Fall 10-16-2022

FROM COMMERCIAL AGREEMENTS TO THE SOCIAL CONTRACT: HUMAN-CENTERED AI GUIDELINES FOR PUBLIC SERVICES

Stefan Schmager
University of Agder, stefansc@uia.no

FROM COMMERCIAL AGREEMENTS TO THE SOCIAL CONTRACT: HUMAN-CENTERED AI GUIDELINES FOR PUBLIC SERVICES

Research full-length paper

Schmager, Stefan, University of Agder, Kristiansand, Norway, stefan.schmager@uia.no

Abstract

Human-centered Artificial Intelligence (HCAI) is a term frequently used in the discourse on how to guide the development and deployment of AI in responsible and trustworthy ways. Major technology actors including Microsoft, Apple and Google are fostering their own AI ecosystems and they are also providing HCAI guidelines. However, these guidelines are mostly oriented to commercial contexts. This paper focuses on HCAI for public services. Approaching human-AI interaction through the lens of social contract theory we identify amendments to improve the suitability of existing commercially-oriented HCAI guidelines for the public sector. Following the Action Design Research methodological approach, we worked with a public organization to apply, assess, and adapt the "Google PAIR guidelines", a well-known framework for human-centered AI development. Three HCAI considerations that are important for public services were identified and proposed as amendments to the existing guidelines: a) articulation of a clear value proposition by weighing public good vs. individual benefit, b) definition of reuse boundaries for public data given the relationship between citizens and their government, c) accommodation of citizen diversity considering differences in technical and administrative literacy. This paper aims to shift the perspective within human-AI interaction, acknowledging that exchanges are not always subject to commercial agreements but can also be based on the mechanisms of a social contract.

Keywords: *AI in public service, Human-centered AI, Responsible AI, AI Guidelines, Social Contract*.

1 Introduction

It is one of the buzzwords of our current times - artificial intelligence (AI) and it enables humans to approach a myriad of societal challenges. From helping medical professionals to detect diseases, to supporting scientists fighting climate change. At the same time, there are also dystopian scenarios describing machine takeover and the excesses of autonomous lethal weapons (Russell et al., 2021). This ever-widening impact has spawned discussions, considerations, and warnings about ethical, responsible, and sustainable implementation of such advanced technologies (Fjeld et al., 2020; Hagendorff, 2020; Hagerty & Rubinov, 2019; Jobin et al., 2019). The major technology actors including Microsoft, Apple and Google have also been engaged in the discourse and in efforts to direct the development of AI towards harm minimization (Amershi, 2020; McAran, 2021; Shneiderman, 2020b). These actors have been fostering their AI ecosystems contributing tools, datasets and also providing guidelines for human-centered AI. The guidelines operationalize theoretical concepts informing the practices of AI development and deployment within and beyond the respective ecosystems.

An area where AI can contribute to significant improvements is the public sector. Governments are major holders of data that can be harnessed to improve the design and provision of public services. There are already various examples of the use of AI systems in the context of public services (de Sousa et al., 2019; Henman, 2020; Misuraca & Van Noordt, 2020). Governments across Europe are working on the development of public sector AI ecosystems leveraging common data source (such as master data and basic data registers) and introducing standards, guides and frameworks. The special role governmental institutions play in the relationship between the public (the ruled) and public governance (their rulers), makes the use of AI technologies especially sensitive and there is a clear need for supporting AI practices within public sector ecosystems with guidelines for responsible and human-centered AI.

Designers, developers and deployers of AI systems need to safeguard against possibilities of privacy violations, opaqueness or lurking biases that may produce inequitable or discriminatory results (The-Alan-Turing-Institute, 2022). Bekker (2021) describes a Dutch court case on the System Risk Indication (SyRI) system which was found to violate human rights by not being transparent about the algorithms used for fighting social security fraud. Jørgensen (2021) examines benefits and disadvantages for individuals as Denmark increasingly uses predictive analytics to identify fraud and relies heavily on processing vast quantities of data about citizens. These examples highlight the need for a responsible and human-centered implementation of AI technologies in the public service.

The fundamental idea of Responsible AI is that "AI systems should be designed and implemented in ways that recognize and are sensitive to human interaction contexts without infringing on core values and human rights" (Dignum, 2019). Ben Shneiderman (2020a; 2020c), uses the term "human-centered AI" (HCAI) to describe a direction of designing AI systems that support human self-efficacy, promote creativity, clarify responsibility, and facilitate social participation. General human-centered AI principles are operationalized in different guidelines developed in the context of different AI ecosystems (Amershi, 2020; McAran, 2021; Shneiderman, 2020b). However, as most of these guidelines are put forth by major technology companies (Wright et al., 2020), they are primarily targeting commercial settings. Public service organizations, as the name suggests, are under a duty to serve the public and are aiming to the benefit of the entire society and communal good. The relationship between the government and people are governed by the "social contract" which assigns legitimacy to governments and their institutions (Rousseau, 1998). In order to use HCAI guidelines that have been developed for commercial settings for public services, the special relationship between the citizens and the government, their different roles, duties, and obligations need to be considered. There clearly is a need for HCAI guidance specifically adapted to the particularities of the public administration. As (Bekker, 2021) puts it, digital welfare states need guidelines to explore the opportunities they offer but also highlight their legitimate boundaries. One way to address this challenge is to start from an existing framework and apply it to a public sector use case by taking a social contract perspective. Such an approach can lead to the identification of needs for adaptation. This approach laid the foundation for our research question:

How can commercially-oriented guidelines for human-centered AI be adapted for the context of public services through a social contract perspective?

