Analysing and organising human communications for AI fairness assessment

Use cases from the Dutch Public Sector

Dankloff, M.; Skoric, V.; Sileno, G.; Ghebreab, S.; van Ossenbruggen, J.; Beauxis-Aussalet, E.

DOI
10.1007/s00146-024-01974-4

Publication date
2025

Document Version
Final published version

Published in
AI and Society

AI & SOCIETY (2025) 40:2347–2367
https://doi.org/10.1007/s00146-024-01974-4

OPEN FORUM

Analysing and organising human communications for AI fairness assessment

Use cases from the Dutch Public Sector

Mirthe Dankloff[1] · Vanja Skoric[2] · Giovanni Sileno[2] · Sennay Ghebreab[2] · Jacco van Ossenbruggen[1] · Emma Beauxis-Aussalet[1]

Received: 9 November 2023 / Accepted: 30 April 2024 / Published online: 10 June 2024

Abstract
Algorithms used in the public sector, e.g., for allocating social benefits or predicting fraud, often require involvement from multiple stakeholders at various phases of the algorithm's life-cycle. This paper focuses on the communication issues between diverse stakeholders that can lead to misinterpretation and misuse of algorithmic systems. Ethnographic research was conducted via 11 semi-structured interviews with practitioners working on algorithmic systems in the Dutch public sector, at local and national levels. With qualitative coding analysis, we identify key elements of the communication processes that underlie fairness-related human decisions. More specifically, we analyze the division of roles and tasks, the required skills, and the challenges perceived by diverse stakeholders. Three general patterns emerge from the coding analysis: (1) Policymakers, civil servants, and domain experts are less involved compared to developers throughout a system's life-cycle. This leads to developers taking on the role of decision-maker and policy advisor, while they potentially miss the required skills. (2) End-users and policy-makers often lack the technical skills to interpret a system's output, and rely on actors having a developer role for making decisions concerning fairness issues. (3) Citizens are structurally absent throughout a system's life-cycle. This may lead to unbalanced fairness assessments that do not include key input from relevant stakeholders. We formalize the underlying communication issues within such networks of stakeholders and introduce the *phase-actor-role-task-skill* (PARTS) model. PARTS can both (i) represent the communication patterns identified in the interviews, and (ii) explicitly outline missing elements in communication patterns such as actors who miss skills or collaborators for their tasks, or tasks that miss qualified actors. The PARTS model can be extended to other use cases and used to analyze and design the human organizations responsible for assessing fairness in algorithmic systems. It can be further extended to explore communication issues in other use cases, design potential solutions, and organize accountability with a common vocabulary.

Keywords Communication framework · Fairness · Transparency · Accountability · Public sector · Qualitative user study

✉ Mirthe Dankloff
m.e.dankloff@vu.nl

✉ Emma Beauxis-Aussalet
e.m.a.l.beauxisaussalet@vu.nl

Vanja Skoric
v.skoric@uva.nl

Giovanni Sileno
g.sileno@uva.nl

Sennay Ghebreab
s.ghebreab@uva.nl

Jacco van Ossenbruggen
jacco.van.ossenbruggen@vu.nl

[1] Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

[2] Informatics Institute, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, The Netherlands

1 Introduction

Algorithms are increasingly being used for predicting various forms of public sector services such as allocating benefits in the domains of education, (mental) health, and detecting fraud in allowances and taxes (Rodolfa et al. 2021, 2020; Williamson 2016; Van Veenstra et al. 2019; Hoekstra et al. 2021). These applications can be beneficial, but can also have detrimental consequences for citizens in high-stake scenarios, such as fraud detection and risk assessment. Notorious examples where incorrect predictions led to wrongful accusations of minorities is the COMPAS case in the US.[1] Other cases in the Netherlands, such as the SyRI-case[2] and the Childcare Benefit Scandal[3] sparked a lot of debate and the latter eventually led to the resignation of the Dutch government.[4]

Nowadays, the problem of fairness in AI,[5] is widely recognized in well-established legal and ethical guidelines (Commission et al. 2020; European Commission and Technology 2019; Commission 2021).

Previous research suggests that practitioners face difficulty in adapting the proposed fairness guidelines to their use cases, and that practical tools and methods are still needed (Floridi et al. 2018; Amarasinghe et al. 2020; Holstein et al. 2019). There can be a disconnection with the fairness guidelines and the additional ethical and legal frameworks that public practitioners already need to deal with (Fest et al. 2022). At the same time, the data literacy at public organizations may not be mature enough to fully recognize ethical issues when applying data practices (Siffels et al. 2022).

According to the European Commission's Ethics guidelines on trustworthy AI (European Commission and Technology 2019), an important step in supporting AI fairness assessment includes involving and educating all stakeholders about their roles and needs throughout the AI system's life-cycle. Algorithmic models are always part of a process driven by the stakeholders' choices and norms (Suresh and Guttag 2021). Throughout the algorithms' life-cycle, design decisions are made that codify crucial social values, from data-oriented phases (eg. which data to collect and how to label them) to model-oriented phases (eg. feature engineering, what is considered a good model), to execution-oriented phases (e.g. how to assess model outcomes) (Lee et al. 2019; Amershi et al. 2019; Haakman et al. 2020). These choices can influence the results and can therefore have an impact on people, including those whose data is used (Barocas et al. 2021).

In the public sector, the need for involving diverse actors and stakeholders in algorithm development is important for ensuring that public interest is prioritized, and potential harms are minimized (Stapleton et al. 2022). For instance, when allocating benefits in the public sector, it has to be decided which data features represent 'eligibility' for a social benefit (Barocas et al. 2019). Also, the punitive (eg. detecting a crime) or assistive (e.g. allocating a benefit) nature of policy interventions might require balancing trade-offs between false-positive or false-negative rates (Saleiro et al. 2018; Rodolfa et al. 2020). Involving diverse stakeholders, roles, and task divisions are also design choices that can be evaluated for fairness assessment.

A solely technical approach to fairness is therefore insufficient when it comes to addressing all the normative decisions made by the involved stakeholders throughout the algorithm's lifecycle (Lee et al. 2019). In this research, we take a socio-technical perspective on the algorithmic system and refer to *fairness assessment* as the process of evaluating the extent to which the algorithm and its output is free from bias, discrimination, and other forms of unfairness towards different groups, individuals, or communities (Barocas et al. 2019; Friedler et al. 2019; Corbett-Davies et al. 2017). We investigate these fairness assessments through the choices and practices applied by stakeholders at each phase of the algorithm's life-cycle that can potentially lead to undesirable effects of bias and unfair outcomes (Suresh and Guttag 2021). We will analyze the internal communication processes[6] between direct stakeholders[7] involved in fairness-related human decisions, by identifying the roles, the divisions of tasks, the required skills, and the potential

[1] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS): the software used to predict the risk of a person recommitting a crime was more inclined to falsely accuse African-American offenders than Caucasian offenders, see for instance (Mehrabi et al. 2021; Fass et al. 2008)

[2] SyRI legislation in breach of European Convention on Human Rights. https://edu.nl/xjubf

[3] The Dutch childcare benefit scandal, institutional racism, and algorithms. European parliamentary questions https://edu.nl/y3h3j.

[4] Dutch Government resigns over Child Benefit Scandal https://www.theguardian.com/world/2021/jan/15/dutch-government-resigns-over-child-benefits-scandal.

[5] Fairness in this context, refers to *fair outcomes* a principle which indicates an absence of prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics through algorithmic decision-making (Mehrabi et al. 2021).

[6] With *internal communication* we refer to communication between collaborating partners such as decision-makers, end-users, developers, etc. as opposed to *external communication*, which refers to communication with the general public (Fest et al. 2022).

[7] Direct stakeholders are the people who use or operate an AI system, as well as anyone who could be directly affected by someone else's use of the system, intentionally or not. In some cases, customers are direct stakeholders, while in other cases they are not. Madaio et al. (2022)

communication challenges between diverse actors occurring throughout the algorithm's life cycle.

The research questions we address are the following:

- *RQ1:* Which actors, roles, tasks, and phases can be identified in multi-stakeholder interactions throughout an algorithm's life-cycle when assessing fairness?
- *RQ2:* How is information about fairness assessments communicated between diverse actors involved throughout the algorithm's life-cycle?
- *RQ3:* What characteristics of actors, roles, and tasks lead to communication issues in fairness assessments?

To answer these questions, we conducted an ethnographic study via 11 semi-structured in-depth interviews with public practitioners working on algorithmic systems in fraud and risk assessment use cases. We identified who makes decisions about what, and in which phase of the algorithm's life-cycle. Additionally, we analyzed the interview transcripts to identify the (co-)occurrences of mentioned actors, roles, tasks, phases, and challenges and labeled them through *in vivo*, descriptive, and process qualitative coding analysis (Saldaña 2013). From the emerging coding patterns, we found a lack of information exchange among different roles and actors regarding the interpretation of the model outcome and a lack of involvement of the relevant roles and actors at the right phase. These communication issues indicate inadequate model governance such as missing roles, skills, and information exchange which can lead to misinterpretation of model outcomes, and potentially to the inability to recognize and address fairness issues throughout the algorithm's life-cycle. These findings are characterized and structured in the *phase-actor-role-task-skill* (PARTS) model to characterize the structure of the communication processes identified from the qualitative coding.

