on the Design Vision
In addition to the differences in use and effect, participants reflected on the posture of the design, the message it appears to carry.
Attempting to answer *RQ 1: What does responsible use of AI entail in policy preparation?* The prototype is based on a particular vision of responsible AI, one that aims to weigh risks and opportunities and support responsible implementation without prescribing a single path. But the results suggest not everyone reads it that way.

These results are solely based on comments made about the goal that the design should be trying to achieve. Some recommendations push the design towards a different vision or are indirect critiques of the stance taken in the design. Several participants made comments on the posture the tool should take on the spectrum from a progressive stance focused on pushing AI implementation to utilise its benefits, and a conservative stance that focuses on warning of the risks and stringent reinforcement of rules and guidelines. These result in five categories: motivate, catalyse, balance, individual responsibility, and caution. The balanced category includes the greatest number of codes. Some participants made separate comments belonging to both conservative and progressive categories, leaning towards preferring a more balanced stance. This suggests that while the prototype intends to occupy a balanced position, that balance might need to be made more explicit or legible to users.

Balancing Encouragement and Restraint
The conservative effect that participants of the evaluation session expect the design to have is in line with the more balanced approach that is based on curiosity and personal responsibility. To those who do not become curious enough to be willing to explore and weigh the risks against the benefits, the tool should be discouraging. Opposite, the design should make as many people curious enough to invest their effort to explore, and potentially responsibly implement AI. This balancing act between thoroughness and the other two themes is also evident in the recommendations.
This effect expectation is in line with the general goal as expressed by the organisation and is consistent with the intended role of the Policy Compass. However, the experience of a restricting or tempering effect could hamper the development of AI curiosity. It is therefore important to strike a balance between enough security and rule enforcement and a playful lightness in the experience of using the tool.

7.1.2 Reflection on Design Principles

The design vision and design of the prototype are based on a number of principles developed through iterative brainstorming and grounded in the literature and fieldwork. Answering *RQ 2: How should a tool for policy preparation professionals be designed to effectively incorporate advice?, These* principles formulate what the design should do and propose a strategy for how to do so. The results of the evaluation show how these principles played out in practice and where tensions emerged.

AI curiosity
A mechanism the design intends to utilise is the mental pathway of curiosity catalysing learning, and moving from passive understanding to acting. The prototype appears effective at inciting curiosity toward AI for policy. Four participants spontaneously offered new to them ways AI

could be used in their practice. Notably, these ideas were about their own practice, and not related to the task in the hypothetical scenario. Most participants asked questions about responsible AI for policy, beyond clarification of the contents of the prototype. Engagement with AI, resulting from AI curiosity, can be a way to ensure critical thinking about AI use, distribute responsibility, and create resilience.

Trust and responsibility
A balance is needed that ensures (enough) risk encapsulation while not crushing the inherently curious spirit of policy officers. The results suggest that the design leans towards caution. Some participants expressed that the prototype would make them, or policy officers, wary about the risks, making them only interested in using AI when all risks are removed by experts and leadership. On the other side of the spectrum, some participants expressed that they, or policy officers, could get motivated and be open to taking some calculated risks in trying to use AI responsibly, even if not fully within the official guidelines.

There are clear links between different ideas about the goal the design should be pursuing and the recommendations. A more conservative, cautious approach to responsibility sharing and reinforcement of rules and guidelines necessitates more effort and restraint on the end of the policy developers and limits the benefits it might have to them. Contrary to that, the more progressive approach that limits responsibility sharing and provides policy developers with more benefits without a lot of effort is favoured by more of these very end-users of the design. This is consistent with earlier findings about Policy Compass, where users feel that it is designed from a perspective that is not theirs.

Anchor function
There is an explosion of different organisations creating recommendations, guidelines and tools. This disjointed nature of the efforts, resources, and expertise relevant to the responsible implementation of AI in the policy process leads to stacking uncertainty and complexity. Participants made recommendations connected to support and terminology, pointing to a need for redirection to existing resources and clarification of what is and isn't included. The design can take on an anchor function [kapstok function], redirecting users to the information, tools, and expertise they require. The Policy Compass is a natural place for this, as it is central to the policy development process.

7.2 Relation to literature and academic context

A key gap in the literature is the need to move from theoretical and speculative perspectives to more situated and practical exploration (Zuiderwijk et al., 2021). This project aims to do just that by testing how a design intervention could support responsible AI in the early stages of policy development. The focus on implementation capabilities, particularly human and organisational AI capacity, complements the existing emphasis on technical capabilities already being developed in the governmental organisation. This responds directly to the strategic alignment gap flagged by van Noordt and Tangi (2023). A key contribution is the idea of AI curiosity as a precursor to AI capability. The prototype does not aim to teach everything at once, but to prompt questions, dialogue, and individual motivation, as a basis for distributed capacity-building. This connects to earlier work on strategic alignment (van Noordt & Tangi, 2023), but shifts the focus to soft capabilities, including user engagement and reflective decision-making.

While the literature notes that AI for policy is still relatively underexplored (Van Noordt & Misuraca, 2022), the results here suggest that low-threshold tools like the prototype can play a catalytic role. Especially generative AI, with its accessibility, can act as a gateway to broader forms of engagement. Its presence in the prototype aligns with the observation that generative models often serve as the first point of contact for policy officers, even if they are not the most impactful technologies long-term. The prototype aims to use this initial engagement to spark curiosity and scaffold further capacity development.

The design draws on institutional theory, not just in how it reflects formal procedures and policy, but in how it attempts to reshape informal cultural practices. Following Koppenjan and Groenewegen (2005), the prototype can be seen as a modest intervention in the larger institutional landscape, one that encourages renegotiation of norms, language, and responsibilities around AI use in policy-making.

The feedback on the stance of the design relates to themes from the literature on democratic governance and responsible AI. The design balances between motivating action and tempering uncritical enthusiasm, a tension reflected in literature on democratic risk, depoliticisation, and technocratic drift (König & Wenzelburger, 2020; Newman et al., 2021). While some participants wanted a stronger push toward innovation, others preferred caution and clearer guardrails. This reflects the value tensions described by Madan and Ashok (2023), particularly between automation and augmentation, and between transparency and performance.

Finally, the project reinforces the importance of embedding contestation and agonistic mechanisms in responsible AI design (Alfrink et al., 2023; Lowndes & Paxton, 2018). Participants' appreciation of support questions and their varied interpretations of the tool’s message illustrate the value of leaving space for pluralism and critical engagement. The aim is not to arrive at one stable definition of responsible AI, but to support a situated, evolving understanding that fits the democratic ethos and institutional dynamics of Dutch policy-making.