To answer this research question, we started from the guidelines provided by Microsoft, Apple and Google. Google´s PAIR guidebook represents the most comprehensive human-centered AI framework (Wright et al., 2020) so we decided to analyze it for its public sector suitability by applying it to a public service scenario. The practical application of the framework revealed a set of needs for adaptation and amendment. Specifically, we found that it is important to a) articulate a clear value proposition by weighing the public good vs. the individual benefit, b) define boundaries for repurposing public data given the relationship between citizens and their government, c) accommodate user group diversity by considering the different levels of technical and administrative literacy of citizens. These findings contribute to the discourse about responsible AI design and development in the public sector.

The remainder of this paper is structured as follows. First, the related background is provided by exploring the use of AI technology within different European public service organizations and by introducing social contract theory as the theoretical lens for this research. In the next section, the organizational context of this research and the Action Design Research (ADR) methodological approach are presented. This is followed by the analysis and the discussion of the findings which are a set of amendments to the framework to improve its suitability for public service contexts. The paper concludes with a section on limitations and directions for further research.

2 Related Literature

2.1 AI in the public service

AI technologies can improve organizational practices and create new capabilities. A recent analysis of AI technologies deployed in the public sector found them to address one or more of the following use cases: performing comprehensive and accurate predictions; detecting anomalies like welfare fraud; collecting and processing information with Computer Vision; processing and understanding audio and text with Natural Language Processing; classification and profiling (Misuraca et al., 2020). AI technologies hold enormous potential but should also be diligently scrutinized and examined. Sætra (2020) discusses different premises against AI including moral, legitimacy and accountability objections. However, he states that if humans can overcome these objections and preserve participation and legitimacy, they can retain the power to define the fundamental goals in politics, decide on values and imagine a good society. In the following section we provide pointers towards real world applications of AI in public services today, showing that the use of AI in the context of the public sector is not a futuristic scenario but it is already part of actual and current public service offerings.

In Europe AI is widely used in the context of public labor and welfare administration services. The Swedish employment service (Arbetsförmedlingen) employs two AI-enabled tools: an assessment tool and a recruitment prediction tool. The assessment tool supports employment officers to make job market assessments and provides suggestions on forms of support that are most suitable for given cases. The second system is called “Spontansökan” and provides job seekers with a prediction on likely employers in a specific sector and geographical area (Görnerup, 2019). The French employment agency (Pôle Emploi) collaborates with the non-profit startup “Bob Emploi” to support job seekers. The agency provides data including information on regional and seasonal labor demand, salaries, vacancies as well as anonymized profiles of other job seekers to Bob Emploi (RSA, 2019). Based on that data Bob Emploi developed an AI tool for citizens which provides a structured assessment of their main challenges, e.g., if salary expectations are unrealistic and also, indicates related industries or neighboring regions, retraining and further education options as well as suggestion for employers which may hire (Bob-Emploi, 2022). The employment officers use a separate version of the same AI tool. The German Federal Employment Agency (Bundesagentur für Arbeit) employs several AI-enabled software systems (Algorithm-Watch, 2019). The “Verbis” software is the standard program for job placement using intelligent matching technology whereas “PP- tools” calculates a job seeker's job market opportunity. The software “Delta-NT” supports preparing psychological assessments and helps with the evaluation of test results assessing the aptitude for professional orientation (Schwär, 2020). The project named

"3A1" is a component of a larger initiative to automate different parts of the processing of unemployment benefit applications (Geck & Seidl, 2020).

However, some applications of AI in the context of public labor and welfare administration services have also received criticism. In 2019, the Austrian Public Employment Agency (Arbeitsmarktservive) started to use the "PA-MAS" (Personalized Labor Market ASsessment) system (Algorithm-Watch, 2019). It uses personal data (e.g., gender, age, and citizenship) as well as employment data of job seekers and local labor market data to group job seekers into different categories depending on their likelihood to find a new placement. This system received a lot of criticism. Lopez (2021) analyzed the AMS algorithm as a case study and found three types of bias (technical, socio-technical, and societal). Researchers from the Vienna University of Economics and Business called the approach of this system a "prime example for discrimination" (Wimmer, 2018). These examples illustrate the growing use of advanced technologies within public services today and emphasize the need for human-centered guidelines for a responsible implementation.