PARTS enables to specify the relations and interactions between the identified concepts. Moreover, PARTS is introduced with the longer term objective to be reused and possibly extended with other use cases concerning multi-stakeholder collaborations, and to promote a common vocabulary around fairness assessments.

2 Related work

In the Dutch public sector, multiple stakeholders from private and public organizations often collaborate at various phases of the algorithm's life-cycle (Hoekstra et al. 2021; Spierings and van der Waal 2020).

Stakeholders can be involved at different moments and places depending on their roles and tasks and, therefore, responsibilities can become diffused across a network of multiple actors simultaneously (Wieringa 2020; Bovens 2007).

Related empirical field research has been conducted to investigate data practices of Dutch local governments (Siffels et al. 2022; Fest et al. 2022; Jonk and Iren 2021). For example, Siffels et al. (2022) argue that with the process of decentralization, many tasks from the central government were delegated to municipalities without giving them more resources and capacities.

As a result, municipalities use data practices to deal with additional tasks and to distribute limited (social) resources. Due to a lack of data literacy, public servants are unable to recognize ethical issues and thus seek collaboration with external partners.

The latter can lead to "The problem of many hands" which is sometimes referred to the decreased ability to be transparent and responsible because parts of the algorithm's life-cycle are outsourced to different stakeholders (Siffels et al. 2022; Cobbe et al. 2023).

In another field research, Jonk and Iren (2021) performed semi-structured interviews with key personnel and decision-makers at 8 Dutch municipal organizations to investigate the actual and intended use of algorithms (Jonk and Iren 2021).

They found that there is a lack of common terminology and algorithmic expertise not only at a technical but also at a governance and operational level.

The authors argue that municipalities would benefit from a governance framework to guide them in the use of tools, methods, and good practices to handle potential risks.

Furthermore, Fest et al. (2022) investigated how higher level ethical and legal frameworks influence everyday practices for data and algorithms used in the Dutch public sector (Fest et al. 2022). They investigated public sector data professionals at Dutch municipalities and the Netherlands Police. They found that the practicality of proposed frameworks remains a challenge for practitioners because they typically do not feel competent or miss the required skill set to make decisions regarding responsible and accountable data practices. Data professionals, as a result, get too much autonomy and discretion in handling questions that belong at the core of public sector operations.

The authors argue that efforts need to be put into the operation and systematization of legal and ethical questions across the data science project life-cycle.

Outside of the Dutch public sector, research on public algorithms has been conducted in the fields of human–computer interaction (HCI), science and technology studies (STS), and Public Administration (PA) (Holten Møller et al. 2020; Lee et al. 2019; Saxena et al. 2021). For instance, Saxena et al. developed a framework for high-stakes algorithmic decision-making in the public sector (ADMAPS) where they qualitatively coded data from in-depth ethnographic study on the daily practices of US child welfare caseworkers and prior literature (Saxena et al. 2021).

Frameworks and theories from various domains have been proposed to characterize the dynamics of interactions amongst a network of actors (Ropohl 1999; Latour 1999). Actor-Network-Theory (ANT) and mediation theory, for example, describe the relations and interactions within a network of (artificial and natural) actors (Latour 1992, 1994, 1999). Following ANT, interaction with technology is never neutral as it influences or mediates the way we carry out our tasks and decisions. On the other hand, technology is continuously mediated by our social aspects, e.g., in formulating goals and other design decisions. A related approach to describe the context of reciprocal interactions between human actors and technology is that of the *socio-technical systems (STS)* (Ropohl 1999). STS is not reserved for technology alone, it rather stresses the interactive nature of social and technical structures within an organization and society as a whole. This term is increasingly used in the field of AI to assess fairness and ethics from a broader normative context in which actors interact and operate as opposed to focusing on individual actors alone (Chopra and SIngh 2018; Dolata et al. 2022). Additionally, frameworks have been proposed to investigate networks and power structures. Following the tripartite model for ethics in technology, three main roles can often be identified through their responsibilities: the developer, who goes about the technical input, the user who goes about the use of the system, and the regulator's role, responsible of taking the "value" decisions (Poel and Royakkers 2011). Research on the use of automated systems for public decision-making has shown to shift discretionary power from the regulator roles to system analysts and software designers, often making them the main decision-makers (Bovens and Zouridis 2002). Whereas decision-makers from public organizations are often involved in the procurement and deployment phase, developers, sometimes from third parties, tend to be more involved in the development phase (Wieringa 2020). When developers become the main decision-makers for design decisions then this can exclude those stakeholders without technical knowledge in important value decisions for the system (Kalluri 2020; Danaher 2016). These imbalanced power dynamics can ultimately lead to a form of technocracy, where governance and (moral) decision-making are based on technological insights and may only yield technological "solutions" (Poel and Royakkers 2011). Lastly, in computer science research, tools such as ontologies[8] and knowledge graphs can be used to characterize communication structures and knowledge exchange between actors (Guarino et al. 2009). The AIRO (AI risks and fairness ontology) have been proposed to underpin the AI Act[9] (Golpayegani et al. 2022). Moreover, the FMO ontology structures knowledge for fairness notions, metrics, and the relations between them (Franklin et al. 2022).

We propose that what is still missing in previous work is a clear framework to characterize the communication structures (and issues) between diverse actors with different skill sets that underlie fairness-related human decisions throughout the algorithm's life-cycle. Following the above frameworks and theories, we assume a perspective of a socio-technical interactive network, in which fairness assessments for algorithms are understood as part of a governance structure where actors with different roles and tasks interact with the system. We conduct interviews to specifically focus on interactions by means of internal communication exchange and challenges that might arise between diverse actors in the Dutch public sector. The communication challenges are finally characterized in a conceptual framework. With this conceptual framework, we aim to provide a structured approach to the planning and auditing of communication processes around technical-decision making in the public sector.

3 Methodology

3.1 Semi-structured interview

We conducted 11 semi-structured interviews with each one divided into three sections and lasting an hour. We formulated the interview transcript in an open-ended manner, where participants were able to share their information in their own words whilst following a general structure of topics (Fujii 2018; Goede et al. 2019). Before conducting the interviews, participants received some example questions and a short description of the research. At the start of the interview, participants gave their consent for collaboration. The questions are added for reference in table 4 in the appendix and can be divided into three main categories:

[8] An ontology can be understood from a philosophical perspective referring to "the nature and structure of reality". From a knowledge engineering perspective, however, it refers to the modeling of a structure of a system—by organizing relevant concepts and relations (Guarino et al. 2009). They are also described as "conceptual schemas", "formal specifications of a conceptualization" and as "the abstract and simplified view of the world we wish to represent and describe in a language that is understandable by humans and/or by software agents" (Antoniou et al. 2005; Guarino et al. 2009).

[9] a regulation proposed by the European Commission to tackle various forms of negative impact caused by the misuse of AI (Commission 2021)

Table 1 Interviewees description

Interviewee	Role (self-identified)	Technical background	Use case domain
PP1	Developer & Researcher	Yes	Risk assessment
PP2	Program Manager	No	Risk assessment
PP3	Program Manager	No	Risk assessment
PP4	Product Owner	Yes	Fraud detection
PP5	Advisor & Researcher	No	Risk assessment
PP6	Researcher & Developer	Yes	Education: face expression detection
PP7	CTO Development Team	No	Risk assessment
PP8	Advisor & Data Science Researcher	Yes	Fraud detection
PP9	Developer & Researcher	Yes	Fraud detection
PP10	Advisor & Researcher	No	Fraud detection
PP11	Innovation Manager	Yes	Fraud detection

1 *General*: the topic of the use case, actors involved, and the content of the respondent's (teams) role. The goal of the system is identified, as well as the envisioned (end) users.
2 *Development process*: type of input, resources, tasks, and roles needed throughout the development process to make informed decisions. The phases of the algorithm's lifecycle are investigated by mapping out the tasks made in five phases: formulation, development, evaluation (go-no-go), deployment, and monitoring phase.
3 *Considerations*: perceived challenges for role and task division, potential improvements or failures of the system, and communication gaps. Questions about assessing errors, bias, as well as the potential negative impact of the model.

The first two sections ask interviewees to describe general procedures and practices in the system's life cycle. In these sections, interviewees got the opportunity to mention internal communication and key elements of the communication processes that underlie fairness-related human decisions. We specifically asked about communication issues in the third section of the interview transcripts. We preliminary tested all interview questions with a pilot on five colleague researchers from different disciplines regarding their use case projects. The questions were deemed suitable for letting interviewees describe their process of communication and related issues. The suitability of the questions was checked in terms of comprehensibility and relevance to the research questions. No questions were altered afterward.

3.2 Case studies

We recruited participants who have been collaborating in multi-stakeholder projects in the Dutch public sector. Participants working in the domains of fraud detection and risk assessment were specifically targeted.