7.3 Future work

Future research

More research is needed to understand and empirically validate ways in which the use of AI systems in policy development can be made responsible, without overburdening policy professionals with responsibilities they are unable to fulfil, through lacking the conditions for moral responsibility in line with Sattlegger et al., (2022), and in the practical sense, based on the experience of practitioners.

8. Concluding remarks

Navigating the value tension in the use of AI for policy preparation requires the tension to be surfaced

Bibliography

Aktinson, P., & Hammersley, M. (1998). Ethnography and participant observation. *Strategies of Qualitative Inquiry. Thousand Oaks: Sage*, 248–261.

Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2023). Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3544548.3580984

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by Design: Towards a Framework. *Minds and Machines*, *33*(4), 613–639. https://doi.org/10.1007/s11023-022-09611-z

Alfrink, K., Keller, I., Yurrita Semperena, M., Bulygin, D., Kortuem, G., & Doorn, N. (2024). Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI. *She Ji: The Journal of Design, Economics, and Innovation*, *10*(1), 53–93. https://doi.org/10.1016/j.sheji.2024.03.003

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, *36*(2), 358–367. https://doi.org/10.1016/j.giq.2018.10.001

Arendt, H. (1990). *On revolution*. Penguin Books. (Original work published 1963)

Autoriteit Persoonsgegevens. (2024). *AI & Algorithmic Risks Report Netherlands Edition 3, summer 2024*. https://www.autoriteitpersoonsgegevens.nl/en/documents/ai-algorithmic-risks-report-netherlands-summer-2024

Bali, A. S., Howlett, M., Lewis, J. M., & Ramesh, M. (2021). Procedural policy tools in theory and practice. *Policy and Society*, *40*(3), 295–311. https://doi.org/10.1080/14494035.2021.1965379

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework[1]. *European Law Journal*, *13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics*, *33*(5), 946–967. https://doi.org/10.1080/01402382.2010.486119

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300271

Buchanan, R. (1992). Wicked Problems in Design Thinking. *Design Issues*, *8*(2), 5. https://doi.org/10.2307/1511637

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration*, *49*(7), 751–761. https://doi.org/10.1177/0275074019856123

Bunders, D. J., & Varró, K. (2019). Problematizing data-driven urban practices: Insights from five Dutch 'smart cities'. *Cities*, *93*, 145–152. https://doi.org/10.1016/j.cities.2019.05.004

Cairney, P. (2022). The myth of 'evidence-based policymaking' in a decentred state. *Public Policy and Administration*, *37*(1), 46–66. https://doi.org/10.1177/0952076720905016

Cambridge University Press. (n.d.a). *Curiosity*. Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/curiosity

Cambridge University Press. (n.d.b). *Responsible*. Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/responsible

Campion, A., Gasco-Hernandez, M., Jankin Mikhaylov, S., & Esteve, M. (2022). Overcoming the Challenges of Collaboratively Adopting Artificial Intelligence in the Public Sector. *Social Science Computer Review*, *40*(2), 462–477. https://doi.org/10.1177/0894439320979953

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2017). Artificial Intelligence and the 'Good Society': The US, EU, and UK approach. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-017-9901-7

Chen, Y.-C., Ahn, M. J., & Wang, Y.-F. (2023). Artificial Intelligence and Public Values: Value Impacts and Governance in the Public Sector. *Sustainability*, *15*(6), 4796. https://doi.org/10.3390/su15064796

Churchman, C. W. (1967). Guest Editorial: Wicked Problems. *Management Science*, *14*(4,), B141–B142.

Coeckelbergh, M., & Sætra, H. S. (2023). Climate change and the political pathways of AI: The technocracy-democracy dilemma in light of artificial intelligence and human agency. *Technology in Society*, *75*, 102406. https://doi.org/10.1016/j.techsoc.2023.102406

Dafoe, A. (2018). *AI governance—A research agenda.pdf*. Centre for the Governance of AI Future of Humanity Institute University of Oxford. fhi.ox.ac.uk/govaiagenda

Desouza, K. C., Dawson, G. S., & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons*, *63*(2), 205–213. https://doi.org/10.1016/j.bushor.2019.11.004

Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, *49*(7), 1043–1064. https://doi.org/10.1177/0095399714555751

Dorst, K. (2004). On the Problem of Design Problems—Problem solving and design expertise. *J. of Design Research*, *4*(2), 0. https://doi.org/10.1504/JDR.2004.009841

Dorst, K. (2015a). Frame Creation and Design in the Expanded Field. *She Ji: The Journal of Design, Economics, and Innovation*, *1*(1), 22–33. https://doi.org/10.1016/j.sheji.2015.07.003

Dorst, K. (2015b). *FRAME INNOVATION: Create new thinking by design*. MIT PRESS.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

ECP. (2018). *Artificial intelligence impact assessment*. https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf

European Parliament. (2023, June 1). *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

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Fatima, S., Desouza, K. C., Buck, C., & Fielt, E. (2022). Public AI canvas for AI-enabled public value: A design science approach. *Government Information Quarterly*, *39*(4), 101722. https://doi.org/10.1016/j.giq.2022.101722

Februari, M. (2023). *Doe zelf normaal: Menselijk recht in tijden van datasturing en natuurgeweld*. Prometheus-Bert Bakker.

Fischer, F., Miller, G., & Sidney, M. S. (Eds). (2007). *Handbook of public policy analysis: Theory, politics, and methods*. Routledge. https://doi.org/10.4324/9781315093192

Flinders, M., & Wood, M. (2014). Depoliticisation, governance and the state. *Policy & Politics*, *42*(2), 135–149. https://doi.org/10.1332/030557312X655873

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical

Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foucault, M. (1997). Polemics, politics, and problematizations. In P. Rabinow (Ed.), *Ethics: Subjectivity and truth* (Vol. 1, pp. 111–119). The New Press.

