Whitepaper

Facilitating Responsible AI with deliberate decision making

Lessons learned in the public sector

Authors

Marianne Witte-Schaaphok
Wouter van der Bij
Marissa Hoekstra
Romy van Drie
Steven Vethman

TNO 2024 P12747

TNO innovation for life

Summary

You are motivated to develop and use AI responsibly in your organization. In 2024 this has become an overwhelming task considering the number of technical, legal, ethical and organizational considerations that need to be accounted for. On the one hand, the development of technology such as Generative AI has created a world of opportunity. On the other hand, the explication of requirements through the AI Act and other associated standards has shifted the focus to the obligations of using AI. Ultimately, it is important that both perspectives are acknowledged such that we can implement AI in a meaningful way, while safeguarding our values and the risks that come with it. How do you ensure that these elements collectively come together to stimulate valuable innovation with AI? How do you strike the balance between value-driven innovation, attaining the necessary resources and weighing the risks of AI for your organization? Lastly, how do you ensure that people from interdisciplinary background can manage these factors in their decision making? The AI Oversight Lab has four years of experience in answering these questions in practice through different use cases. Based on these experiences, we have drafted eight key recommendations that help guide the practice of AI decision making. These recommendations go beyond considerations of the technical aspects and way of working, to include the practice of responsible decision making about AI.

1. Formulate an explicit objective and consider AI only as a tool
2. Ensure a structured development process including explicit decision moments
3. Discuss and assess the BVRR-balance at key decision moments
4. Invest in an interdisciplinary approach
5. Provide a safe experimentation environment
6. Develop an AI vision and clarify ownership of AI applications
7. Invest in AI knowledge and skills of employees
8. Start and learn with low-risk applications

The rise of AI

The number of applications of Artificial Intelligence is growing rapidly, as is the number of tools to deploy it responsibly. Yet in practice, it remains difficult to shape responsible AI development properly.

Artificial Intelligence (AI) has made a rapid advance in recent years with the promise that organisations can work more efficiently, effectively and objectively with AI. Notably the emergence of generative AI added particular momentum to the incorporation of AI in workplace activities. In the public sector, as well, AI brings opportunities to work more effectively and has the potential to improve public services to make it easier for citizens to get what they need[1,2]. With the ambition for a stronger information-driven way of working in government processes, there has been an increase in experimentation with AI[3,4]. Thanks to the accessibility of the AI software embedded by commercial tech companies, organisations can easily begin using AI in their procured systems and processes. However, considering the public nature of government organisations, it is important that they deploy AI in a responsible manner. To support this deployment of AI, it must be based on interdisciplinary decision making, considering the explicit trade-off between added value, public values, resources and risks of using AI.

In practice, organisations run into many challenges that make the responsible adoption of AI challenging. This includes limited resources in the organisation, poor data quality, and difficulty in finding the right expertise. As a result, innovations are often sidelined, delayed or left unable to scale up. In addition, there are applications of AI in the market that have not taken the risks into proper consideration[5], resulting in potential harmful consequences. Unfortunately, there are many examples of such cases, for example the reports of discriminatory algorithms at DUO[6], UWV[7] and municipality of Rotterdam[8].

Even if organisations pay deliberate attention to their algorithm, the effects could still show unintentional discriminatory bias, robustness issues or other unwanted effects[9]. A more structured development approach and continuous monitoring during deployment is essential to decrease these risks.

The European Union is working to understand how to regulate the responsible development of AI with the AI-Act10. In doing so, the EU has stipulated that using AI in an organisation requires a certain level of AI literacy in addition to a set of guardrails for high-risk applications, among other things. The AI Act went into effect in August 2024 and has shifted the responsible AI conversation within many organisations to one about compliance with AI regulation. Being compliant with the AI Act, however, does not always guarantee responsible AI innovation.

The growing focus on responsible AI and the introduction of the AI Act are leading to an overflowing number of technical solutions, handbooks, guidelines and resources that support the responsible development and deployment of AI. From the technical perspective, there are many accurate[11] , fair[12] and explainable[13] methods being developed and researched to help the development of responsible AI. The ethical perspective also offers a plethora of tools to support ethical considerations such as the non-discrimination handbook[14] , DEDA[15] , IAMA[16] , Approach to Guidance Ethics[17] , ethical guidelines for reliable AI[18] , ALTAI[19] and the Algorithm Framework[20]. With this toolbox for AI development, developers and deployers can make considerations beyond the workings of the technology. It supports considerations on goals, responsibilities and the impact on stakeholders and society. Despite the wide range of tools and support available, in practice we have found that it remains challenging to successfully shape responsible AI innovation in the context of a real life use case in a real-world organization.