We aimed to investigate participants that are or were involved in the choices and practices applied at each phase of the algorithm's life cycle. To this effect, we used a repository of use cases[10] from the Dutch Ministry of interior affairs[11] and the snowball sampling technique. These use cases were of particular interest because the outcomes can directly affect citizens.[12]

For this research, (*N = 11)* interviews were conducted with (*N=10*) actors involved in Dutch social domain use cases, and *(N = 1)* actor in the Dutch educational domain. Table 1 describes the participants (PP), the use case, and the (self-identified) role they filled at the time of involvement. Also, it is indicated if they mentioned to have a technical background. With a technical background, we refer to those who are (not) educated or have (no) experience in technical science. Topics discussed for the use cases were predicting fraud amongst citizens (e.g. when applying for public service) *(N = 5)* or predicting the need for social benefits amongst citizens in the coming years *(N = 5)*. The difference between risk and fraud cases is not always clear-cut. A fraud detection model can also be used for risk assessment and vice-versa. The categorization of the use cases is

[10] Some examples of public domain use cases in the Netherlands can also be found via the Artificial Intelligence Netherlands Coalition (NL AI Coalitie) website https://nlaic.com/use-cases and in (Van Veenstra et al. 2019; Hoekstra et al. 2021).

[11] Dutch Ministry of Interior Affairs and Kingdom Relations https://www.rijksoverheid.nl/ministeries/ministerie-van-binnenlandse-zaken-en-koninkrijksrelaties.

[12] Moreover, in 2019 inspection and enforcement use cases were identified as the biggest categories for AI usage in the Dutch social domain which refers to the prediction of (security) risks by identifying patterns of behavior (Van Veenstra et al. 2019).

Fig. 1 Example of interview Q &A (left) and the corresponding coding labels (right) from the qualitative coding analysis

therefore based on the punitive (fraud detection) or assistive (risk assessment) nature of the policy intervention (Rodolfa et al. 2020).

3.3 Qualitative coding analysis

To answer RQ1 *Which actors, roles, tasks, and phases can be identified between multi-stakeholder interactions throughout the algorithm's life-cycle?*, we performed a qualitative coding analysis by labeling key *codes* from the interview output[13]. The results can be found in Sect. 4.1. We used in vivo,[14] descriptive[15] and process coding[16] to identify the process of communication exchange between diverse actors, as well as the practices and choices made at each stage of the algorithm's life-cycle.

The coding analysis was done in multiple cycles. In each round of coding, pieces of text are merged or split into categories. This was repeated until all codes could be placed under bigger categories. Two of the authors perform a separate coding analysis to reduce the impact of personal bias. We performed coding analysis by hand and using a coding analysis tool.[17] Beforehand, both analysts agreed that particular attention should be drawn to identifying the roles, tasks, phases, and challenges from the interview transcript. For example, if a participant were to mention that "person X is a developer and performs bias analysis in the development phase" the actor, the role, the task, and the phase would be labeled. We also paid attention to the direction of communication between actors (eg. person X hands over model outcomes to person Y). Furthermore, communication challenges, decisions, and information that were repeated or pointed out by participants as essential for fairness-related decisions were coded.

In Fig. 1, an example is given for the interview output (left) and the corresponding codes (right). The figure shows the colored group labels for *challenge, roles, task*, and *phase*. On the right, an example of the corresponding descriptive codes can be found. For example, "it's hard to get a focused answer" was summarized as an information exchange challenge of the type where "more input is needed". We added the corresponding role(s) to the codes in brackets "[]". If the code concerned multiple roles, we added "-" to indicate a relation and arrows if there was a direction of information exchange. In this example, more input is needed from the end-user to the developer.

For the sake of supporting a common terminology, we took a broad range for the formulation of the codes. For example, participants that mention roles for engineers, coders, and data scientists are all considered as the role of developers. Also, for the phases, we considered the testing phase, modeling, and experimentation phase as the development phase. These are grouped because they were not purposely

[13] Saldaña (2013) describes that "A code in qualitative inquiry is most often a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data." In addition, a code can be understood as a researcher-generated construct that symbolizes the construct and assigns an interpreted meaning.

[14] In Vivo coding is also named "literal coding" and refers to a word or short phrase from the actual language found in the qualitative data record, e.g. terms used by participants themselves (Strauss 1987; Saldaña 2013).

[15] Descriptive coding refers to summarizing the basic topic of a passage of qualitative coding in a word (noun) or short phrase (Saldaña 2013).

[16] Process coding refers to "action coding" which implies action from more simple observable activity (eg. reading) to more general conceptual action (such as adapting) (Saldaña 2013).

[17] Atlas.ti: The Qualitative Data Analysis & Research Software https://atlasti.com/

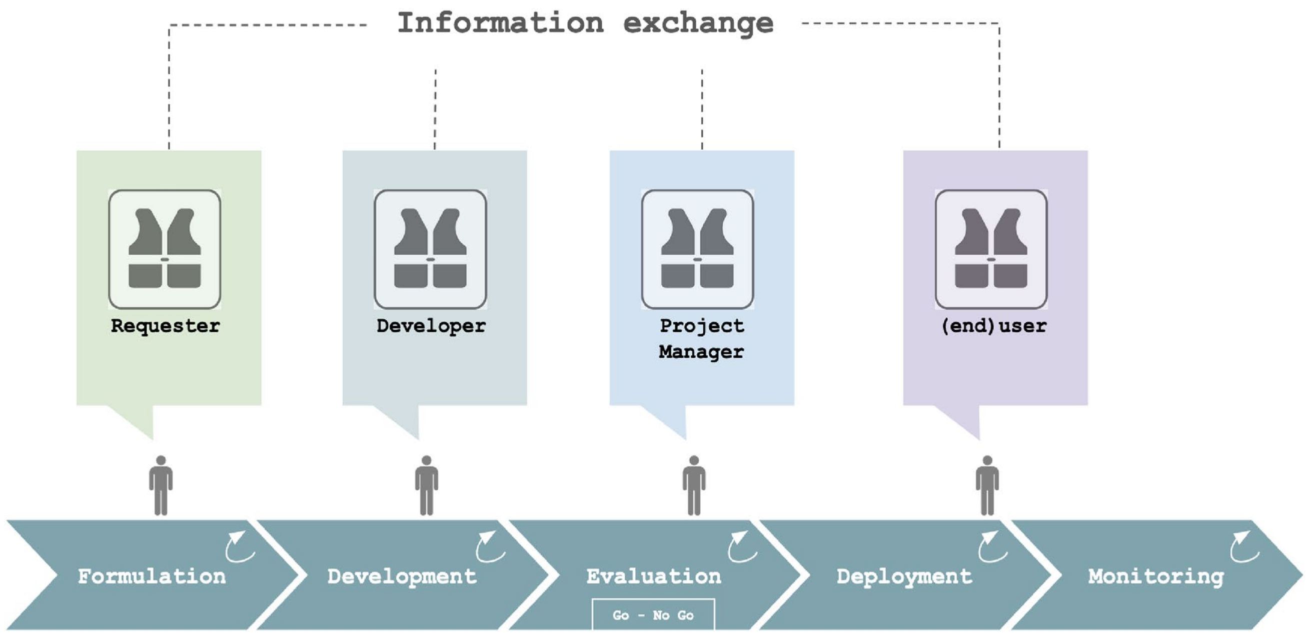

Fig. 2 Key *phases* of a system's life-cycle, and their *actors* and *roles*, identified from the interview transcripts

separated but used interchangeably by participants in the interviews.

In Fig. 2, an overview can be found of examples for the main identified phases, actors, and roles from the interview transcripts. Moreover, it can be seen that we focus on the (missing) information exchange between roles throughout the phases. The arrows in the phases indicate the iterative nature of a system's life-cycle. We compared both coding analyses to identify discrepancies or alignments. Finally, we performed a co-occurrence analysis on the identified codes by counting which roles occurred the most per phase, task, and challenge throughout the interviews. We purposely separated actors from roles to analyze how many roles actors can take on for a certain task.

3.4 Constructing a conceptual model

The co-occurrence analysis alone did not capture the relation between codes, i.e. "end-user is missing in the development phase" would still count as a co-occurrence although the role was missing. We therefore built a model to contextualize the identified codes by describing the relations and characteristics. The model can be used to represent the general communication processes between diverse actors that underlie fairness-related decisions. By constructing a conceptual model, we answer RQ2: *How is information about fairness assessments communicated between diverse actors involved throughout the algorithm's life cycle?* Subsequently, we applied the model to the challenges identified in the qualitative coding analysis and answer RQ3: *Which communication challenges can be identified in fairness assessment processes?*

3.4.1 Creating concepts from codes

The identified codes from the interview output formed the concepts for the communication model. The construction of the model resembled that of an ontology by building it in an iterative manner (Noy and McGuinness 2001; van Hage et al. 2011; Golpayegani et al. 2022; Franklin et al. 2022). This means that we continuously adjusted the model until it would represent every piece of code identified from the qualitative coding analysis. We also presume that the model can be revised and extended for future work when more input is available.

3.4.2 Defining descriptions, properties, and relations between concepts

The internal structure and relation between the concepts are characterized by adding definitions, characteristics, and properties. By doing this step, we described the concepts in relation to each other for the generic process of communication.