Frayling, C. (1993). *Research in art and design*. Royal College of Art.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, *5*(3), 143–164. https://doi.org/10.1037/dec0000068

Gonen, H., & Goldberg, Y. (2019). *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them* (No. arXiv:1903.03862). arXiv. https://doi.org/10.48550/arXiv.1903.03862

Gray, M., & Mcdonald, C. (2006). Pursuing Good Practice?: The Limits of Evidence-based Practice. *Journal of Social Work*, *6*(1), 7–20. https://doi.org/10.1177/1468017306062209

Gray Widder, D., West, S., & Whittaker, M. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4543807

Guenduez, A. A., & Mettler, T. (2023). Strategically constructed narratives on artificial intelligence: What stories are told in governmental artificial intelligence policies? *Government Information Quarterly*, *40*(1), 101719. https://doi.org/10.1016/j.giq.2022.101719

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

Hekkert, P. (2023). *Vision in product design (ViP) elective, Q4 2022/23 [lecture slides]*.

Hekkert, P., & van Dijk, M. (2011). *Vision in design: A guidebook for innovators* (2nd printing). BIS.

High Level Expert Group on Artificial Intelligence (HLEGAI). (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthyai

Hill, M. J., & Varone, F. (2021). *The public policy process* (Eighth edition). Routledge.

Hood, C. (2011). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton University Press. https://doi.org/10.1515/9781400836819

Howlett, M., & Ramesh, M. (2023). Designing for adaptation: Static and dynamic robustness in policy-making. *Public Administration*, *101*(1), 23–35. https://doi.org/10.1111/padm.12849

Inglehart, R., & Norris, P. (2016). Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2818659

Jacobs, K. (2018). Electoral Systems in Context: The Netherlands. In E. S. Herron, R. J. Pekkanen, & M. S. Shugart (Eds), *The Oxford Handbook of Electoral Systems* (pp. 556–580). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190258658.013.44

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, *40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Kamal, M. M. (2006). IT innovation adoption in the government sector: Identifying the critical success factors. *Journal of Enterprise Information Management*, *19*(2), 192–222. https://doi.org/10.1108/17410390610645085

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*(8), 963–973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

Kashdan, T. B., & Silvia, P. J. (2009). Curiosity and Interest: The Benefits of Thriving on Novelty and Challenge. In S. J. Lopez & C. R. Snyder (Eds), *The Oxford Handbook of Positive Psychology* (pp.

366–374). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195187243.013.0034

KCBR. (n.d.). *Beleidscyclus_Webtoegankelijk-v4*. Retrieved 15 February 2024, from https://www.kcbr.nl/sites/default/files/2023-04/253.127_Beleidscyclus_Webtoegankelijk-v4%5B47%5D%20een%20pagina.pdf

Kitchin, R. (2016). The ethics of smart cities and urban science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160115. https://doi.org/10.1098/rsta.2016.0115

König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly*, *37*(3), 101489. https://doi.org/10.1016/j.giq.2020.101489

Koppenjan, J. F. M., van Popering-Verkerk, J., van der Meer, J., Vroon, C., Spekkink, W., & Duijn, M. (2024). *Het gebruik van het Beleidskompas binnen de Rijksoverheid*. Erasmus Universiteit - Vakgroep Bestuurskunde & Sociologie. http://hdl.handle.net/20.500.12832/3394

Koskinen, I. K. (Ed.). (2011). *Design research through practice: From the lab, field, and showroom*. Morgan Kaufmann/Elsevier.

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in Policy versus Practice. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. https://doi.org/10.1145/3375627.3375835

Kuhn, T. S. (with Ralph Ellison Collection (Library of Congress)). (1970). *The structure of scientific revolutions* ([2d ed., enl). University of Chicago Press.

Kulk, S., & van Deursen, S. (2020). *Juridische aspecten van algoritmen die besluiten nemen. Een verkennend onderzoek*. WODC.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, *44*(6), 101976. https://doi.org/10.1016/j.telpol.2020.101976

Lampo, A., Mancarella, M., & Piga, A. (2018). *(Non)-neutrality of science and algorithms: Machine Learning between fundamental physics and society*. https://doi.org/10.13131/1724-451x.labsquarterly.axx.n4.117-145

Levi, M., & Stoker, L. (2000). Political Trust and Trustworthiness. *Annual Review of Political Science*, *3*(1), 475–507. https://doi.org/10.1146/annurev.polisci.3.1.475

Linder, S. H., & Peters, B. G. (1989). Instruments of Government: Perceptions and Contexts. *Journal of Public Policy*, *9*(1), 35–58. https://doi.org/10.1017/S0143814X00007960

Lipset, S. M. (1959). Some Social Requisites of Democracy: Economic Development and Political Legitimacy. *American Political Science Review*, *53*(1), 69–105. https://doi.org/10.2307/1951731

Liu, X., & Dijk, M. (2022). How more data reinforces evidence-based transport policy in the Short and Long-Term: Evaluating a policy pilot in two Dutch cities. *Transport Policy*, *128*, 166–178. https://doi.org/10.1016/j.tranpol.2022.09.022

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Loukis, E., Charalabidis, Y., & Androutsopoulou, A. (2017). Promoting open innovation in the public sector through social media monitoring. *Government Information Quarterly*, *34*(1), 99–109. https://doi.org/10.1016/j.giq.2016.09.004

Lowndes, V., & Paxton, M. (2018). Can agonism be institutionalised? Can institutions be agonised? Prospects for democratic design. *The British Journal of Politics and International Relations*, *20*(3), 693–710. https://doi.org/10.1177/1369148118784756

Madan, R., & Ashok, M. (2022). A Public Values Perspective on the Application of Artificial Intelligence in Government Practices: A Synthesis of Case Studies. In J. R. Saura & F. Debasa (Eds), *Advances in Electronic Government, Digital Divide, and Regional Development* (pp. 162–189). IGI Global. https://doi.org/10.4018/978-1-7998-9609-8.ch010

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*(1), 101774. https://doi.org/10.1016/j.giq.2022.101774

Madan, R., & Ashok, M. (2024). Making Sense of AI Benefits: A Mixed-method Study in Canadian Public Administration. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-024-10475-0

Margetts, H., & Dorobantu, C. (2019). Rethink government with AI. *Nature*, *568*(7751), 163–165. https://doi.org/10.1038/d41586-019-01099-5

Marquet, L. D. (2019). *Turn the ship around! A true story of turning followers into leaders*. Penguin Business.

Meijer, A., & Ruijer, E. (2021). *Code Goed Digitaal Openbaar Bestuur (CODIO): Borgen van waarden bij de digitalisering van het openbaar bestuur*. USBO Advies. https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/codio

Merriam-Webster. (n.d.). *Liable*. https://www.merriam-webster.com/dictionary/liable

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, *58*(3), 103434. https://doi.org/10.1016/j.im.2021.103434

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., Gupta, M., & Niehaves, B. (2022). Enabling AI capabilities in government agencies: A study of determinants for European municipalities. *Government Information Quarterly*, *39*(4), 101596. https://doi.org/10.1016/j.giq.2021.101596

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Moore, M. H. (2014). Public Value Accounting: Establishing the Philosophical Basis. *Public Administration Review*, *74*(4), 465–477. https://doi.org/10.1111/puar.12198

Mouffe, C. (1993). *The return of the political*. Verso.