Responsible decision making

In practice, more attention is required to facilitating responsible decisions. People within the organisation need to be empowered to make the trade-offs in development and deployment explicit, including the benefits-value-resource-risk balance.

Organisations experience challenges in leading the development process, validating whether the innovation is justified, and moving to implementation. Two aspects play an essential role here. Firstly, too little attention is paid to facilitating responsible decision making. Secondly, too limited attention is paid to the explicit trade-off between 1) the benefits and purpose, 2) public values at play, 3) available resources, and 4) residual risks. We call this the benefits-values-resources-risks balance, or BVRR balance for short.

Facilitating responsible decisions

The focus of responsible AI development tends to center around two perspectives 1) responsible AI technology, and 2) responsible AI use. The technology perspective focuses on the development of technical methods that support certain aspects of responsible AI, such as fairness or transparency. These include increasing general performance, developing bias detection and mitigation techniques, increasing the explainability of algorithms and developing hybrid AI methods. From the second broader perspective, research directs at the interaction with and deployment of AI. Here, the main focus is on the work process in which the algorithm is used and the interaction between human and algorithm. This includes how the outcomes of the AI are used, what the role of the user is, how the user can remain critically reflective, what are the measures to avoid unwanted bias and how the deployment of the AI is communicated. Both perspectives involve people making different decisions to ensure that technology and the way of use are implemented in a responsible manner.

We state that a third perspective on facilitating responsible and deliberate decisions is essential to include in this process. This perspective focuses on the explicit trade-offs that are considered behind the scenes and drive the development of value driven, responsible AI innovation. From this perspective we push organisations to be aware and explicit when making decisions about AI. Consider decisions about the intended goals, underlying (public) values, the technical design, the interaction with humans and the way of working. Moreover, it includes higher level decisions about the development process, such as stakeholder engagement, investments and key go/no-go decisions. Besides making the decisions explicit, it is important that accountability over these decisions is guaranteed through proper consultation and documentation. Many of these decisions are highly dependent on the context of the application and therefore require the right expertise and interdisciplinary collaboration. Organisations need to prepare and facilitate their teams to make these decisions. While all the tools mentioned (technical, legal and ethical) help kickstart this process, the deployment of responsible AI requires sufficient time, resources and expertise to apply these tools in practice.

“Pay more attention to responsible decisions in the balance between benefits, public values, resources and risks.”

The BVRR-balance

We pose the BVRR-balance as an important framework for responsible decision making. The BVRR balance is the explicit trade-off between benefits (B), public values (V), available resources (R) and residual risks (R). The figure on the next page shows the four aspects of the balance and the interdisciplinary decision based on the assessed aspects. In the end the benefits and alignment with public values[21] must outweigh the resource commitment and potential risks. Performing all the requirements and impact assessments for an (high-risk) algorithm requires a lot of resources, such as time, money, capacity, capabilities, data and technical infrastructure. This trade-off must be at the heart of the AI development process. At each stage of the development process, it is important to make this balance explicit at defined decision moments. Between every phase there should be a new deliberate decision moment (go/no-go gate) whether the AI application is worth to continue developing or implementing. At this gate you should also determine what resources can and should be allocated to responsibly progress in the next phase. This decision will be made based on information collected through experiments and other sources during the development process. Throughout the AI lifecycle the different aspects of the balance will be extended with more detailed knowledge based on research and experiments, allowing for a properly justified decision before the deployment stage and deliberate monitoring and evaluation while AI solution is in real life production.

Having a clear AI vision and strategy on an organisational level that provides guidance is an essential prerequisite for constructing this balance. This vision should provide guidance on responsibilities and pose AI as a tool rather than a goal. Through a clear vision that fosters critical reflection, an organisational culture can develop that allows for experimentation and learning, which strengthens decision making.

Illustration: BVRR-balance in the AI lifecycle

After the problem definition phase you decide whether AI is a potential solution for the problem. In the BVRR balance you note the purpose, based on the problem definition, and the intended benefits. You broadly identify potential risks and the alignment with various public values, based on exploration and interdisciplinary discussions. You determine which information you need to gather in the design phase to validate hypotheses about feasibility, benefits and risks. You make an estimate of required resource investments for the upcoming phase but also to bring the AI solution to production responsibly. Together you decide whether you want to continue to the “design” phase and which investments will be made.