We added descriptions to each concept to agree on common definitions. Descriptions are based on answers we got from the interview output, the definitions we found from documents provided by the European Commission on Trustworthy AI, and from other sources in the literature (Commission et al. 2020; European Commission and Technology

2019; van Hage et al. 2011; Golpayegani et al. 2022; Tamburri et al. 2020; Poel and Royakkers 2011). The properties “affiliation” and “is missing” were added as properties to the model. By describing the type of private or public affiliation (national institute, ministry, or municipality) we can contextualize how tasks and roles are divided within multi-stakeholder collaborations. Moreover, relations are added between codes. For example, an actor always “has” a certain role whereas a task “requires” a role “during” a phase. This step resulted in the Phase, Actor, Role, Task, Skills (PARTS) model and can be found in Results Sect. 4.2.

3.4.3 Applying the model to identify patterns for communication challenges

Lastly, we characterized the challenges to answer RQ3: *Which communication challenges can be identified in fairness assessment processes?*. To this effect, we added two extra concepts, one for skills and one for information exchange. By adding these, we were able to distinguish challenges that occurred because either an actor or role with the right skills was missing for a certain task or phase or if (more) information was needed between actors to do a certain task.

4 Results

4.1 Qualitative coding analysis

4.1.1 Semi-structured interviews

In all use cases multiple stakeholders were involved with varying technical expertise—from social workers to developers, researchers, program managers, and advisors from third parties. 6 interviewees mentioned having a technical background themselves, but not necessarily in AI or software development. For the majority of use cases *(N = 10)*, the procurement for the algorithm came from government organizations and municipalities. Furthermore, the envisioned end-users of the systems were in most cases policy-makers or other social workers at municipalities with minimal or no technical expertise. There was one exception for the educational domain, where teachers were the envisioned end-users. Figure 3 provides an overview of the main group codes under which all codes are categorized. These group codes are ‘Phase’, ‘Role’, ‘Task’, ‘Skill’, and ‘Actor’. For instance, it can be seen that we identified the code “Define use case” under the group code “Task” which happens during the code formulation under the group code of “Phase”.

4.1.2 Qualifying the stakeholders

Figures 4 and 5 illustrate which roles were mentioned the most based on co-occurrence per Phase and Task. Overall, the developer role was mentioned the most *(N = 189)*, followed by End-users *(N = 107)* and Policy Makers *(N= 92)*. More formulation of roles, phases, and tasks were identified but mentioned less and thus not shown here.

Figure 4 shows that developers are most prominent in the development, evaluation, and formulation phases and less in the deployment and monitoring phases. PP1, who filled a developer role reported *“we don’t monitor what the municipalities are doing with the results.”* and *“feedback is needed on how the results will be used in deployment”*.

End-users and policy-makers were reportedly the same in our use cases. This means that public partners working at municipalities or other public organizations were in almost all cases also the envisioned client and those that would (in) directly engage with the system. Both roles were also the second highest in occurrences for phases. Moreover, Fig. 4 demonstrates that the monitoring phase *N = 10* was mentioned the least throughout the interviews whereas the evaluation phase was mentioned the most *N = 90*.

The data subject role occurred the least for phase and task and was in almost all cases filled by citizens. On involving citizens, a researcher (PP10) working on fraud detection, reported that *“it depends on the type of AI. If it has an impact on citizens or uses a lot of data from citizens, it would be relevant to include a focus group of citizens from the beginning but it is less relevant for road repairs.”* Other roles are also mentioned that need to be more involved throughout the phases. Developer (PP9), reported that *“For the future, we could incorporate stakeholders at earlier stages in the development to see what the potential sources of bias are.”*

Additionally, requester roles were rarely mentioned as being involved throughout the phases. They were only mentioned in the formulation phase for funding or initiating the project. In our use cases the request for the model often came from ministries. The data subject role as well as the requester role were never described as the end-users.

In figure 5 it is shown that the developer role was mentioned most for tasks, e.g. technical decision-making, consulting, dealing with fairness and risks, and the least for model usage. This indicates that actors taking on developer roles were the most prominent in making decisions throughout the algorithm’s life-cycle. On what type of decisions one makes, PP9 (developer) said that they *“decided on how to improve accuracy and handling issues. For instance, gathering more data diverse to handle bias.”*. Developer PP1 mentioned that they *“define and chose metrics for the models.”* and that these*"are defined in collaboration with the*

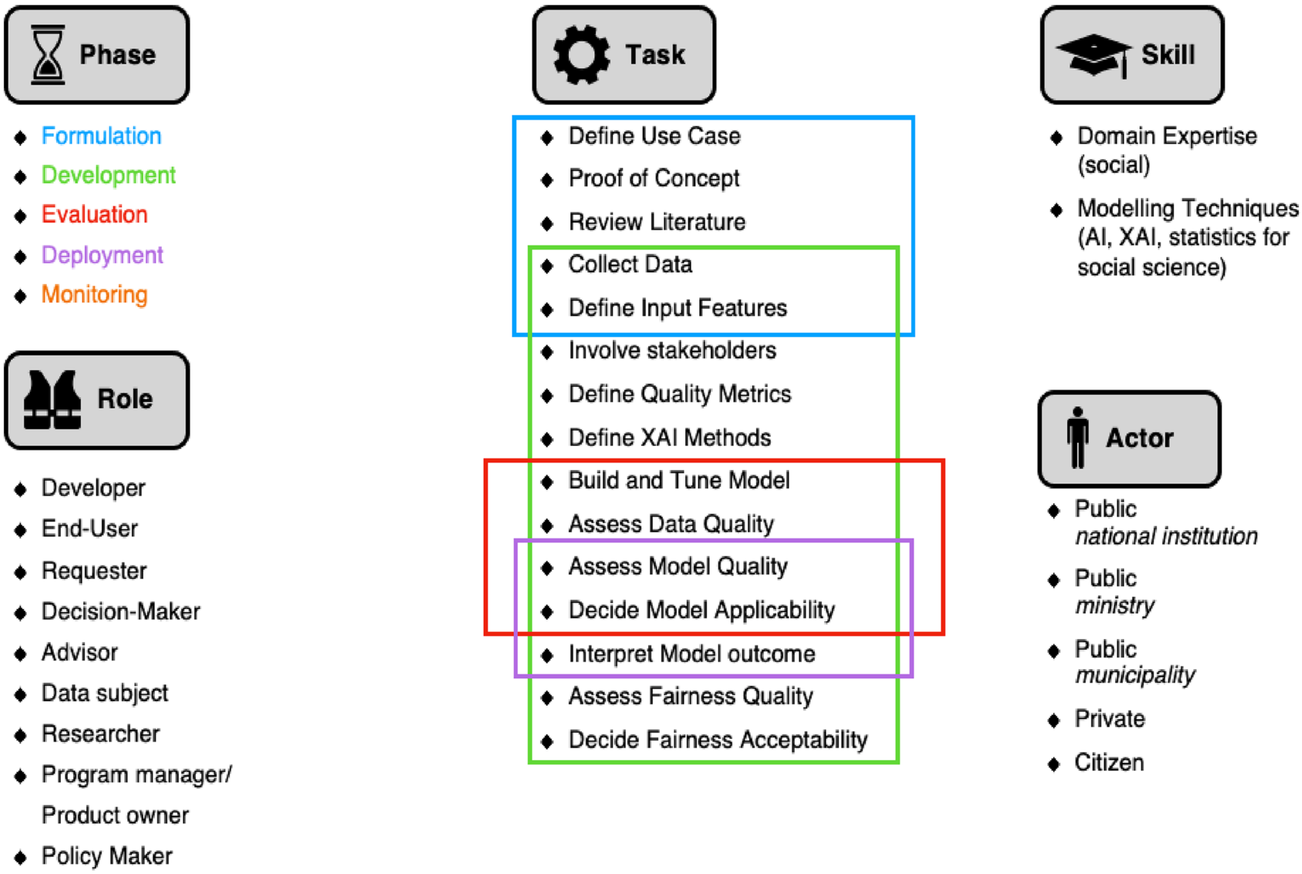

Fig. 3 Codes categorized by main group codes identified in interviews. As the colored boxes show, *Tasks* can occur during different *Phases* and these can overlap. Instances of *Actors* are not generic, they represent public affiliations (national, ministry of the municipality), private or individual citizens

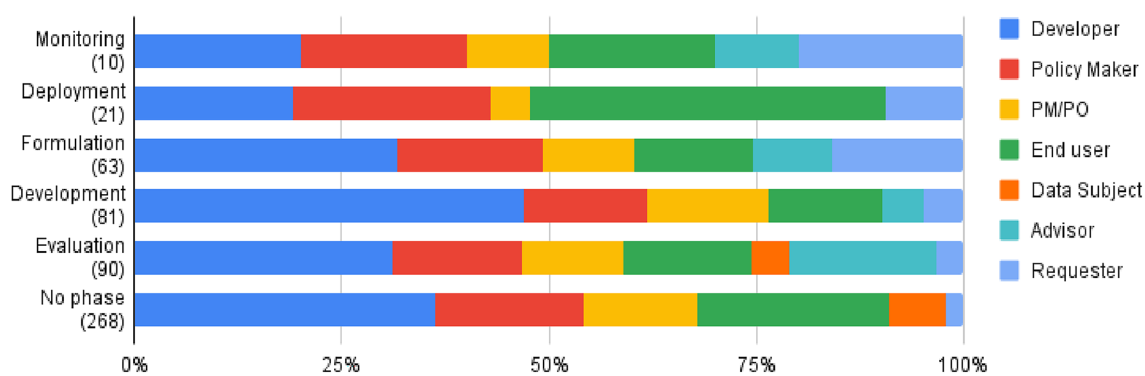

Fig. 4 Mentions of roles (relative) per phases. The total amount of occurrences per phase is mentioned on the *y*-axis (from least to most). Note that the developer role (blue) is mentioned most in all phases, followed by end-users (green) and policy-makers (red) and data subjects (orange) least

municipality but choosing metrics and trimming down after input was decided by the two of their team."