2.2 Frameworks for responsible AI implementation

A good algorithm or machine-learning model is not enough by itself. It needs to be incorporated into an intuitive user-centric experience, considering workflows and mental models. Developing and iterating together with users and other involved stakeholders helps to spur appropriate development and adoption (Dhasarathy et al., 2020). Existing frameworks and guidelines are building upon this premise, to not only consider the sustainable and responsible development from a purely technical perspective but acknowledge the fact that the model and its output are only pieces of the overall interaction a user has with a system (Amershi et al., 2019; European-Commission, 2019; ICO, 2022; Smith, 2019). These frameworks provide guidance and help to consider all relevant factors, like stakeholders, relationships and processes along the human-centered AI development process (Shneiderman, 2020ac). Wright et al. (2020) analyzed the Human-AI interaction guidelines put forth by three major technology companies Apple, Google, and Microsoft. In their study they developed a unified structure to compare the over 200 guidelines and to identify and evaluate the emphases of each set of guidelines. The analysis revealed that Google's PAIR guidelines (which consist of 6 categories, 20 subcategories and 113 guidelines in total), are the most comprehensive and also, the most balanced in terms of their relative emphasis in the themes of "interface", "deployment", "initial" and "model" compared to the other guidelines. The Human-AI guidelines released by Microsoft and Apple focus mostly on interface and deployment aspects but pay next to no attention to the algorithmic model.

Interestingly, despite the fact that the Google PAIR guidelines are designed to raise socio-technical concerns in the development of novel AI systems, they have not been discussed much in Information Systems literature. Dellermann et al. (2019) remarked that we still lack domain-specific design guidelines which will allow humans to understand processes and needs of AI systems. They mention Google's PAIR guidelines as a development to build appropriate trust and ensure interpretability and transparency of AI. The goal of the guidelines is to focus on the relationship between users and technology. The guidelines consist of a set of recommendations, methods and best practices for designing with AI, which are based on data and insights from Google's experience, other industry experts, and academic research (Google-PAIR, 2022). The guidelines are meant to support developers, designers and deployers of AI products from the inception of a project and throughout the process. Although they are certainly valuable, their application within public services is not straightforward. The guidelines are developed mostly targeting commercial settings not for public services governed by the social contract.

2.3 Social Contract Theory

Social contract theory (SCT) provides diverse accounts of human nature and the social processes that shape conflict, cooperation, and compliance, specifically when applied to the challenges of contemporary public administration (Jos 2006). The rights and obligations of public service organizations to execute governmental functions and enforce legislations are derived from the

conceptual idea of a social contract. A social contract describes a structure of power relationships between governments, its institutions and the people, which assigns legitimacy to governments and institutions based on the consent by the people they govern – the general will (Rousseau, 1998). In Rousseau's understanding consensus is built on complex social processes. It is not merely an intellectual task but requires social, political, and organizational transformation (Jos 2006). Believing that a government, and its institutions will act in the interest of the general will is an integral part of social contract theory and can be seen as a fundamental requirement for a society to work. As a consequence, this engenders a responsibility for the rulers to act in the interest of the ruled, but also ascribes certain rights and obligations to the latter which are manifested in the partial relinquishment of freedom. Governments are required to provide the same services to all people within their jurisdiction, regardless of educational, financial, social, ethnic, or religious background. This implies that public organizations cannot choose whether they want to offer a service or not – unlike private companies. But just like a governmental institution cannot pick and choose with whom to engage with, citizens are also bound to the public service organizations as sole providers for specific services (Junginger, 2016). These provided services are to society first and to individuals secondarily and necessary to make society function as a whole (Junginger, 2016). Furthermore, public service organizations have access to large amounts of citizen data which makes particular rigor, caution, and care necessary to be exercised when considering the purpose of data collection and usage. It is crucial that governments use of AI has good oversight procedures to ensure that the use of AI is done in accordance with overall collective objectives for how they use AI (Henman, 2020). The overall collective objectives are conceptually alike to the idea of a general will. These interrelations are specific to the dynamics between citizens and public service organizations, which renders SCT as a suitable lens for this research into the responsible design, development and deployment of AI and machine learning technologies in the public service domain.

Having its origin in political philosophy, SCT has been less discussed within the Information Systems field compared to other, more business and commercially focused theories. However, although less common, SCT has served as a theoretical vantage point in several studies. In an attempt to create an IS-compatible definition, Grazioli (1998) defines a social contract as an interaction between two parties characterized by a conflict of interest: one party is obligated to satisfy a requirement of some kind, usually at some cost, in order to receive a benefit from the second party. At a higher societal level, a social contract might involve the advantages offered by a firm to society in exchange for the right to exist and even prosper (Dunfee et al., 1999). In their MISQ article, Smith and Hasnas (1999) examine SCT as a potential guiding theory to form "normative business ethics". Describing that when following a SCT approach, the involved actors need to be aware of their contractual roles. Managers should aim to increase social welfare above what it would be in the absence of the existence of corporations, yet without violating basic canons of justice. And in turn, society will be willing to authorize corporate existence only if corporations agree to remain within the bounds of justice. Also, Vial (2019) suggested a turn toward normative theories of ethics, including social contract theory to research the challenges in the context of technology design and use by multiple parties. Building on the work by Martin (2016), Cardon et al. (2021) used a social contract perspective to examine potential issues about consent, data use and privacy in the context of analyzing work meeting recordings. Following a SCT perspective, privacy is contextually and relationship dependent norms which are negotiated agreements within a particular community or situation as microsocial and macrosocial contracts (Dunfee et al., 1999).

3 Context & Methods

3.1 Empirical context

This study is part of a wider research project in collaboration with a public organization in Norway where the government aims to use data and AI in a responsible way to improve services (Norwegian Government, 2019). Within this project we developed a fictitious but realistic scenario for the use of an AI system in the public welfare services context. This allowed us to explore the applicability of the Google PAIR human-centered AI framework for public services. We believe that the design of a

technological system should not neglect any perspective, but rather be understood as a holistic process. Therefore, and based on the analysis conducted by Wright et al. (2020), we selected Google's Human-AI guidelines as the most suitable set of guidelines for our research. In our study we aimed to identify how Google´s PAIR guidebook can be adjusted to become suitable for the public sector context.