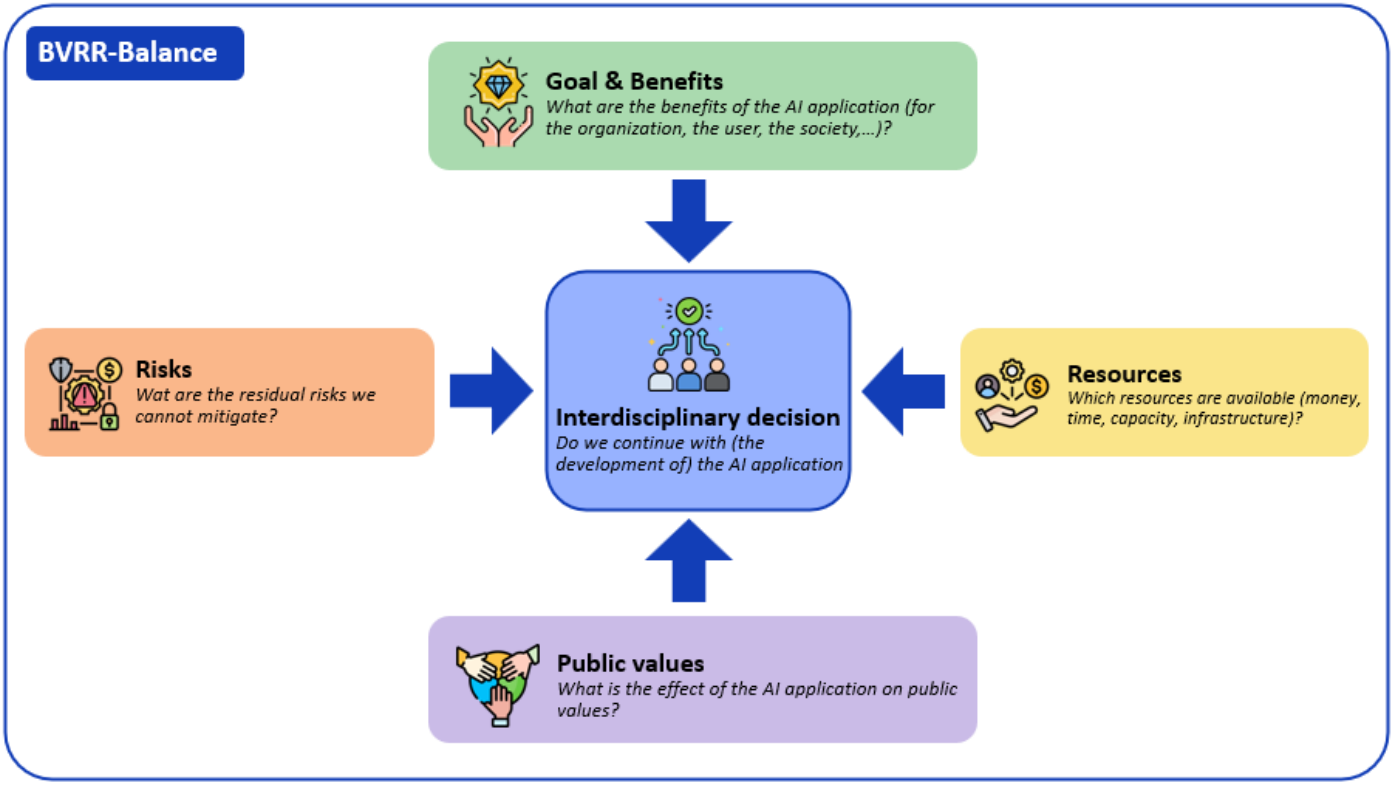
After this design phase, where you have built a proof-of-concept (PoC), you decide whether to run a proof-of-value or pilot. During the PoC-phase you have gained new insights on benefits and potential risks, and you have run the determined tests to validate the hypotheses from the first phase. These new insights are added to the balance. A better understanding of the alignment with public values is gained to strengthen the decision and follow-up actions. Based on the updated balance you determine whether to run a pilot and which hypotheses you need to test in the pilot phase. These hypotheses should cover the various aspects of the balance.

After the pilot phase you decide whether to implement, discard or further develop the application. During the pilot phase you have collected evidence for claims on benefits, risks and value alignment. The balance is now far more detailed and complete than the first phase. Based on these aspects you can decide whether the obtained benefits are in line with the residual risks. This provides a well structured and detailed overview for the decision on deployment.

Note that also during deployment, regular monitoring of the application and assessment of the balance is required to maintain responsible and justified use.