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social Research: An International Quarterly*, *66*(3), 745–758.

Mul, S., & Werkhorst, K. (2020, June 11). *De digitale transformatie vraagt om een flexiebele wetgever.pdf*. https://esb.nu/de-digitale-transformatie-vraagt-om-een-adaptieve-wetgever/

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Neumann, O., Guirguis, K., & Steiner, R. (2022). Exploring artificial intelligence adoption in public organizations: A comparative case study. *Public Management Review*, *26*(1), 114–141. https://doi.org/10.1080/14719037.2022.2048685

Newman, J., & Mintrom, M. (2023). Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis. *Journal of European Public Policy*, *30*(9), 1839–1859. https://doi.org/10.1080/13501763.2023.2193223

Newman, J., Mintrom, M., & O’Neill, D. (2022). Digital technologies, artificial intelligence, and bureaucratic transformation. *Futures*, *136*, 102886. https://doi.org/10.1016/j.futures.2021.102886

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

OECD. (2021). *Recommendation of the Council for Agile Regulatory Governance to Harness Innovation* (No. OECD/LEGAL/0464). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0464

Organisation for Economic Co-operation and Development (OECD). (2024). *Recommendation of the council on artificial intelligence (OECD/LEGAL/0449)*. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Otjes, S., & Louwerse, T. (2018). Parliamentary questions as strategic party tools. *West European Politics*, *41*(2), 496–516. https://doi.org/10.1080/01402382.2017.1358936

Oudeyer, P.-Y., Gottlieb, J., & Lopes, M. (2016). Intrinsic motivation, curiosity, and learning. In *Progress in Brain Research* (Vol. 229, pp. 257–284). Elsevier. https://doi.org/10.1016/bs.pbr.2016.05.005

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review*, *83*(4), 863–877. https://doi.org/10.1111/puar.13615

Pettit, P. (2004). Depoliticizing Democracy. *Ratio Juris*, *17*(1), 52–65. https://doi.org/10.1111/j.0952-1917.2004.00254.x

Pixabay. (2014). *[Photograph of a girl playing with bubbles]*. https://pixabay.com/photos/young-girl-child-playing-bubbles-388661/

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton university press.

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Raad van State. (2021). *Toetsingskader – Digitalisering en wetgeving [Assessment framework – Digitalisation and legislation]*. Afdeling advisering van de Raad van State. https://www.raadvanstate.nl/publish/library/13/raad_van_state_uitgave_toetsingskader_-_digitalisering_en_wetgeving_def_losse_pags.pdf

Rijksoverheid. (2024). *Overheidsbrede visie Generatieve AI [Government-wide vision on Generative AI]*. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

https://www.rijksoverheid.nl/documenten/rapporten/2024/01/01/overheidsbrede-visie-generatieve-ai

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, *4*(2), 155–169.

Roozenburg, N. F. M., & Cross, N. G. (1991). Models of the design process: Integrating across the disciplines. *Design Studies*, *12*(4), 215–220. https://doi.org/10.1016/0142-694X(91)90034-T

Roozenburg, N. F. M., & Eekels, J. (2016). *Productontwerpen, structuur en methoden* (2e druk). Boom Lemma uitgevers.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.

Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society*, *75*, 102372. https://doi.org/10.1016/j.techsoc.2023.102372

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). *AI Watch. Defining Artificial Intelligence: Towards an operational definition and taxonomy of artificial intelligence* (No. EUR 30117 EN). Publications Office of the European Union. https://doi.org/10.2760/382730

Satariano, A., & Mozur, P. (2024, August 14). The Global Race to Control A.I. *The New York Times*. https://www.nytimes.com/2024/08/14/briefing/ai-china-us-technology.html

Sattlegger, A., & Bharosa, N. (2024). Beyond principles: Embedding ethical AI risks in public sector risk management practice. *Proceedings of the 25th Annual International Conference on Digital Government Research*, 70–80. https://doi.org/10.1145/3657054.3657063

Sattlegger, A., Van Den Hoven, J., & Bharosa, N. (2022). Designing for Responsibility. *DG.O 2022: The 23rd Annual International Conference on Digital Government Research*, 214–225. https://doi.org/10.1145/3543434.3543581

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2021). A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-

Welfare. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–41. https://doi.org/10.1145/3476089

Schakel, W. (2021). Unequal policy responsiveness in the Netherlands. *Socio-Economic Review*, *19*(1), 37–57. https://doi.org/10.1093/ser/mwz018

Schmidt, V. A. (2017). Britain-out and Trump-in: A discursive institutionalist analysis of the British referendum on the EU and the US presidential election. *Review of International Political Economy*, *24*(2), 248–269. https://doi.org/10.1080/09692290.2017.1304974

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Schreuder, Y. (2001). The Polder Model in Dutch Economic and Environmental Planning. *Bulletin of Science, Technology & Society*, *21*(4), 237–245. https://doi.org/10.1177/027046760102100401

Selten, F., & Klievink, B. (2024). Organizing public sector AI adoption: Navigating between separation and integration. *Government Information Quarterly*, *41*(1), 101885. https://doi.org/10.1016/j.giq.2023.101885

Sienkiewicz-Małyjurek, K. (2023). Whether AI adoption challenges matter for public managers? The case of Polish cities. *Government Information Quarterly*, *40*(3), 101828. https://doi.org/10.1016/j.giq.2023.101828

Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. *AI and Ethics*, *2*(4), 763–770. https://doi.org/10.1007/s43681-022-00136-w

Stompff, G., Van Bruinessen, T., & Smulders, F. (2022). The generative dance of design inquiry: Exploring Dewey’s pragmatism for design research. *Design Studies*, *83*, 101136. https://doi.org/10.1016/j.destud.2022.101136

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., {Gonzalez Zelaya}, C., & {van Moorsel}, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Twizeyimana, J. D., & Andersson, A. (2019). The public value of E-Government – A literature review. *Government Information Quarterly*, *36*(2), 167–178. https://doi.org/10.1016/j.giq.2019.01.001