Program managers and product owners (PM/PO) often work in the same research teams as developers and are either hired externally or internally by a public requester. They are the main decision-makers for involving (external) stakeholders such as (external) advisors, requesters, and end-users, and sometimes are the only ones in direct contact with them to answer questions. PM/PO's were often reported to supervise developers in technical decision-making but rely on

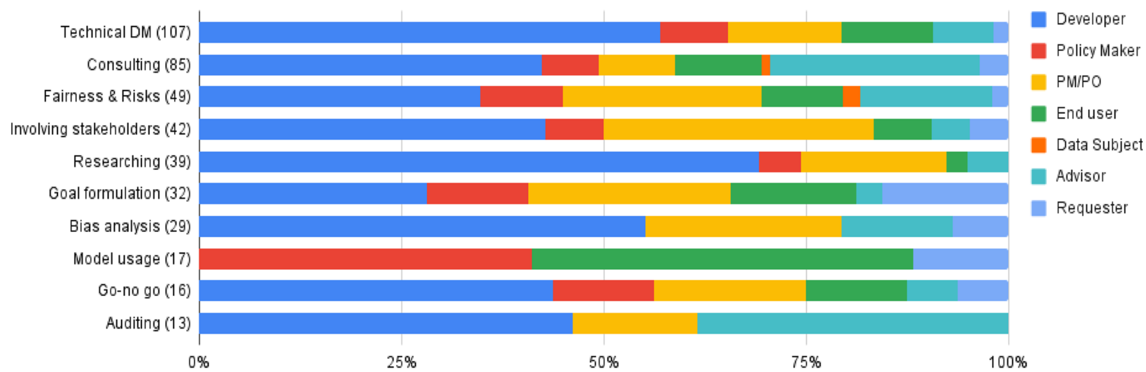

Fig. 5 Mentions of roles (relative) per task. The total amount of occurrences per task is mentioned on the *y*-axis (from most to least). Note that technical decision-making is mentioned the most. Developer roles (blue) are mentioned most for all tasks except for model usage

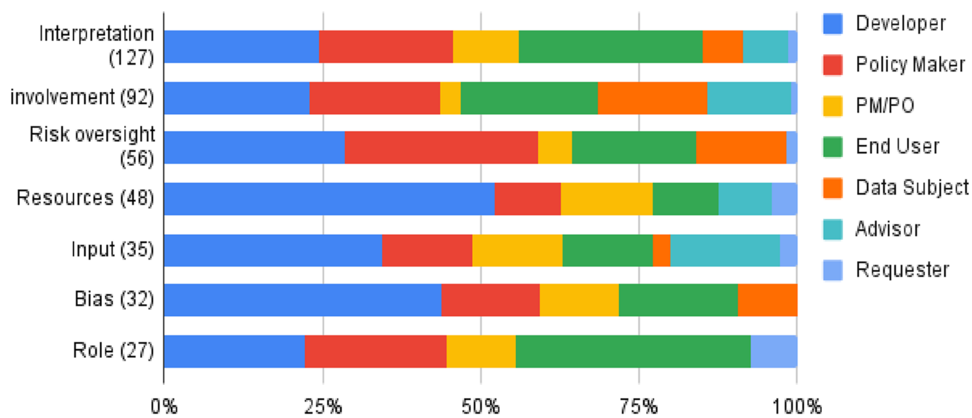

Fig. 6 Mentions of roles (relative) per challenge. The total amount of occurrences per challenge is mentioned on *the y*-axis (from most to least). Note that interpretation and involvement issues were mentioned most. Moreover, developer (blue) is mentioned most followed by end-users (green) and policymakers (red)

their judgment for bias, fairness, and risk factor analysis. Program manager (PP2) said that for handling error rates and biases they *"rely on the technical teams' judgment"*. Also they mentioned that *"the technical colleagues give advice when the model is good enough, but it's a bit of a grey area. We also rely on literature"*. PP3 mentioned that they *"don't go to users"* for error or bias analysis but *"talk to the data scientist"*. Product owner (PP4) said that *"it is time intensive to explain [bias analysis] to stakeholder users. Bias analysis is sometimes so complex, even as an expert I sometimes don't understand it, and it takes a lot of time"*.

Advisors were often mentioned to consult on the system in the evaluation phase before deployment, or when the project is halted. Advisors were presented as giving advice on (1) domain knowledge, (2) technical knowledge, or (3) ethical knowledge. Actors filling developer roles also sometimes filled advisor roles. A third-party developer could be hired to analyze the code, give technical advice or even build the model. PP3 added that they *"hired an external bureau for auditing and investigating the algorithm"*. Also because they *"could not get reliable predictions because the social domain changes all the time, and it's hard to keep track of these changes—for example in social support—and how that impacts the system"*. PP10 mentioned that *"an external company was hired to develop the model for the municipality"*, which made the *"data ecosystem quite complex"*. A social domain expert and researcher at the municipality (PP5) mentioned that they *"were involved to give feedback as an involved bystander. But it was hard for someone like me to understand what the difference between implementation and design is and what that means for real-life implications"*.

In Fig. 6, the relative occurrences for roles per challenge are demonstrated. Most communication challenges were reported amongst end-users, policymakers and developers due to lack of interpretation, missing involvement, and (risk) oversight in a certain phase. Various interviewees reported that more input is needed on the interpretation and use of the results from end-users in the deployment phase. Vice versa it was reported that input is needed on feature selection and bias analysis in the development phase from developers to end-users. Program manager (PP3) mentioned that it is challenging that *"we don't know if governments and municipalities can understand the model"*. Also, PP1 mentioned that *"it's hard to get a focused answer on how they are going to use the model and what the results will be."* On the matter of involvement, (PP2) reported that *"More frequent and streamlined collaboration with the municipality is needed"* and that they *"would like to have closer contact with the municipality."*. PP1 confirmed that in their case *"the municipality is too loosely involved in the project,* and that *more involvement is needed"*.

End-users and policy-makers are often reported to miss the technical skills to understand the uncertainty of predictions and limitations of the model in real-world settings. For example, PP10 mentioned that the most difficult challenge is the *"gap between data scientists and policymakers. How to make sure that what is developed is being well understood and useful for those of non-tech background."*. PP1 also confirmed that the *"Most important risk is that the model will not be used or is misinterpreted. For example, mixing up correlation and causality might lead to not helping people at risk of poverty."*. PP6 mentioned that *"People could trust the model blindly and mistake it for a decision-making tool"*. and lastly, PP8: *"Not sure if the inspectors fully understood why certain cases were flagged as misuses or put on the list."*

With regard to biased decisions, another participant (PP9) mentioned that *"there should be more focus on asking users what policymakers perceive as risks and biases"* and that it is *"difficult for them to understand that there are many different interpretations. What it really means to be a 'true positive', is this person really a fraud, or was this person not able to fill in the forms properly?"*. Misinterpretation and misuse of the system by end-users and policy-makers were mentioned most in the monitoring and deployment phases. PP6 confirms that *"training for users is needed, to remind users not to rely on the tool but that the decision is up to them."*. This also explains why policy-makers and end-users are mentioned a lot throughout all phases because of their co-occurrence with challenges in certain phases.

Lastly, participants reported the need for more citizen involvement and being more transparent to citizens throughout the phases algorithms' life-cycle. PP2, a program manager, mentioned that *"the plan is to check, e.g. with research labs how citizens can give feedback on the model. But currently, nothing is envisaged"*. Program managers and product owners are also hesitant to involve citizens for a lack of appropriate frameworks, and potentially hostile attitudes. "*There is a long history with the citizen council for consultation and it is usually conflict-based. It's hard to make fruitful collaboration, getting them to understand the issues and getting them out of anger mode.'* said a product owner (PP4) working at the municipality. PP7, on previous involvement of a citizen council:*"they said no on the feasibility of the model from the municipality. They did not get it. It was more of a general no to technology instead of asking a targeted question"*.

4.2 Building PARTS

From the results of the qualitative coding analysis specific patterns of communication issues emerged. The codes that were identified are critical to characterize the communication processes that underlie fairness-related human decisions. However, these communication processes are not easily described with qualitative coding analysis only. By solely counting the codes for (co-)occurrences, we could not capture the type of relation between codes that are important for determining the process of communication. For instance, a communication challenge could be that a role is missing in a certain phase, which would also count as a co-occurrence for a role within a phase. The communication patterns are also not easy to document in written form only. Therefore, we documented the communications patterns we identified in a conceptual model.