To prepare the BVRR balance properly, it is important for the teams to have a common language and an interdisciplinary way of working. The BVRR balance combines various disciplines and an interdisciplinarity approach ensures that all perspectives and knowledge are taken into account. The decision whether to continue with the development or deployment of the AI application, should be made with an interdisciplinary mindset and should be documented to increase accountability and transparency.

In the next section, we elaborate on the practical challenges of making responsible decisions based on our experiences in public sector. In the recommendations section, we propose recommendations that support organisations in discussing and assessing the BVRR balance and facilitating responsible decisions.



Challenges of responsible decisions about AI

We see in practice that many organisations encounter challenges in the responsible development of AI, despite the introduction of legislation, guidelines, manuals and tools. In this section, we elaborate on the challenges in practice and how they contribute to the call for making the BVRR balance explicit and facilitating responsible decisions.

Limited availability of time and resources

Working out all aspects of responsible AI and the responsible deployment of AI requires time and resources, which are often limited in practice. For example, in the development and evaluation of an AI application, the input of users and domain experts is crucial. In practice, these people often have a high workload. As a result, they are sometimes involved less or later than desired. Besides, a lack of budget or capacity sometimes makes it difficult to go into the desired level of detail for exploration or evaluation.

To transition from development to successful implementation, decisions must be made in practice about what will and will not be done. These decisions must be made in such a way that the added value can be demonstrated and the risks can be identified. There must be commitment from management with sufficient resources and capacity to facilitate a responsible decisions on continued development or deployment.

Conflicts between different public values

Every AI application touches multiple public values, such as efficiency, effectiveness, safety, transparency, accountability, privacy, autonomy and inclusiveness.[22] Several methodologies are now available that can help identify and specify public values, such as ECP's Guidance Ethics Approach[23] , value-based sensitive design[24] and the data dialogue[25]. Public values can however come into conflict. For example, consider weighing the need for transparency and security when deploying AI models or the degree of inclusiveness and efficiency throughout the development process of an AI application. Achieving an inclusive development process for an AI application can be time-consuming and resource-intensive to engage the right stakeholders, which can come at the expense of efficiency. In practice, it remains difficult to make trade-offs between public values. On top of that, in some cases, prioritising a public value may have a positive impact on one group but a (potentially) negative impact on another.

We observe that organisations often want to get started with AI because they hope it can make their work processes more efficient and effective. In doing so, they set efficiency and effectiveness as central public values, often without explicit consideration of trade-offs with other values. It is important to assess whether efficiency and effectiveness are the central values for the intended application and whether they can be realised within a responsible BVRR balance.

Limited AI literacy within organisations

We We see that there is often limited AI literacy within organisations. AI literacy means that users have knowledge and skills to work responsibly with the AI application and can critically evaluate AI outcomes26. AI literacy is needed to make employees aware of the capabilities and limitations of AI and empower them to critically reflect on AI outcomes. In both development and use of AI, it is important that employees have the knowledge and skills to make responsible decisions. Note that people involved in development require higher levels of AI literacy than users. In development, this involves decisions for the algorithm design, work process design and opportunities for appropriate human control. In addition, suitable knowledge and skills are required for decisions about ethical considerations and the impact of AI use on the organisation. There must be sufficient expertise within the organisation in these areas.

Difficult interaction between different areas of expertise

The required knowledge is often spread across different people with different areas of expertise. In practice, we found that in many cases the interaction between different areas of expertise remains difficult. Often AI development begins with a focus on technical design decisions. Some of those design decisions involve ethical or organisational issues. Due to limited interaction of areas of expertise, the right people are not involved in making those decisions in time, leaving those decisions (without realising) to the data scientist, for example. When the timely interaction between people with different areas of expertise is facilitated, mutual understanding has proven to be challenging. Another consideration it that it is often unclear who should make the final decisions during development. This makes it more difficult to get the people with the relevant expertise around the table. Responsible decision making requires an interdisciplinary approach. People with different perspectives and expertise need to be able to work together and be aware of each other's responsibilities. This also links to the previous point on improving AI literacy across the board. Ultimately effective collaboration is the only way to incorporate different expertise's in this decision-making process.

(Administrative) focus on AI regulation does not ensure responsible deployment

Much of the administrative focus is on complying with AI regulation, which can create blind spots in the responsible deployment of AI. Naturally, compliance is highly important and it provides valuable guardrails. However, complying with the AI Act is no guarantee of ensuring responsible deployment.