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Valle-Cruz, D., Criado, J. I., Sandoval-Almazán, R., & Ruvalcaba-Gomez, E. A. (2020). Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Government Information Quarterly*, *37*(4), 101509. https://doi.org/10.1016/j.giq.2020.101509

van Boeijen, A. G. C., Daalhuizen, J. J., Zijlstra, J. J. M., & van der Schoor, R. S. A. (Eds). (2013). *Delft design guide*. BIS Publishers.

van den Berg, C., Schmidt, A., & Van Eijk, C. (2015). Externe advisering binnen de Nederlandse overheid: Naar een empirisch en theoretisch onderbouwde onderzoeksagenda. *Bestuurskunde*, *24*(3), 17–31. https://doi.org/10.5553/Bk/092733872015024003003

van der Staaij, K., & Sneller, J. (2023, December). *Werken aan wetten; Praktische handreiking wetgevingskwaliteit voor Kamerleden*. Tweede Kamer der staten-generaal. https://www.tweedekamer.nl/downloads/document?id=2023D47970

van Noordt, C. (2023). *Public Value Creation with Artificial Intelligence Technologies in Public Administration* [[object Object]]. https://doi.org/10.23658/TALTECH.58/2023

van Noordt, C., Medaglia, R., & Tangi, L. (2023). Policy initiatives for Artificial Intelligence-enabled government: An analysis of national strategies in Europe. *Public Policy and Administration*, *40*(2), 215–253. https://doi.org/10.1177/09520767231198411

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3), 101714. https://doi.org/10.1016/j.giq.2022.101714 van Noordt, C., & Tangi, L. (2023). The dynamics of AI capability and its influence on public value creation of AI within public administration. *Government Information Quarterly*, *40*(4), 101860. https://doi.org/10.1016/j.giq.2023.101860

Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. *Government Information Quarterly*, *36*(4), 101383. https://doi.org/10.1016/j.giq.2019.05.010

Watson, R., & Dorst, K. (2022, June 25). *Pragmatism, design and public sector innovation: Reflections on action*. DRS2022: Bilbao. https://doi.org/10.21606/drs.2022.778

Weedon, S. (2019). The Core of Kees Dorst's Design Thinking: A Literature Review. *Journal of Business and Technical Communication*, *33*(4), 425–430. https://doi.org/10.1177/1050651919854077

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (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

WRR. (2021). *Opgave AI: De nieuwe systeemtechnologie*. Wetenschappelijke Raad voor het Regeringsbeleid. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 535–563. https://doi.org/10.1145/3531146.3533118

Zimmerman, J., & Forlizzi, J. (2008). The Role of Design Artifacts in Design Theory Construction. *Artifact*, *2*(1), 41–45. https://doi.org/10.1080/17493460802276893

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577

APPENDIX

[Redacted]

IDE Master Graduation Project

Project team, procedural checks and Personal Project Brief

In this document the agreements made between student and supervisory team about the student's IDE Master Graduation Project are set out. This document may also include involvement of an external client, however does not cover any legal matters student and client (might) agree upon. Next to that, this document facilitates the required procedural checks:

- Student defines the team, what the student is going to do/deliver and how that will come about
- Chair of the supervisory team signs, to formally approve the project's setup / Project brief
- SSC E&SA (Shared Service Centre, Education & Student Affairs) report on the student's registration and study progress
- IDE's Board of Examiners confirms the proposed supervisory team on their eligibility, and whether the student is allowed to start the Graduation Project

STUDENT DATA & MASTER PROGRAMME

Complete all fields and indicate which master(s) you are in

Family name	Mieras	IDE master(s)	IPD □ DfI □ SPD ☑
Initials	D.L.	2nd non-IDE master	
Given name	David	Individual programme *(date of approval)*	
Student number		Medisign	□
		HPM	□

SUPERVISORY TEAM

Fill in he required information of supervisory team members. If applicable, company mentor is added as 2nd mentor

Chair	Lianne Simonse	dept./section	DOS / DIVE
mentor	Kars Alfrink	dept./section	SDE / IoT
2nd mentor	Suzanne van Melis		
client:			
city:		country:	
optional comments			

! Ensure a heterogeneous team. In case you wish to include team members from the same section, explain why.

! Chair should request the IDE Board of Examiners for approval when a non-IDE mentor is proposed. Include CV and motivation letter.

! 2nd mentor only applies when a client is involved.

APPROVAL OF CHAIR on PROJECT PROPOSAL / PROJECT BRIEF -> to be filled in **by the Chair** of the supervisory team

Sign for approval (Chair)

Name Lianne Simonse Date 18 Apr 2024 Signature

CHECK ON STUDY PROGRESS

To be filled in **by SSC E&SA** (Shared Service Centre, Education & Student Affairs), after approval of the project brief by the chair. The study progress will be checked for a 2nd time just before the green light meeting.

Master electives no. of EC accumulated in total ______ EC

Of which, taking conditional requirements into account, can be part of the exam programme ______ EC

□	**YES**	all 1st year master courses passed
□	**NO**	missing 1st year courses

Comments:

Sign for approval (SSC E&SA)

Name ______ Date ______ Signature ______

APPROVAL OF BOARD OF EXAMINERS IDE on SUPERVISORY TEAM -> to be checked and filled in by IDE's Board of Examiners

Does the composition of the Supervisory Team comply with regulations?

YES	□	Supervisory Team approved
NO	□	Supervisory Team not approved

Comments:

Based on study progress, students is …

□	**ALLOWED** to start the graduation project
□	**NOT** allowed to start the graduation project

Comments:

Sign for approval (BoEx)

Name ______ Date ______ Signature ______

Personal Project Brief – IDE Master Graduation Project

Name student David Mieras

Student number

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT

Complete all fields, keep information clear, specific and concise

Project title Navigating value tensions in the use of AI for policy preparation; towards guidelines and a practical tool

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

In recent years large scandals, like the Dutch childcare scandal, which led to the resignation of the Dutch cabinet, have illuminated the risks of using artificial intelligence (AI) in policy execution (Peeters & Widlak, 2023). This project focuses on a subject that has received less attention, but is fundamentally challenging to our democratic system: the use of artificial intelligence systems in policy preparation.