In the next section, we provide a more structural analysis of the communication patterns using the Phase-Actor-Roles-Tasks-Skills (PARTS) model. With PARTS the process of communication is characterized which is needed to understand how fairness assessments is communicated between diverse actors involved throughout the algorithm's life cycle. PARTS was specifically designed for—but not restricted to—modeling the results of our qualitative coding analysis.

4.2.1 Qualifying the communication process

The PARTS model structures the communication frameworks and the basic concepts for the communication issues identified in our qualitative coding analysis. The six concepts selected from the group codes from the qualitative analysis are provided in Fig. 7 and their corresponding descriptions, relations and properties are provided in Table 2.

It comprises generic concepts, such as *Phase* and *Actor*. We added an extra concept for *Information Exchange* to characterize the communication needs between actors in certain phases to complete a certain task. The model requires at least two generic concepts to materialize the communication process: the stakeholders who exchange information (represented with the concept *Actor*), and the act of communicating (represented with the concept *Information Exchange*). The relevant context of the communications can be modeled with four additional concepts (*Phase, Task, Role, Skill*) that underlie fairness related human decision-making. We decided to add *Skill* as a specific concept because the outcome of the qualitative coding analysis showed the importance of identifying and differentiating individual stakeholders by their particular *roles* and specific *skills* within the process.

With the properties *during* and *involves*, we can model *Tasks, Information Exchanges* and *Actors* to the *Phases* of

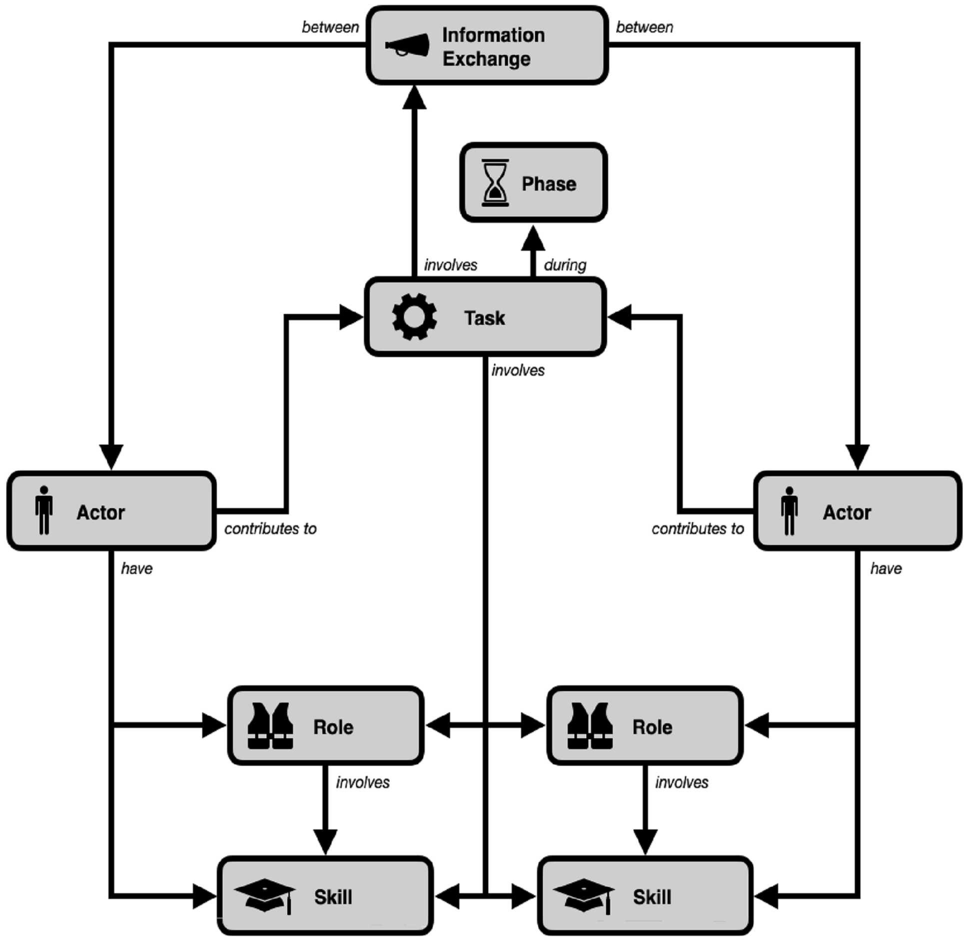

Fig. 7 Basic concepts of the PARTS model for characterizing the structure of communication frameworks, and the issues identified. For example, fairness assessment *Tasks* may be missing at certain *Phases* of a system's life-cycle. Or *Actors* may not have the right *Skills* or *Role* when performing fairness assessments

Table 2 Concept descriptions

Concept	Description	Relation	Property
Task	Actions carried out by Role	Requires Role, Skill, Info exchange, during Phase	*Is missing*
Role	Function or a part filled by Actor	Actor Has Role, Task requires Role	*Is missing*
Actor	Entities that perform tasks, actively or passively	Has role, skill, phase involves actor, info exchange between actor	*Affiliation*: public (national institution, ministry, municipality) or private, *is missing*
Phase	Indicates the evolution of the system from conception through retirement	Involves actor, task during phase, info exchange during phase	*Is missing*
Skill	Professional ability, expertise, or knowledge needed in practice to successfully complete a specific task	Info exchange requires skill, task requires skill, actor has skill, role requires skill	*Is missing*
Info exchange	Communication transfer between actors	requires skill, between actor, during phase, task requires info exchange	*Is missing*

the system's life-cycle. This provides a generic temporal overview of the communication and assessment processes, e.g., to reflect on the assessment *Tasks* or *Actors* (e.g., stakeholders from specific institutions) that may be missing at specific *Phases*.

Our main finding is that adding the property *is_missing* to any of the six concepts in the communication model is of great interest for expressing the issues that interviewees mentioned. We represented the elements of communication frameworks using these six concepts precisely because issues arise if any of these types of elements are missing. For example, an *Actor* may be missing specific *Skills*, or a fairness assessment *Task* may be entirely missing. Communication issues also arise from missing a *Information Exchange*.

Table 3 Main roles description

Role types	Description
Developer	Research, design, and/or develop algorithms
Policy-maker	Responsible for designing and overseeing the carrying out of policy and social decisions
Project manager/product owners (PM/PO)	Supervise the projects for the development of the system and oversee documentation checks and balances
End-user	(In)directly engage with the system and use algorithms within their business processes to offer products and services to others
Data subject	Organization or entity that is impacted by the system, service, or product
Advisor	Give constructive feedback on the system throughout the life-cycle
Requester	Who are the main client and investor for the use case

The added properties and relations between concepts were sufficient to represent the communication challenges we observed in the interviews. The property *requires* assigned to the fairness assessment *Tasks* allows representation of the *Roles, Skills* and *Information Exchange* needed to perform the task. This part of the PARTS model is essential to define the high-level requirements for accountability and transparency. The practical implementation of these requirements can be modeled by associating *Actors* with specific *Skills* to the *Roles* they take in fairness assessment *Tasks*.

In Table 3, descriptions are given for the main group codes for roles. The concepts we selected, and their descriptions, are in correspondence with those provided by the European Commission on trustworthy AI, and other sources in the literature (Golpayegani et al. 2022; van Hage et al. 2011; Commission et al. 2020; European Commission and Technology 2019; Poel and Royakkers 2011; Tamburri et al. 2020). For example, the developers and end-users (as instances of *Actor*) are, respectively, those who "research, design, and/or develop algorithms" and those who "(in)directly engage with the system and use algorithms within their business processes to offer products and services to others" (European Commission and Technology 2019). Data subjects (also instances of *Actor*) are "an organization or entity that is impacted by the system, service or product" (Golpayegani et al. 2022).

We found that these concepts were sufficient for modeling generic communication challenges that we observed in the interviews. The model can be extended with additional concepts, e.g., to represent the affiliation of *Actors*. Further research is needed to further refine the model. Adding more concepts at this stage comes at the risk of making it less applicable and harder to generalize to new contexts where different local, organizational values and norms may be at play.

4.3 Applying the PARTS model: characterizing the communication challenges

Finally, we apply the PARTS model to characterize the three most prominent findings for challenges. From the qualitative coding analysis it was identified that interpretation and involvement issues were mentioned most between policy-maker, end-user and developer roles. Also, data subjects were reported to be missing throughout all the phases of the life cycle. In this section, we provide an example for each of these challenges through visualizations from the PARTS model.

4.3.1 Challenge Finding 1: end-users and policy-makers often lack the technical skills to interpret the system's output or to estimate potential fairness issues

From the interview, we identified that more input is needed between developer, end-user and policy-making roles. End-users need input on the explanation of the model for its limitations and performance from developers in the deployment phase to be able to interpret the model outcomes. Developers need more input from end-users on how the results of the model are used in deployment. In Fig. 8, we provide an example of this where it can be seen that the task of "interpreting the model results" during the deployment phase requires the skill of "understanding the model performance". However, this skill seems to be missing for the actor at the municipality that has the policy-maker and end-user role. Moreover, more input is needed between the developer roles and the policy-maker and end-user roles. Here it can be seen that the information exchange for "discussing the model limitations & performance" is missing, which challenges the task of "interpreting the model results".