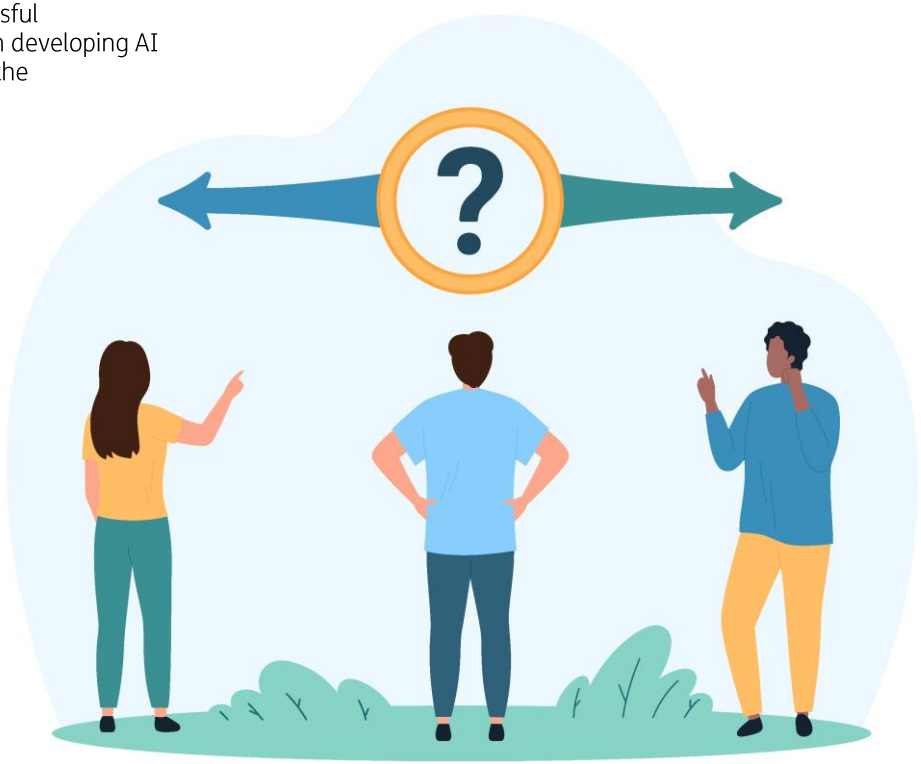
The AI Act has a product perspective, providing high-level product safety requirements in line with a risk-based approach. There are many requirements for technical development and documentation related to performance, robustness and bias, risk management, and human control of the system. Due to the higher level of these requirements, the challenge remains how to implement this in practice and how to enable people to carry out this development and evaluation responsibly. In this respect, many organisations look hopefully to the arrival of harmonized norms and standards from the EU and standardisation organisations that can help them in this respect. At the same time, these standards will not be able to fully address trade-offs for specific use cases. These trade-offs will still call on the ability of staff involved in development to make responsible and informed decisions.

Besides complying with the AI Act, there is a need to look beyond it. Organisations with a public mission, in particular, need to be critical and careful in their processes and the deployment of AI. If an AI system does not qualify as high-risk, it does not mean that it can have high impact or all the requirements of the AI Act are no longer desirable. It is therefore important to be able to make responsible decisions about all types of AI applications, regardless of the type of risk classification prescribed by the AI Act.

Insufficient preparation of the organisation to adopt AI

There are many organisations in the public sector experimenting with AI development, but only a small proportion of AI applications lead to successful implementation. We found many efforts on developing AI applications, but not enough on preparing the organisation to adopt AI.

For example, in many organisations, an AI vision has not yet been established, making it unclear what the organisation does or does not want to do with AI. Insufficient preparation also manifests itself in the lack of resources and appropriate knowledge and skills among employees to develop and use AI responsibly. Finally, we noted that in many cases there are no clear agreements on employee responsibilities. This has the consequence that employees are not empowered to make the responsible decisions when it comes to AI.



Recommendations for responsible decisions about AI

In this section, we provide eight recommendations to help organisations make responsible decisions about AI. First, we recommend setting an explicit objective and seeing AI not as a goal but as a resource. In development, a clear process that includes explicit decision points makes a significant contribution. To facilitate these explicit decision points, discuss and assess the BVRR-balance at each of these point. Then, the importance of interdisciplinary teams must be considered as they bring together different perspectives that contribute to more informed decisions. A safe experimentation environment provides the space to do evaluations, collect substantiating information, and learn lessons on deployment. On organizational level is essential to establish an AI vision and ownership within the organisation, making intended objectives, relevant public values, roles and responsibilities clear. Besides organisations need to invest in the knowledge and skills of its employees. Finally, starting with low-risk applications helps organisations build the right expertise and experience. In the paragraphs below, we explain the recommendations in more detail.