The developed scenario is described as follows: the public service organization aims to deploy a novel algorithmic tool capable of predicting the expected duration of an unemployment period for a citizen. For the prediction tool to become fully functional, it requires citizen specific, personal data as input to calculate a probable unemployment duration as output. Within a prototyped web interface of the AI application, a citizen will be provided with information about the AI used and get asked whether she consents to its use to support the decision-making process of the responsible employee by providing a prediction for the unemployment duration. This scenario was selected because it involves sensitive personal data which is a common hurdle in the deployment of AI technologies in the public service.

3.2 Action Design Research

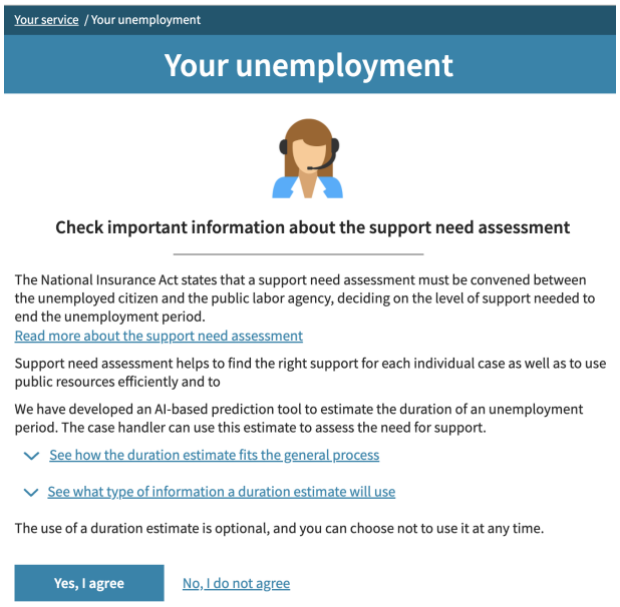
The Action Design Research (ADR) approach, described by Sein et al. (2011) provides the methodological framework for this research. By following ADR, close collaboration with an organization enables continuous involvement and reflection within the organizational context with the understanding that the specific field problem is a knowledge-creation opportunity. ADR allows both the research team as well as the organizational actors (e.g. stakeholders, end-users, practitioners) to shape artifacts over the research lifecycle (Haj-Bolouri et al., 2018). Assessment and iteration happened in bi-weekly collaboration sessions. In each session, design choices were presented, scrutinized, and discussed. This mode of close collaboration and coordination was the basis for advancement through the different stages of the ADR process. ADR emphasizes that the ensemble artifact creation and evaluation is informed by theory. We picked social contract theory as our theoretical lens and existing human-centered AI guidelines as a starting point to guide the initial design. In iterative "building", "intervention" and "evaluation" (BIE) cycles, the design was adapted and improved in close collaboration with the organization. In a process of continuous learning and improvement we went through several iterations over the set of HCAI guidelines by directly applying them to an evolving ensemble artifact, embodied in an interactive prototype to facilitate debate and feedback. The developed prototype consists of a user interface, mimicking a public employment agency portal, with a particular interaction sequence starting from a notification about the optional use of an AI-based prediction, to different types and levels of information about the prediction model as well as a consent decision. This evolvement is described as "guided emergence", as it captures underlying tensions between the stakeholders enabling reciprocal shaping of the artifact (Haj-Bolouri et al., 2018). During ADR projects reflections and learnings are conceptually generalized to be applicable to a broader class of problems. Conscious reflection on the problem understanding, the applied theoretical lenses and the emerging design is required to evaluate possible generalizations and knowledge contributions (Sein et al., 2011). The process of generalization led to formulating specific PAIR amendments for the context of public services. Work on conceptual generalizations is planned to continue within the overarching research project.

4 Analysis: from a commercial relationship to the social contract

To be able to assess the suitability of the Google PAIR guidelines for the public service context, we applied the guidelines within a fictitious scenario defined together with the collaborating organization. During iterative BIE cycles, we used the guidelines to inform the concept, design, and development of an interactive prototype for a system addressing the defined scenario. For each iteration, the guidelines´ manifestation was evaluated. This practical examination of the guidelines revealed three suggested amendment themes to the guidelines to improve their suitability within the public service domain: 1) articulation of a clear value proposition by weighing the public good versus the individual benefit; 2) definition of boundaries for repurposing public data given the relationship between citizens and their

government; 3) accommodation for user group diversity by considering the different levels of technical and administrative literacy of citizens. A screenshot from the prototype is provided in figure 1.

Figure 1 Prototype screenshot



The first chapter of the Google PAIR guidebook discusses the identification of a user need and if this need can be addressed using AI capabilities. Further, it provides guidance on how the decision of the AI system can provide value and what type of enhancement would be appropriate. In the public welfare service use case, this framing needs to be adjusted, from "user need" to "citizens benefit". Although this change in perspective seems minor, it addresses the consideration if an AI infused system is appropriate in the first place. As an example, efficient use of public resources is less of a "user need", but it certainly can be seen as a benefit for the individual citizen as well as for the wider society.