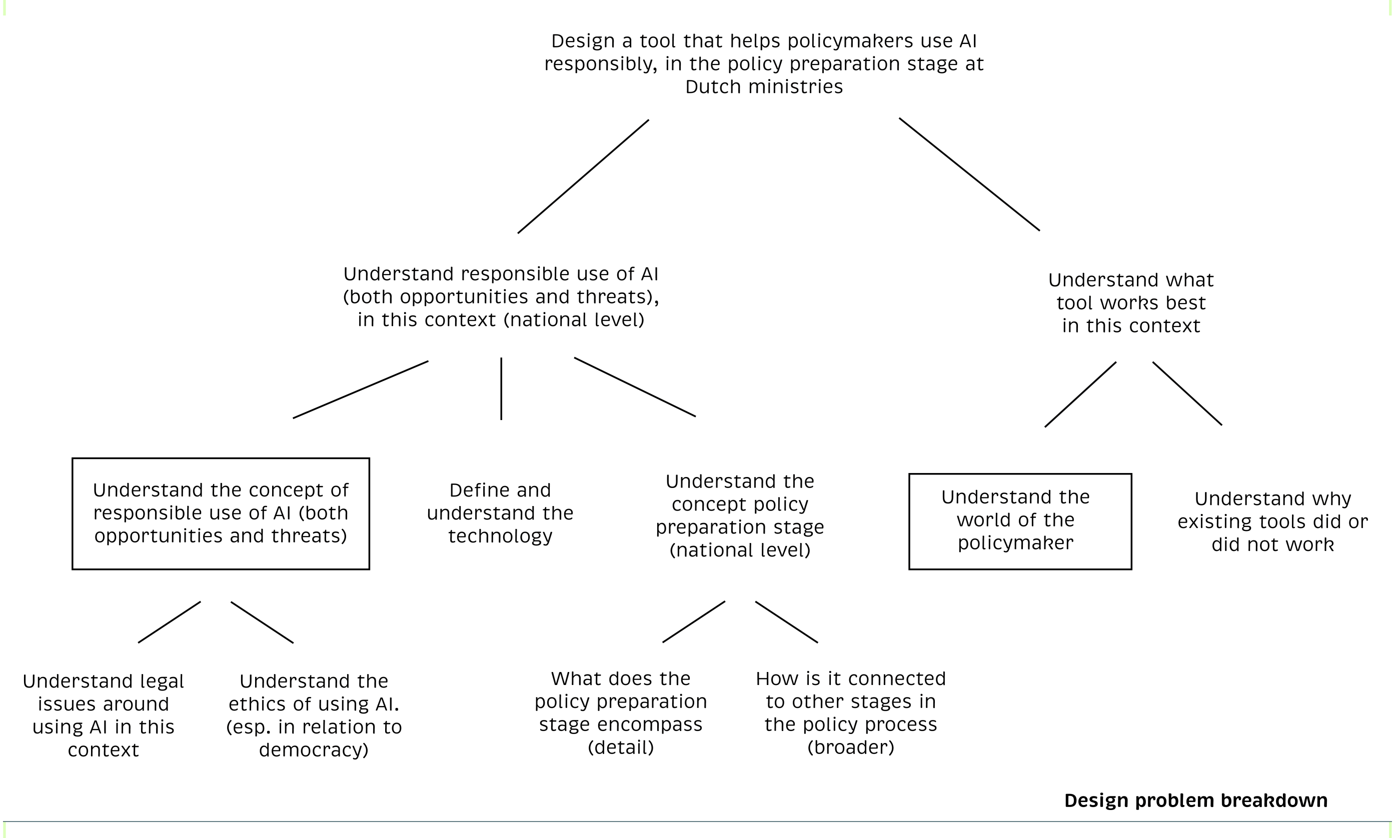
AI is a disruptive technology with great potential benefits. However, due to the novelty of the technologies, the effects of using AI in the development of policy are uncertain (Prins, 2021). What is certain is that AI affects the processes it's introduced to, AI is non-neutral (Janssen et al., 2022; Kuhn, 1970; Stinson, 2022). This applies unabatedly to the adoption of AI in the democratic institutions at the core of our society (Dafoe, 2018; Newman & Mintrom, 2023). Results provided by AI can for instance have an anchoring- or depoliticizing effect (Pettit, 2004), even when policy preparation professionals critically weigh an AI's output, hurting the quality and democratic legitimacy of policy.

To mitigate the risks and bring the benefits of AI in the policy preparation stage within reach, ministries need practical guidance on the responsible use of AI for policy preparation professionals, considering both the procedural- (using AI for the right thing) and technical (AI doing what it is intended to do) proper functioning of AI.

The application domain of this project is the policy preparation practice at Dutch ministries. The main stakeholders are civil servants involved in policy preparation and the users of the design. Other stakeholders are civil service leadership, political leadership and citizens. Whose interest(s) and support are crucial for the adoption of the design.

➔ space available for images / figures on next page

introduction (continued): space for images



Design a tool that helps policymakers use AI responsibly, in the policy preparation stage at Dutch ministries

Understand responsible use of AI (both opportunities and threats), in this context (national level)

Understand what tool works best in this context

Understand the concept of responsible use of AI (both opportunities and threats)

Define and understand the technology

Understand the concept policy preparation stage (national level)