Formulate an explicit objective and consider AI only as a tool

It is essential to have a strong focus on the purpose of the AI application. Deploying AI is not a goal alone, it is a resource. This objective of the AI should therefore be explicitly defined: For what organisational or societal problem does the AI application offer a (partial) solution? AI applications are often developed from a technical analysis of the possibilities of AI. This can lead to a situation where it is not sufficiently clear what the purpose of the AI system is and how it will fit into the work process or people having a different understanding of the purpose. This affects the development as the purpose has implications for the technical optimisation of the system, the design of the interface with humans, and the choice of evaluation methods. Moreover, a clear purpose is necessary when preparing the BVRR balance, for impact assessments and the final decision on whether to deploy the AI application. A lack of a clear purpose description can lead to different opinions on the purpose of the system, design choices that do not align with each other and inadequate evaluation. This highlights the importance of establishing a clear collaborative description of the intended purpose. For an AI application that has multiple goals, make explicit what the goals are and make the various trade-offs for each goal explicit. Clear goals and trade-offs assist the different stakeholders to make suitable decisions regarding the development, evaluation and implementation of the AI application.

Ensure a structured development process including key decision moments

Although many structured development processes27 exist, they do not always pay attention to the explicit decision moments. Install a ‘go/no-go’ between the different phases of the development process. For each decision moment, it should be clear what will be decided at this point, who will make this decision, what information is needed to make this decision and where the decision will be recorded. In this decision moment, there needs to be room for insights and deliberate considerations from different disciplines but also the joint responsibility to (re)define the BVRR balance. In addition, utilize these moments to align with key stakeholders within the organization (strategic, tactical levels) or outside the organization (for instance citizens or interest groups).

Installing these deliberate decision moments helps in the development process to prioritise what needs to be researched in the following phase and project/allocate required resources in an informed manner. In addition, it prevents important decisions from not being explicitly evaluated, decisions being made by people without mandate or decisions not being properly documented.

Discuss and assess the BVRR-balance at key decision moments

To support key decision moments, it is essential to make the BVRR balance concrete for the intended application. In every phase of the development process this balance needs to be discussed with and reviewed by different disciplines. The level of detail will vary, however, per phase. In the initial phase, which asks for a decision whether to investigate AI as a possible solution for the problem at hand, the input for the BVRR balance will be more hypothetical and high-over. An interdisciplinary discussion and concretization will give insight in the potential benefits and risks and therefore the required resources that need to be invested. In each of the following phases, based on questions and assumptions from the earlier BVRR balance, information will be collected, hypotheses will be tested and risks will be addressed where possible. Hence the decision to continue or not will be better grounded and justified than before, overall supporting AI accountability. In the final stage before deployment the necessary insights and information for a legitimate and justified decision should be available.

Invest in an interdisciplinary approach

Interdisciplinary teams are of great value in the development and decision-making process of AI. In the development and deployment of AI, many considerations and decisions must be made. These may be technical considerations, but also, decisions about implementation in the policy process or the risks involved in deploying AI.

If the decision-making process does not involve the necessary disciplines, it can lead to critical points only being identified at a later stage or that ethical considerations in development fall to the technical team. Moreover, various aspects of the organisation need to be prepared for AI adoption, which requires different disciplines to be involved.

In interdisciplinary teams, an issue is analysed from the combined technical, ethical, organisational and legal expertise, where there is a common language and mutual understanding. Learning from each other's expertise and views creates a better understanding of AI in the organisation. Integrating different perspectives in an interdisciplinary team improves the decisions made in the development and evaluation process. This contributes to the smooth and responsible development of new AI applications. Invest in bringing them together and developing mutual understanding.

Provide a safe experimentation environment

To make responsible decisions about the deployment of AI in practice, it is important to facilitate a safe experimentation environment. Such an environment allows you to approximate reality as closely as possible during the test phase, without putting citizens and professionals at risk. An experimentation environment allows you to develop and evaluate an AI application and evaluate its impact on the organisation, user and citizens. An experimentation environment and culture in which learning, evaluation and reflection are central is crucial to build the right level of knowledge and experience in the organisation. By a safe experimentation environment, we also mean a place where people can ask critical questions about the AI application.

Essential requirements are a mindset allowing for innovations to "fail", working with different implementation forms (AI or alternative solution) and continuously making hypotheses and design decisions explicit. Finally, the safe experimentation environment should enable the collection of the necessary information for the final consideration in the BVRR balance and the decision whether or not to deploy the AI solution. The BVRR balance is the trade-off to be repeatedly made between 1) added value and purpose, 2) public values, 3) resources and 4) risks.