Google PAIR Consideration	**Amendment for public service context**
Find the intersection of user needs & AI strengths. Solve a real problem in ways in which AI adds unique value.	Does the use of an AI system consider the implications for the citizen as well as the common good? Is there a conflict of interest between the benefits for the public and potential relinquishment of freedom for the individual.
Assess automation vs. augmentation. Automate tasks that are difficult, unpleasant, or where there's a need for scale. Augment tasks that people enjoy doing, that carry social capital.	Does the envisioned system leverage AI capabilities and serve the needs of citizens, the public service employee or maybe even both?
Design & evaluate the reward function. Optimize for long-term user benefits by imagining the downstream effects of your product. Share this function with users if possible.	When defining "right" and "wrong" predictions, consider how to weigh "precision" vs. "recall". E.g., would a "false" prediction disqualify a citizen from an eligible service?

Table 1. User Needs + Defining Success

In the second chapter of the guidebook, the concepts data collection and quality evaluation are discussed. In general, proper diligence and care are the basic premises for any responsible AI system that uses data. However, in the case of the public welfare system, scrutiny is even more important due to the special relationship between the citizens, their data and the governmental institutions using it. The exchange between both parties is not subject to any commercial agreement but underlies the general mechanisms of the social contract. As described by (Junginger, 2016), both parties don't have a free choice whether to interact with each other or not. This composition engenders a responsibility for the rulers to act in the

interest of the ruled and due to this responsibility, particular rigor, caution, and care need to be exercised when considering the purpose of data collection and usage.

Google PAIR Consideration	**Amendment for public service context**
Plan to gather high-quality data from the start. Planning for data gathering and preparation, to avoid the effects of poor data choices further downstream in the AI development cycle.	Evaluate the availability of existing data. Should existing data be used? Are the data needs aligned? Should novel data be collected with the data needs understood? The diversity within the audience also needs to be reflected in the data collection.
Translate user needs into data needs. Determining the type of data needed to train the model. Considering predictive power, relevance, fairness, privacy, and security.	Consider "features" and "labels" of the public data which will affect the accuracy and robustness of the model in the light of governmental boundaries and obligations. Is the data suitable for the use case?
Source your data responsibly. Ensuring the data and collection method is appropriate for the project.	Has the dataset, or the data collection method, the required quality for the use case? Was the purpose of data collection compatible with the initial data collection? Is there additional consent required from citizens for the data to be used?
Prepare and document your data. Preparing the dataset and documenting its contents and the decisions made while gathering and processing the data.	Data sources and data changes history need to be transparently documented to ensure a responsible and appropriate use of public data.
Design for labelers & labeling. For supervised learning, having accurate data labels is crucial to getting useful output from your model. Thoughtful design of labeler instructions helps to get better quality labels and better output.	Data labeling needs to undergo the same scrutiny as the data sourcing itself. When repurposing existing data, critically assess whether the labeling was done with the same goals in mind.
Tune your model. Once the model is running, interpret the output to ensure it is aligned with product goals and user needs. If not, you explore potential issues.	Model tuning and the decision of deployment requires careful consideration. Changes in the model might result in different treatment between citizens, although equal treatment is their legal right.

Table 2. Data Collection + Evaluation

Because public welfare organizations need to serve the entirety of the public, by nature this covers a very diverse group of people. This means that the design and development of AI systems used in the public service needs to cater for diverse levels of technical literacy, as well as administrative literacy. Döring (2021) describes administrative literacy as the competence to obtain, process, and understand information and services provided by public organizations to draw appropriate conclusions and take informed decisions. To address different levels of technical literacy, it is important to build upon common and established mental models and leverage them for the implementation of AI systems. A mental model is a cognitive shorthand, helping people to make sense of a concept in the world (Cooper et al., 2014). Using potentially existing mental models will help a broader group of citizens to transfer existing knowledge and understanding into a new situation and create a more familiar context. Understanding the dynamics of the interactions is key to shaping HCAI experiences to create benefits for both sides of the interaction. One approach to address these types of literacy can be to position the system into a larger context of administration as well as interactions. This helps to explain how the input and output of the system fits the overall process. This overview can help the citizens to not only see the technology use in isolation but fathom its wider application.

Google PAIR Consideration	**Amendment for public service context**
Set expectations for adaptation. AI allows systems to adapt, optimize, and personalize. Familiar mental models help users feel comfortable.	Provide insight into the governmental mandate related to the use of the system. Explain the role an AI algorithm has within the overall process. Explain the benefit, not the technology.

Onboard in stages. Introducing users to an AI product, explaining what it can and can't do, how it may change, and how to improve it.	Provide an adequate level of information in stages. Be transparent about potential personal data use. Explain which data would be used, and how.
Plan for co-learning. Feedback to an AI product will adjust models and change how people interact with it. Mental models will change similarly.	Introduce feedback mechanisms and explain how the feedback will inform and adjust the model. Citizen feedback provides a format of direct participation.
Account for expectations of human-like interaction. Communicate the nature and limits of the AI product to set realistic expectations and avoid deception.	Full disclosure of the AI system about its use, form, and constraints. Avoid any form of deception and support citizens to form intelligible mental models.