4.3.2 Challenge Finding 2: developers take on extra roles such as advisor and decision-maker while potentially lacking the business logic and domain expertise required

Fairness assessments are also impacted due to involvement issues. For example, when individual actors with specific roles and skills are not involved—then other roles without the right skills might make unjust decisions. In Fig. 9,

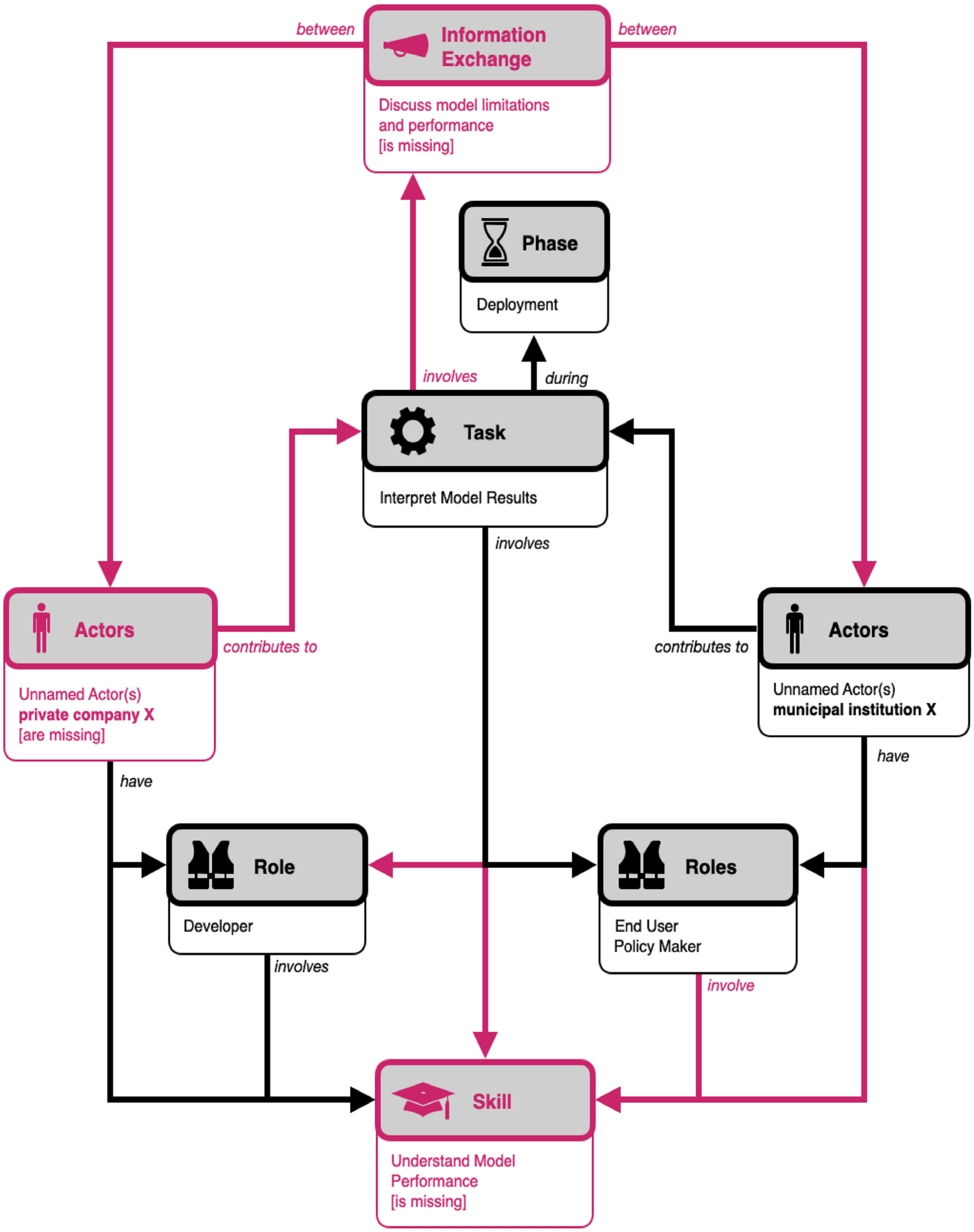

Fig. 8 The PARTS model applied to challenge finding 1: end users interpretation issues of the model outcome. Note that when the information exchange between roles is missing and the technical skills are missing the task of interpreting model outcomes becomes challenging

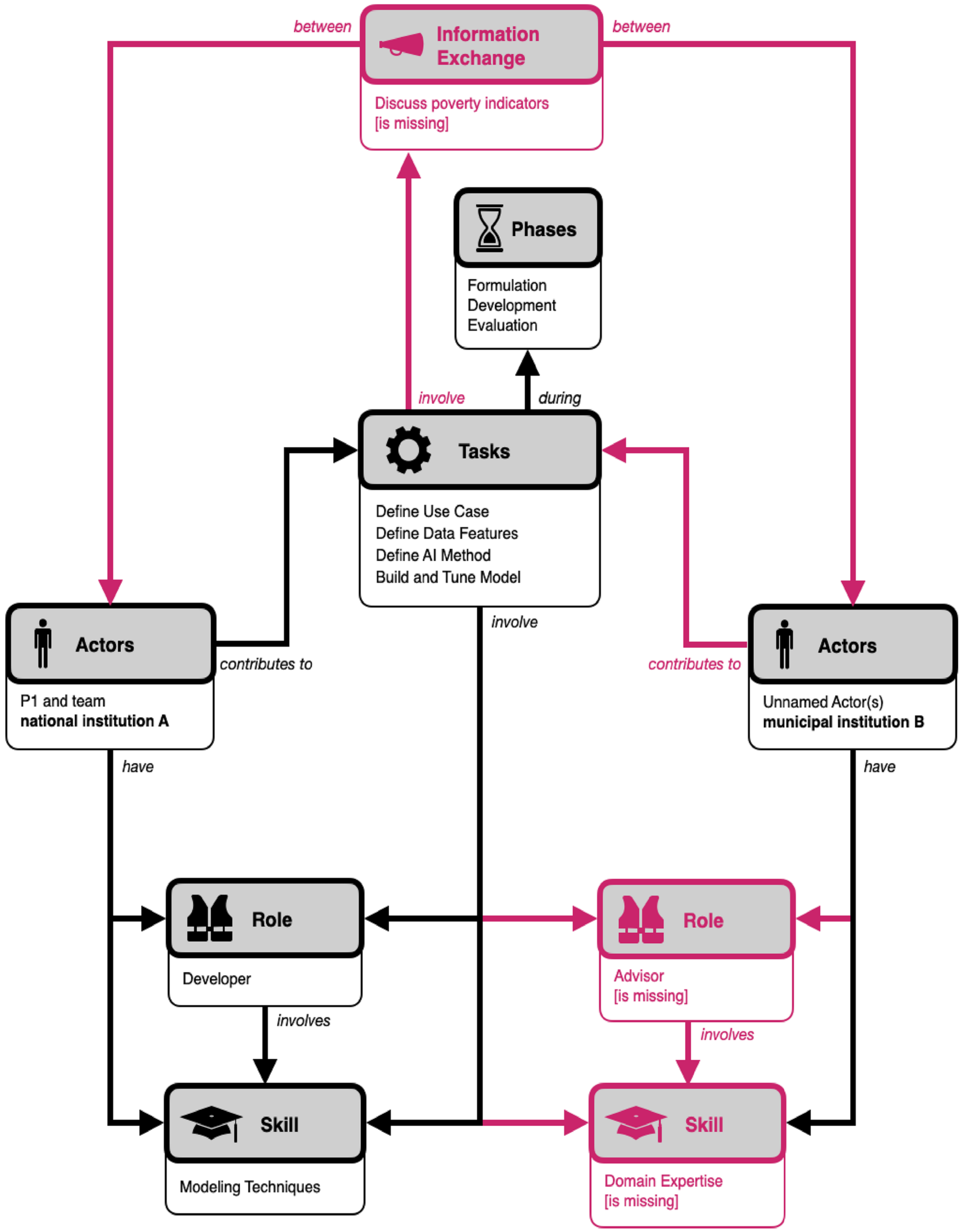

Fig. 9 The PARTS model applied to challenge finding 2: Developers take on extra roles without domain expertise due to missing involvement from domain experts in design decisions

we use the PARTS model to illustrate that an actor from the municipality with the right domain expertise is missing during formulation, development and evaluation phase. When the advisor and domain expertise roles are missing then there is no information exchange regarding the discussion for poverty indicators for a poverty prediction model. Here a consequence might be that actors with technical skills becomes the main decision-maker but may miss the required skills in the domain of application. Examples of domain skills are, e.g. expertise in demographics, finance,