The following considerations help set up a safe experiment environment. Determine which aspects of the intended implementation are essential to replicate during the experiment. Consider setting up a representative sample or adding additional noise or relevant scenarios to the experiment. Determine where it can be deployed in the normal way of working, whether the way of working requires redesign or if it should be disconnected from the reallife way of working to protect citizens/professionals. Determine what data needs to be collected in the experiment environment and what experiments are needed to create this information. Create an open and critical working environment with moments of reflection.

Develop an AI vision and clarify ownership of AI applications

Establish an AI vision and associated governance that makes clear what the organisation wants to achieve with AI applications, what it can and cannot be used for, who has what responsibilities and what skills are needed within the organisation to use AI responsibly. In the AI vision, the roles of those who have ownership for the development and implementation of AI applications should be mentioned explicitly. The vision should address these aspects at strategic, tactical and operational levels. This way, the appointed people in the organisation are empowered to make responsible decisions and it is clear how accountability of decisions is facilitated. In addition, an AI vision enables the organisation to prepare for the development and deployment of AI, allowing experiments to successfully land in the organisation. Creating a sense of ownership in the organisation ensures that development and implementation are in line with the AI vision and the reality on the work floor. In addition it ensures there is clarity on responsibilities and mandates. Making a RACI-overview for your organization makes ownership and responsibilities clear and helps to initiate the needed interactions between stakeholders.

Invest in AI knowledge and skills of employees

Ensure employees have a broad understanding of AI and the right skills to work with AI. This is essential to enable employees to make responsible decisions about AI. Note that different roles require different skillsets and try to make a distinction between awareness and capability. Users need a basic understanding of AI to interact responsibly with the AI application (awareness). Employees involved in development need broader and more tactical knowledge and skills to make the decisions in development, to assess the BVRR balance and to be able to effectively interact with other disciplines. This means that the organization requires sufficient technical knowledge, ethical, legal and organisational expertise. In addition, the skills of employees need to be developed to make complex trade-offs, so they can handle cases where they may be conflicting public values or complex considerations in the BVRR balance. Ensure that staff involved in development has the basic understanding of other areas of expertise in addition to their own to promote interdisciplinary working. Ensure that all staff, including legal and policy staff, have a realistic understanding of what AI is and can do. Take note that those using an AI application, are obligated to have a minimum level of AI literacy according to the AI Act

Start and learn with low-risk applications

Start small to gain experience, to learn and to build trust. Here, we mean applications with low risk for citizens and users and with limited changes in organisational and technical processes. Making responsible decisions requires suitable knowledge, experience and expertise. By building this experience within your organisation from a low-risk development, you quickly learn which people should be involved, when which decisions should be made and what kind of information is needed for these decisions. Lower-risk applications offer the benefit of simpler decisions. One is therefore not directly faced with highly complex ethical and organisational considerations. In addition, the risks are more limited if mistakes do appear. At the same time low-risk applications can still have a highly positive impact on the organization, which does make it worthwhile to invest. We strongly recommend to document and share lessons learned and adopt them in following projects. This supports the learning process and enables successful innovations to come to fruition and scale up. Successful innovations boost enthusiasm and trust within and outside the organization.

Conclusion

The use of AI is becoming more common and offers an ever-increasing level of opportunity. In the public sector, steps are being taken to transform this technology into applications to harness this potential. However, we see that in many cases organisations face challenges in facilitating the responsible deployment and development of AI.

We advocate a more explicit focus on the balance between 1) benefits and purpose, 2) public values, 3) available resources, and 4) residual risks. In doing so you can ensure that the added value and alignment with public values outweighs the resource commitment and residual risks. In addition, employees should be better facilitated in making responsible decisions. To that end, interdisciplinarity and mutual understanding are essential. The responsible development and deployment of AI hinges on the decisions made by the people involved.

We therefore urge organisations to implement eight recommendations that contribute to making responsible decisions about AI. This requires decisive action from strategic, tactical and operational levels, an open mind from everyone and good cooperation between disciplines.

1. Formulate an explicit objective and consider AI only as a tool
2. Ensure a structured development process including explicit decision moments
3. Discuss and assess the BVRR-balance at key decision-moments
4. Invest in an interdisciplinary approach
5. Provide a safe experimentation environment
6. Develop an AI vision and clarify ownership of AI applications
7. Invest in AI knowledge and skills of employees
8. Start and learn with low-risk applications

When organisations address these points, they empower their employees to make responsible decisions in the development and deployment of AI. This encourages valuable AI innovations in light of both risks and opportunities and supports in successful adoption in practice.