Table 3. Mental Models

The PAIR guidebook discusses furthermore concepts like Explainable AI and Trustworthy AI. Both are discussed extensively in other places and would warrant an even more detailed inspection of themselves (Doran et al., 2017; Ehsan & Riedl, 2019; Glikson & Woolley, 2020). But this would exceed the purpose of this study. Therefore, this work focuses only on the specific recommendations from the guidebook. Trust calibration is a key concept in human-AI interaction. One dimension for citizens to calibrate their trust is making data usage transparent in an understandable format. This means the system needs to expose which data is used, why it is used and where it comes from. Giving a global explanation, e.g., with providing feature importance visualizations can be a useful approach to help citizens calibrate their trust in the system. However, since machine learning models usually use a large number of different features, it might be expedient to combine types of information into higher level groups, which could be explored and drilled into if desired. The "Shneiderman mantra", which states "Overview first, zoom and filter, details on demand" suggests an easy to navigate arrangement which caters for different needs of information (Shneiderman, 2003). Those needs for information can also be affected by the overall trust in a government and its institution (Grimmelikhuijsen et al., 2013). Therefore, when defining which level of information detail would be appropriate, such characteristics need to be considered as well.

Google PAIR Consideration	**Amendment for public service context**
Help users calibrate their trust. Since AI products are based on statistics and probability, users shouldn't trust the system completely. Provide explanations, to help them understand when to trust system predictions and when to use own judgment.	Trust is a multidimensional construct, which can be connected to all the defined amendments. It is important to clarify the value proposition of a system, its boundaries, and any legal obligations on both sides. In addition, incorporate ways to contest the outcome of a system to provide a sense of control.
Plan for trust calibration throughout the product experience. Establishing the right level of trust takes time. AI can change and adapt over time, and so will the user's relationship with the product.	Add explanations within the relevant context. Reuse established elements and patterns to manage trust evolvement.
Optimize for understanding. There may be no explicit, comprehensive explanation for the AI output. Or the reasoning behind a prediction may be knowable, but difficult to explain in simple terms.	Finding the balance between technical accuracy and understandability is critical. Serving citizens with a vast variance of technical literacy requires careful information architecture. Involving professional copywriters, to collaborate with data scientists crafting simple but accurate explanations.
Manage influence on user decisions. If, when, and how the system calculates and shows confidence levels can be critical in informing the user's decision making and calibrating their trust.	Highlight (partial) power of decision of the citizen. From consenting to the use of personal data to the contestability of a systems output.

Table 4. Explainability + Trust

It is important to gather different types of feedback to be able to draw the right conclusions about expectations, values, and concerns of citizens. Again, the special relationship between the sovereign, i.e., the governmental agency and the citizens influences these, as in a commercial context, there are different considerations than when technology is applied within the public sector. In the example scenario, the prediction model works as an additional source of information for the public service employee, which only receives a recommendation. It is therefore not an automated decision-making

system and can be understood as an "Human in the loop" system. However, it still needs to be considered, if seeking consent for the use of personal data from the citizen, would be appropriate and useful. Although dissenting to the use of an AI system, can already be seen as implicit feedback, implementing a consent mechanism creates a natural opportunity and timing to ask for further feedback. Asking for feedback should be framed as a request for help and participation, explaining that any details about why a citizen has opted out of the use of the prediction model will help to improve the system for everyone. Yet again, it should not be tried to enforce any feedback, as it will impact the general perception of the system and the organization in general. Also, it would compromise the integrity of the feedback, as citizens might end up giving any feedback just to be able to complete the interaction.

Google PAIR Consideration	**Amendment for public service context**
Align feedback with model improvement. Clarify the differences between implicit and explicit feedback and ask useful questions at the right level of detail.	Implement mechanisms for the citizen to provide feedback, but do not enforce it. Plan for further investigation into the feedback.
Communicate value & time to impact. Understand why people give feedback so you can set expectations for how and when it will improve their user experience.	Explain how the provided feedback can be used to improve the model.
Balance control & automation. Give users control over certain aspects of the experience and allow them to easily opt out of giving feedback.	As with the general use of an AI system, a citizen should feel in control about the amount and type of feedback she wants provide.

Table 5. Feedback + Control

To help citizens understand the capabilities but also the constraints of an AI system, it is crucial to provide them with the necessary information. This is especially relevant when planning for potential errors or expectation failures evoking disagreement. Citizens need to be able to identify, contest or report deficient outcomes or flawed data. The large variance in technical literacy needs to be a key consideration to avoid feelings of belittlement or intimidation and dissent or opt-out mechanisms should be planned and designed as early as possible.

Google PAIR Consideration	**Amendment for public service context**
Define "errors" & "failure". What the user considers an error is deeply connected to their expectations of the AI system. How these interactions are handled establishes or corrects mental models and calibrates user trust.	Be transparent about system constraints. Provide the citizen with the necessary information to identify, contest or report erroneous data or assessments.
Identify error sources. With AI systems, errors can come from many places, be harder to identify, and appear to the user and to system creators in non-intuitive ways.	Use an appropriate level of technical detail to explain where an error could come from, without being patronizing and/or intimidating. Provide ways to give feedback on errors.
Provide paths forward from failure. Creating paths for users to take action in response to the errors they encounter encourages patience with the system, keeps the user-AI relationship going, and supports a better overall experience.	Provide ways to contest and report errors. Allow users to opt-out of the automated suggestion process anytime, to take back full control.