Understand the world of the policymaker

Understand why existing tools did or did not work

Understand legal issues around using AI in this context

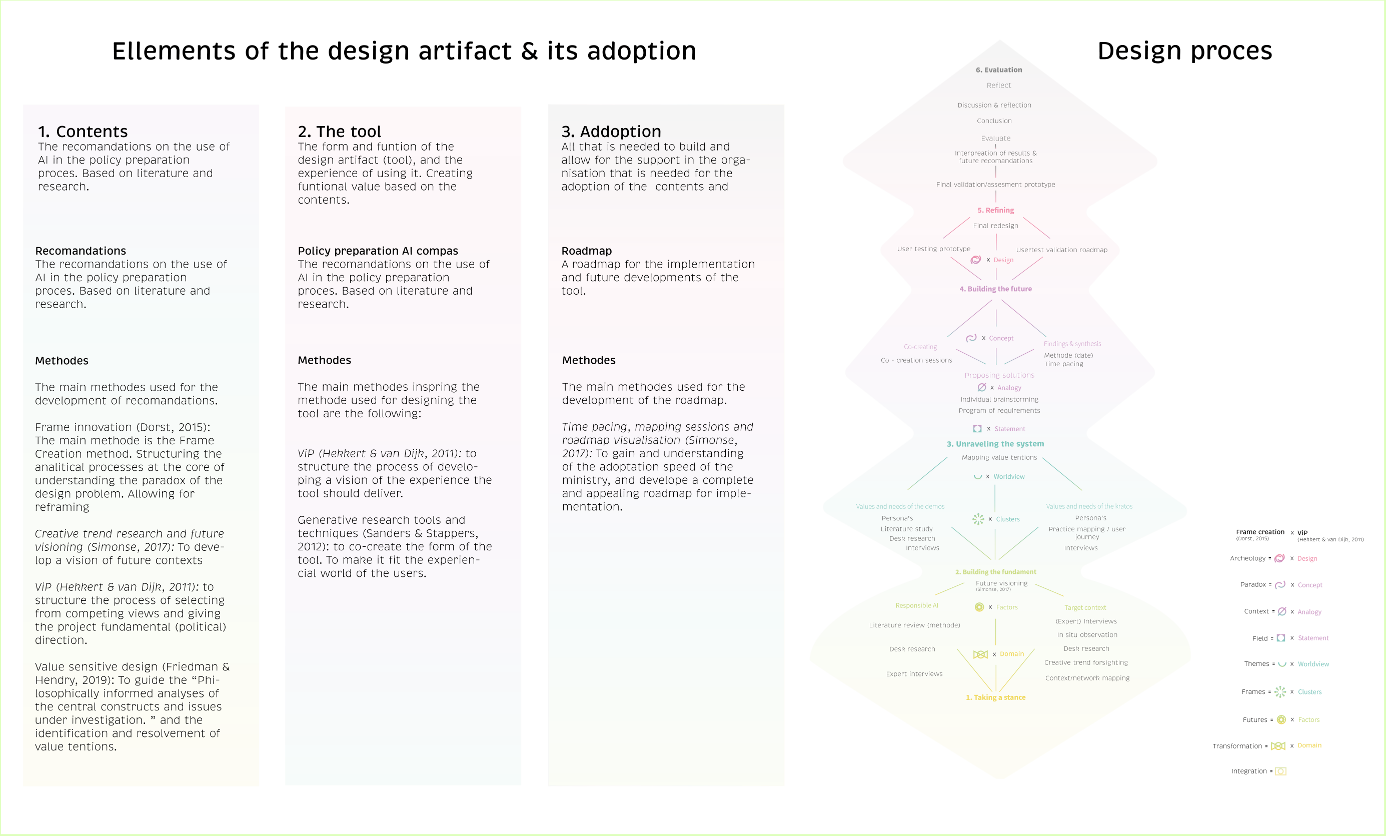
Understand the ethics of using AI. (esp. in relation to democracy)

What does the policy preparation stage encompass (detail)

How is it connected to other stages in the policy process (broader)

Design problem breakdown

image / figure 1 Design problem breakdown



Ellements of the design artifact & its adoption

1. Contents
The recomandations on the use of AI in the policy preparation proces. Based on literature and research.

Recomandations
The recomandations on the use of AI in the policy preparation proces. Based on literature and research.

Methodes

The main methodes used for the development of recomandations.

Frame innovation (Dorst, 2015): The main methode is the Frame Creation method. Structuring the analitical processes at the core of understanding the paradox of the design problem. Allowing for reframing

Creative trend research and future visioning (Simonse, 2017): To develop a vision of future contexts

ViP (Hekkert & van Dijk, 2011): to structure the process of selecting from competing views and giving the project fundamental (political) direction.

Value sensitive design (Friedman & Hendry, 2019): To guide the "Philosophically informed analyses of the central constructs and issues under investigation. " and the identification and resolvement of value tentions.

2. The tool
The form and funtion of the design artifact (tool), and the experience of using it. Creating funtional value based on the contents.

Policy preparation AI compas
The recomandations on the use of AI in the policy preparation proces. Based on literature and research.

Methodes

The main methodes inspring the methode used for designing the tool are the following:

ViP (Hekkert & van Dijk, 2011): to structure the process of developing a vision of the experience the tool should deliver.

Generative research tools and techniques (Sanders & Stappers, 2012): to co-create the form of the tool. To make it fit the experiencial world of the users.

3. Addoption
All that is needed to build and allow for the support in the organisation that is needed for the adoption of the contents and

Roadmap
A roadmap for the implementation and future developments of the tool.

Methodes

The main methodes used for the development of the roadmap.

Time pacing, mapping sessions and roadmap visualisation (Simonse, 2017): To gain and understanding of the adoptation speed of the ministry, and develope a complete and appealing roadmap for implementation.

Design proces

image / figure 2 Design methods

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

Practice | The governmental organisation is presented with vast opportunities of using AI in the policy preparation stage, but she is under-prepared for the adoption of these novel technologies. In part because she lacks standards and guidance on the responsible, safe and effective, use of AI. While governmental organisation (the collective) is aware of the general risks involved in the use of AI, the average policy preparation professional has little insight into these risks. Fundamental risks like depoliticization as a result of AI adoption are largely overlooked and require practical translation to be put on the agenda. The foregoing suggests that the governmental organisation needs guidelines and a practical tool to support the responsible use of AI by policy preparation professionals, informed by existing frameworks and insights from international organizations such as the OECD (2021), the ECP (2018), and academic literature and frameworks like the value-based assessment model (Yurrita et al., 2022), and the five loops model for the systemic integration of contestability in AI based policy execution (Alfrink et al., 2023).

Academic | A gap in academic literature is the general lack of application to and testing in practice. Insight into the relation between frameworks developed in science and implementation in practice (in the complex and dynamic political environment of government practice) is necessary to align future scientific research with the needs of practice.

Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Design a set of guidelines and a prototype for tool assisting professionals within Dutch ministries in the ethical and effective utilization of AI in the policy preparation stage . The tool should fit the existing organizational environment and resonate with the practical experiences of these professionals, ensuring they can capture the benefits of AI while effectively mitigating associated risks.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

This design problem is "wicked" (Buchanan, 1992; Churchman, 1967), situated around a rapidly developing technology in an open and dynamic governmental context with great societal impact. The project is undertaken in a dynamic and complex context (Howlett & Ramesh, 2023), with a fast amount of stakeholders with contradicting interests, making the design problem especially "open, complex, and dynamic" (Dorst, 2015, p. 11), centers around paradoxes and value tensions. This necessitates that problem, approach and solution co-evolve throughout the project.
To allow this co-evolution of the approach the design methodology for this project is rooted in three design approaches, covering the wider field of challenges that can be expected to arise throughout the project (see Figure 2).

The frame creation method (Dorst, 2015) is the main methodological grounding, in conjunction with a value sensitive design (Friedman et al., 2002) approach. Additionally, the design problem poses philosophical, political, and experiential questions. Questions that have no definitive answers. To answer these questions, visions need to be developed. To this end, the vision in product design method (ViP) (Hekkert & Dijk, 2011) is invoked together with a roadmapping approach (Kim et al., 2022; Simonse, 2017).