The AI Oversight Lab

The AI Oversight Lab supports public organisations in making responsible and legitimate decisions about the development and use of algorithms and AI in their work processes. In the AI Oversight Lab, we have been working in an interdisciplinary team with municipalities, government organisations and public organisations for four years. We focus on creating actionable insights from these collaborations and closing knowledge gaps in science and policy. Sharing experiences and lessons within the learning community plays an essential role here.

In the four years since establishing the AI Oversight Lab, we have focused our work on public sector organisations. This has led to broad practical knowledge and scientific contributions of the responsible adoption and use of AI in the public sector. Examples of the challenges central to the collaborations are:

- Technical evaluations of accuracy, robustness and bias/non-discrimination;
- Exploration of preconditions and requirements for organisational embedding of AI within an organisation;
- Research on how public values can be included and evaluated in the development process;
- Developing a holistic evaluation methodology for LLMs, including technical, ethical, legal and organisational aspects;
- Shaping the involvement of users in the development process.
- Shaping the proportional governance of (gen)AI use in the public sector

The lessons and recommendations in this whitepaper are based on our collaborations and scientific developments.

Getting started with responsible AI?
Do you want to successfully deal with challenges you encounter in practice? Increase interdisciplinarity collaboration in your organisation? Or prepare your organisation to implement AI? To sum up, do you want to take the next step in the responsible development of AI with your organisation?

Get in touch to discuss how together we can facilitate your organisation to make more conscious decisions. With our interdisciplinary team we can support on technical, ethical, organisational and interdisciplinary issues. In our work we are committed to enhancing professionalism and knowledge within your organisation.

The AI Oversight Lab is an initiative of TNO, a Dutch independent not-for-profit research institute.

Acknowledgement

- TNO Appl.AI program for the financial contribution to our research;
- Our partners in the AI Oversight Lab, who we thank for their openness, commitment and enthusiasm to innovate responsibly with AI. It is through these collaborations that these key findings have come about.
- All AI Oversight Lab team members for their contributions in the various research projects, and specifically Cor Veenman, Anne Fleur van Veenstra and Marianne Schoenmakers for sharpening our findings and Pippa Jones for her contribution in the English translation.

References

1. World Economic Forum. Unlocking Public Sector Artificial Intelligence. cking Public Sector Artificial Intelligence | World Economic Forum
2. European Commission. (2024). Adopt AI Study.
3. TNO (2024). Quickscan AI in de publieke dienstverlening III. https://publications.tno.nl/publication/34642601/SASNc3ZW/TNO-2024-R11005.pdf
4. Algemene Rekenkamer. (2024). Focus op AI bij de rijksoverheid.
5. TNO (2021). Op zoek naar de mens in AI. Betrek de burger en experimenteer op een verantwoorde wijze.
6. NOS. (2023). DUO mag algoritme niet gebruiken totdat meer bekend is over mogelijke discriminatie
7. Autoriteit Persoonsgegevens. (2023). AP ziet toe op hersteloperatie UWV na illegale inzet algoritme.
8. Wired. (2023). Inside the Suspicion Machine.
9. European Parliament. (2024). EU AI Act: first regulation on artificial intelligence
10. Consider for example recent developments in LLMs.
11. For example: Gemeente Amsterdam. (2022). The Fairness Handbook.
12. For example: UXAI: Design Strategy.
13. Handreiking non-discriminatie by design | Rapport | Rijksoverheid.nl
14. Universiteit Utrecht. De Ethische Data Assistent (DEDA).
15. Universiteit Utrecht. Impact Assessment Mensenrechten en Algoritmes
16. ECP. Aanpak Begeleidingsethiek.
17. High Level Expert Group. (2019). Ethische richtsnoeren voor betrouwbare KI
18. High Level Expert Group. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.
19. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2024). Algoritmekader.
20. For example transparency, privacy, sustainability and fairness. WRR (2011). iOverheid. Opgevraagd van: iOverheid | Rapport | WRR
21. ECP. Aanpak Begeleidingsethiek.
22. M. Steen and I. van de Poel, "Making Values Explicit During the Design Process," in IEEE Technology and Society Magazine, vol. 31, no. 4, pp. 63-72, winter 2012, doi: 10.1109/MTS.2012.2225671.
23. IBDS. (2024). Datadialogen: doordacht van start met data
24. Long, D. & Magerko, B. (2020). What is AI literacy? Competencies and Design Considerations.
25. For example CRISP-DM, SEMMA of NIST

Curious how TNO can support your organisation in responsible development and use of AI?

Contact marianne.schaaphok@tno.nl or wouter.vanderbij@tno.nl

TNO innovation for life