Table 6. Errors + Graceful Failure

5 Discussion

Available frameworks, guidelines and checklists for responsible AI implementation provide useful resources for the discussion about how to design, develop and deploy these modern technologies. The major technology actors including Microsoft, Apple and Google have provided such guidelines (Amershi, 2020; McAran, 2021; Shneiderman, 2020b) however, a commonality across these guidelines is their orientation to commercial contexts. We believe that the use of AI within public services requires a tailored angle. As a theoretical contribution, this work adds to the academic discourse by applying social Contract Theory as a novel lens to the timely topic of responsible AI development specifically for the public sector. Social Contract Theory recognizes a different relationship between governments, their

institutions and citizens compared to the relationship between a commercial company and its customers. Public service organizations, as the name suggests, are under a duty to serve the public. They obtain their legitimization by the consent of the public (Rousseau, 1998), which implies their interests should be driven by the benefit for the entire society and communal good. Governmental organizations are required to weight the public good against the individual benefit, which has fundamental implications to the design and deployment of algorithmic systems, as potentially conflicting interests need to be considered. In addition, as described by Junginger (2016), the concept of a "market" doesn't exist in the interaction between the governmental organization and the citizens. Neither in the form of a "target group", which usually allows institutions to provide services only to those who fit a pre-defined profile, nor as competition, providing a customer with a variety of service providers to choose from. This mutual dependence yields responsibilities in relation to the repurposing of available data and draws new boundaries for the processing of sensitive citizen information (de Sousa et al., 2019; Misuraca & Van Noordt, 2020). Legal frameworks and requirements add a dimension to the design and development process which applies to all parties involved.

We applied the original Google PAIR guidelines to the design and development of an interactive prototype addressing a fictitious scenario in the public sector. Through several BIE cycles, the prototype evolved and improved. Within each iteration, the implementations of the guidelines were discussed and gaps as well as necessary enhancements of the guidelines identified. Following the ADR methodology, these reflections and learning needs to be translated into a broader conceptualization (Sein et al., 2011). Building upon the findings from the BIE cycles, we formulated three amendment themes, which can be understood as generalized considerations that are required for a responsible AI implementation specifically within the public service context. These amendment themes are defined as follows: a) the articulation of a clear value proposition by weighing public good vs. the individual benefit, b) the repurposing of public data subject to special scrutiny due to the responsibilities public organizations inherit, c) the diversity of citizen needs that must be reflected in the design of an AI system. We believe that these identified amendments can contribute to the responsible design, development, and deployment of AI systems in the public service context. By leveraging an already defined set of HCAI guidelines within an ADR project, we also contribute methodologically by demonstrating how generalized outcomes need not always start from a "blank page" but can also be performed by revisiting existing principles and recommendations and amending or adapting them based on the ADR learnings. This work can be useful for practitioners that work on the introduction of AI in public sector contexts. Although our findings originate from one specific framework, their applicability is not restricted to it. In their current state, the amendments should be understood as an additional set of considerations, supporting practitioners developing or evaluating the design of an AI system through the lens of social contract theory.

6 Limitations and Further Research

In the research reported in this paper we adapted and amended the guidelines from a well-established HCAI framework (PAIR guidebook) to shift the perspective from a commercial relationship between providers and buyers to one following the dynamics of a social contract. We applied the HCAI framework to a fictitious, yet realistic use case from public welfare. An acknowledged limitation of this work is the use of a fictional scenario instead of a real-world use case. We aim to apply our findings to actual use cases in later phases of the project. A real case can reveal more nuances, particularities and special public sector conditions. Additionally, investigating the adaptability of HCAI guidelines across different types of public services is another important and exciting avenue for further research. Public services can vary significantly in terms of their sensitivity and citizen impact. Such variations call for different, service-specific configurations and it is important to research how this can be reflected in HCAI guidelines.

References

Algorithm-Watch. (2019). *Labor*. Retrieved 02. Mar from https://atlas.algorithmwatch.org/report_en/labor/

Amershi, S. (2020). Toward Responsible AI by Planning to Fail. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., . . . Inkpen, K. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 chi conference on human factors in computing systems,

Bekker, S. (2021). Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. In *Netherlands Yearbook of International Law 2019* (pp. 289-307). Springer.

Bob-Emploi. (2022). Retrieved 21 Feb from https://www.bob-emploi.fr/

Cardon, P., Ma, H., Fleischmann, A. C., & Aritz, J. (2021). Recorded work meetings and algorithmic tools: Anticipated boundary turbulence. Proceedings of the 54th Hawaii International Conference on System Sciences,

Cooper, A., Reimann, R., Cronin, D., & Noessel, C. (2014). *About face: the essentials of interaction design*. John Wiley & Sons.

de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly*, *36*(4), 101392.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637-643.

Dhasarathy, A., Jain, S., & Khan, N. (2020). When governments turn to AI: Algorithms, trade-offs, and trust. *https://www. mckinsey. com/industries/public-and-social-sector/our-insights/when-governments-turn-to-ai-algorithms-trade-offs-and-trust*.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Döring, M. (2021). How-to bureaucracy: A concept of citizens' administrative literacy. *Administration & Society*, *53*(8), 1155-1177.