Project planning and key moments

*To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a **kick-off meeting**, **mid-term evaluation meeting**, **green light meeting** and **graduation ceremony**. Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting 11 mrt 2024

Mid-term evaluation 16 mei 2024

Green light meeting 11 juli 2024

Graduation ceremony 15 aug 2024

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	✔
For how many project weeks	20
Number of project days per week	1,0

Comments:
Will be performing additional internship tasks for governmental organisation .

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

I believe in a fair and honest liberal democratic system of governance, based on (individual) freedom, honesty, problematization and politicization. For governance that shapes society to the will of the people, for the good of the collective, and protecting vulnerable groups.

As a young strategic designer, at a governmental organisation, I will be in the best place to take part in the continued improvement of our democratic governance system. With this project I challenge myself to use all strategic design competencies I have acquired, to deal with an "open, complex, and dynamic" (Dorst, 2015, p. 11) problem, in a political organizational context.

My personal goals:

1. I aim to gain in-depth knowledge of what is needed to use AI responsible in the policy preparation phase, in a democratic system.
2. I want to prove that I can do rigorous academic research;
3. Further develop my ability to design strategically for wicked problems in a complex, networked and political environment;
4. Enhance my capacity to manage time constraints;
5. Foster active participation and garner support from stakeholders to create engagement and buy-in for the project.

Planning

Stage	Date	Goal	Activities	Methodes	Outcome
0. Preparation					
Week 0	*29-Jan*				
		Finding a mentor + organising start of the project			
1. Taking a stance		*Map the problem space, Iterate the research questions and approach*	*Literature study, supplemented with academic- and field experts*		
Week 0.1	*05-Feb*	Identify topics and works of intrest	Literature study	Snowball sampling literature study	Literature problem space overview
Week 0.2	*12-Feb*	Identify key words and open reserch question, itterate research question and methodes + potential publishers			Research questions & research methodes
2. Building the fundament		*Literature search*	*Literature study Academic- and field expert interviews*		*Literature study paper*
Week 0.3	*19-Feb*	Intruduction at the ministry			
Week 1	*11-Mar*				
Kick off					
Week 2	*18-Mar*				
Week x	*25-Mar*	*Sick			
Week 3	*01-Apr*				
Week 4	*08-Apr*				
Week 5	*15-Apr*				
Week 6	*22-Apr*	Writing findings and conclusion			Literature study paper + practically relevant work summarised

3. Unraveling the system		*Understanding the needs in the system*		*Qualitative interview, value mapping*	*A overview of relevant values & value tentions*
Week 6	*29-Apr*				
Week 1					
Week 7	*06-May*				
Week 2					
Week 8	*13-May*				
Mid term					
Week 19	*20-May*				
Week 4					
Week 10	*27-May*				
Week 5					
4. Building the future		*Proposing and cocreating the design*			*The design*
Week 11	*03-Jun*				
Week 1					
Week 12	*10-Jun*				
Week 2					
Week 13	*17-Jun*				
Week 3					
Week 14	*24-Jun*				
Week 4					
Week 15	*01-Jul*				
Week 5					
Week 16	*08-Jul*	**Last internship week**			
Green light					
5. Refining		*Polish the design, thesis and papers*			*The design, the papers and the thesis*
Week 17	*15-Jul*				
Week 1					
Week 18	*22-Jul*				
Week 2					
Week 19	*29-Jul*				
Week 3					
Week 20	*05-Aug*				
Week 4					
Week 21	*12-Aug*	**Monday: Graduation deadline**			

References

Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2023). Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3544548.3580984

Buchanan, R. (1992). Wicked Problems in Design Thinking. *Design Issues*, *8*(2), 5. https://doi.org/10.2307/1511637

Churchman, C. W. (1967). Guest Editorial: Wicked Problems. *Management Science*, *14*(4,), B141–B142.

Dafoe, A. (2018). *AI governance—A research agenda.pdf*. Centre for the Governance of AI Future of Humanity Institute University of Oxford. fhi.ox.ac.uk/govaiagenda

Dorst, K. (2015). *FRAME INNOVATION: Create new thinking by design*. MIT PRESS.

ECP. (2018). *Artificial Intelligence Impact Assesment*. https://ecp.nl/artificial-intelligence-impact-assessment/

Kim, E., Simonse, L. W. L., Beckman, S. L., Appleyard, M. M., Velazquez, H., Madrigal, A. S., & Agogino, A. M. (2022). User-Centered Design Roadmapping: Anchoring Roadmapping in Customer Value Before Technology Selection. *IEEE Transactions on Engineering Management*, *69*(1), 109–126. https://doi.org/10.1109/TEM.2020.3030172

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Friedman, B., Kahn, P. H., & Borning, A. (2002). *Value Sensitive Design: Theory and Methods*.

Hekkert, P., & Dijk, M. van. (2011). *Vision in design: A guidebook for innovators* (2nd printing). BIS.

Howlett, M., & Ramesh, M. (2023). Designing for adaptation: Static and dynamic robustness in policy-making. *Public Administration*, *101*(1), 23–35. https://doi.org/10.1111/padm.12849

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, *40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Kuhn, T. S. (1970). *The structure of scientific revolutions* ([2d ed., enl). University of Chicago Press.

Newman, J., & Mintrom, M. (2023). Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis. *Journal of European Public Policy*, *30*(9), 1839–1859. https://doi.org/10.1080/13501763.2023.2193223

OECD. (2021). *Recommendation of the Council for Agile Regulatory Governance to Harness Innovation* (OECD/LEGAL/0464). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0464

Pettit, P. (2004). Depoliticizing Democracy. *Ratio Juris*, *17*(1), 52–65. https://doi.org/10.1111/j.0952-1917.2004.00254.x

Prins, C. (2021). *Opgave AI: De nieuwe systeemtechnologie*. Wetenschappelijke Raad voor het Regeringsbeleid.

Sanders, E. B.-N., & Stappers, P. J. (2012). *Convivial design toolbox: Generative research for the front end of design*. BIS.

Simonse, L. W. L. (2017). *Design roadmapping* (J. Whelton, Ed.). Bis Publishers.

Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. *AI and Ethics*, *2*(4), 763–770. https://doi.org/10.1007/s43681-022-00136-w

Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 535–563. https://doi.org/10.1145/3531146.3533118