Dunfee, T. W., Smith, N. C., & Ross Jr, W. T. (1999). Social contracts and marketing ethics. *Journal of marketing*, *63*(3), 14-32.

Ehsan, U., & Riedl, M. O. (2019). On design and evaluation of human-centered explainable AI systems. *Glasgow'19*.

European-Commission. (2019). *Ethics guidelines for trustworthy AI*. Directorate-General for Communications Networks, Content

Technology. https://doi.org/doi/10.2759/177365

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*(2020-1).

Geck, B., & Seidl, S. (2020). *Automatisierung in der Arbeitslosenversicherung*. https://www.egovernment-wettbewerb.de/praesentationen/2020/BA_Modernisierer.pdf

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Google-PAIR. (2022). *People+AI-Guidebook*. Retrieved 23 February from https://pair.withgoogle.com/

Görnerup, O. (2019). *Ai och Dataanalys gör spontana jobbansökningar smartare*. Retrieved 05. Oct from https://www.ri.se/sv/berattelser/ai-och-dataanalys-gor-spontana-jobbansokningar-smartare

Grazioli, S. (1998). Facilitating the Detection of Strategically Manipulated Information: A Field Test of Social Contract Theory.

Grimmelikhuijsen, S., Porumbescu, G., Hong, B., & Im, T. (2013). The effect of transparency on trust in government: A cross-national comparative experiment. *Public administration review*, *73*(4), 575-586.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99-120.

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*.

Haj-Bolouri, A., Purao, S., Rossi, M., & Bernhardsson, L. (2018). Action Design Research in Practice: Lessons and Concerns. ECIS,

Henman, P. (2020). Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration*, *42*(4), 209-221.

ICO. (2022). *Explaining decisions made with AI* Retrieved 4 October from https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Jørgensen, R. F. (2021). Data and rights in the digital welfare state: the case of Denmark. *Information, Communication & Society*, 1-16.

Junginger, S. (2016). *Transforming Public Services by Design: Re-orienting policies, organizations and services around people*. Routledge.

Lopez, P. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, *10*(4), 1-29.

Martin, K. (2016). Understanding privacy online: Development of a social contract approach to privacy. *Journal of business ethics, 137*(3), 551-569.

McAran, D. (2021). Privacy, Ethics, Trust, and UX Challenges as Reflected in Google's People and AI Guidebook. International Conference on Human-Computer Interaction,

Misuraca, G., & Van Noordt, C. (2020). AI Watch-Artificial Intelligence in public services: Overview of the use and impact of AI in public services in the EU. *JRC Working Papers*(JRC120399).

Misuraca, G., van Noordt, C., & Boukli, A. (2020). The use of AI in public services: results from a preliminary mapping across the EU. Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance,

Norwegian Government. (2019). A common ecosystem for national digital collaboration. In *One digital public sector*. https://www.regjeringen.no/en/dokumenter/one-digital-public-sector/id2653874/?ch=6

Rousseau, J.-J. (1998). The social contract (HJ Tozer, Trans.). *Hertfordshire: Wordsworth Classics of World Literature*.

RSA. (2019). *Bob Emploi: An open source application that uses AI to give jobseekers personalized, data-driven advice and coaching*. Retrieved 21 Feb from https://www.thersa.org/projects/archive/economy/future-work-awards/winners/bob-emploi

Russell, S., Aguirre, A., Javorsky, E., & Tegmark, M. (2021). Lethal Autonomous Weapons Exist; They Must Be Banned. *IEEE Spectrum*,(June 16, 2021).

Sætra, H. S. (2020). A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society*, *62*, 101283. https://doi.org/10.1016/j.techsoc.2020.101283

Schwär, H. (2020). *Jobcenter Setzen auf künstliche Intelligenz - Die Folgen für bewerber könnten fatal sein.* Retrieved 22 Feb from https://www.businessinsider.de/tech/jobcenter-setzen-laut-forschern-auf-kuenstliche-intelligenz-die-folgen-fuer-bewerber-koennten-fatal-sein-2019-4/

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. *MIS quarterly*, 37-56.

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization* (pp. 364-371). Elsevier.

Shneiderman, B. (2020a). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

Shneiderman, B. (2020b). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

Shneiderman, B. (2020c). Human-centered artificial intelligence: three fresh ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109-124.

Smith, C. J. (2019). Designing trustworthy AI: A human-machine teaming framework to guide development. *arXiv preprint arXiv:1910.03515*.

Smith, H. J., & Hasnas, J. (1999). Ethics and information systems: The corporate domain. *MIS quarterly*, 109-127.

The-Alan-Turing-Institute. (2022). *Public Policy*. Retrieved 20 February from https://www.turing.ac.uk/research/research-programmes/public-policy

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The journal of strategic information systems*, *28*(2), 118-144.

Wimmer, B. (2018). *Der Ams-Algorithmus ist ein „Paradebeispiel für Diskriminierung"*. Retrieved 22. Feb from https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421

Wright, A. P., Wang, Z. J., Park, H., Guo, G., Sperrle, F., El-Assady, M., . . . Chau, D. H. (2020). A comparative analysis of industry human-AI interaction guidelines. *arXiv preprint arXiv:2010.11761*.