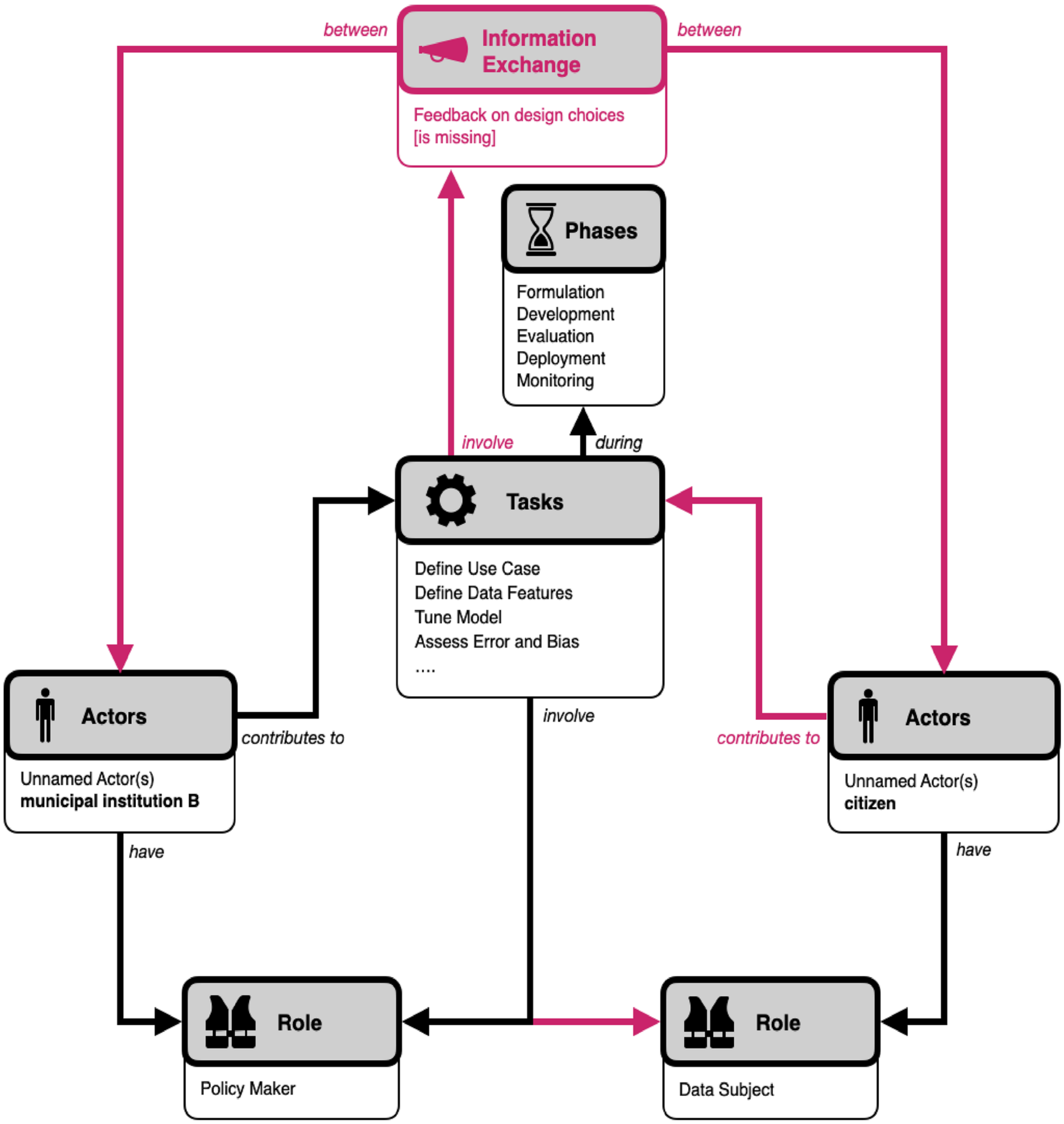

Fig. 10 The PARTS model applied to challenge finding 3: citizens feedback is structurally missing throughout the algorithm's life cycle

and forensics to ensure that their technical decisions are fair or allow them to assess fairness outcomes appropriately.

4.3.3 Challenge Finding 3: citizens are structurally absent throughout the algorithmic life-cycle

The third challenge indicated a lack of involvement from data subjects which may lead to unbalanced fairness assessments that do not include key input from relevant stakeholders. Interviewees reported that the right frameworks to collaborate with citizens were missing. In Fig. 10, it is shown that a citizen is missing that has the role data subject. Because the citizen is missing there is also no information exchange between municipalities and citizens regarding feedback on the model.

5 Discussion

The patterns identified from the coding analysis indicate a heavy reliance on the (technical) skills of the developer role in the development, formulation, and evaluation phases. These results are based on the co-occurrence analysis performed on the interview outputs. Developers are often the main decision-makers for most tasks throughout the life cycle, including but not limited to bias analysis, analysis of fairness and risk factors as well as data selection and model training. These tasks together can influence the extent to which algorithms and their outputs are free from bias, discrimination, and other forms of unfairness towards different groups, individuals, or communities. These tasks are therefore an important part for the assessment of fairness.

Moreover, the coding patterns demonstrated a growing information gap between technical experts on one side and domain experts, policy-makers, and end-users on the other side. There is a need for increased input between end-users, policy-makers, and developers at the right phase, in particular during design choices and output interpretation. Even when there is involvement, participants indicate that the necessary skills can be missing to provide informed decisions. Developers have limited supervision, advice, and lack the necessary feedback loops. For end-users there is uncertainty about whether the algorithm actually delivers what it is supposed to perform, coupled with weak support and guidance. Presenting algorithms and features in an understandable way to end-users and policy-makers is increasingly needed to make fairness assessments for potential risks, failures, and errors of the system.

These findings show that in many use cases there is no clear governance regarding roles, actors, information flows, and responsibilities during the algorithm's life-cycle. It was not always clear who makes final (mostly policy) decisions on the further development or use of algorithms, or what is the (legal, procedural, information) basis for such decisions. This seems to be due to the lack of actors filling the right roles at the right phase. As a consequence, actors will take on multiple roles at once for which they may not be fully equipped, which can lead to a discretionary imbalance. We saw that mainly developers tend to have multiple roles such as advisor, researcher, and (main) decision-maker, without being fully aware of the business logic and domain expertise. In addition, policy-maker or social worker roles tend to coincide with end-user roles without having the technical skill or understandable assessment tools set to interpret model outcomes and to make informed value decisions. We want to emphasize that it is not always a choice made by individual roles to not communicate or involve other roles, but rather a consequence of an institutional setup and a lack of established internal processes.

It is possible that interviewees forgot to mention involved stakeholders in the development process or didn't see some as influential for choices, practices, and protocols. Forgetting a particular role or actor does not necessarily reflect the actual governance structure or experienced communication challenges. Participants may not have oversight, be unwilling to provide specific details, or were perhaps steered by how the interview questions were formulated.

We also recognize that in practice, formulations for roles and groups can vary and can be diffused. For instance, in some cases, actors identified as developers would primarily identify themselves as a researcher who would also carry out "developer tasks". Moreover, counting (co-)occurrences is not enough to assess the structure of communication process between multi-stakeholder collaborations. As mentioned in the results, sometimes an occurrence would be counted for role and phase when actually 'role X was missing in phase Y' and thus was more an occurrence for a challenge than for a phase. Therefore, the PARTS model was needed to characterize the relations between concepts. Regardless of the stated limitations, our findings confirm and complement earlier work where the growing autonomy and discretion of developers in public sector operations is emphasized as well as unclear role divisions for automated decision tool usage (Fest et al. 2022; Kalluri 2020; Bovens and Zouridis 2002; Poel and Royakkers 2011).

The questions in our semi-structured interview were not limited to questions regarding fairness directly. We intentionally wanted to ask about all choices and practices that may relate to fairness-related decisions, including key elements of the communication processes. However, we do not claim that fairness or other communication issues are resolved with PARTS as the model is based on a specific and limited *(N = 11)* target group of participants involved in Dutch multi-stakeholder collaborations for fraud and risk assessment use cases. A limitation of our findings is that we did not talk to citizens. We chose to describe and investigate only those stakeholders that were involved in the choices and practices throughout the algorithm's lifecycle. Through the snowball sampling technique, no referrals were made to citizens. Future work should focus more on the co-design and participatory approaches with citizens regarding algorithms in the public sector.

The PARTS model is purposefully kept as simple as possible as every (public sector) use case will have its own (normative) context, specific governance, and communication structure. However, in this research we tried to investigate exactly those local conditions—the communication process underlying the choices and criteria for what is considered to be (un)fair—and resist "the portability trap" of stating that every AI use case will function the same from one context to another (Selbst et al. 2019; Barocas et al. 2021). More research is needed to investigate how local governance structures and communication processes influence fairness assessments outside of the Dutch context.

With PARTS we show that if these governance structures—such as roles and responsibilities—are unclear, undefined, and if there is a lack of communication or miscommunication, then this can lead to misinterpretation of the system output and even to misuse of the algorithmic system. With the PARTS framework we aimed to make a start at structuring and outlining these information gaps for missing elements such as missing roles, skills, and involvement. Furthermore, it promotes a common terminology beyond solely the technical perspective of fairness assessments by highlighting the governance of algorithms throughout the lifecycle.

6 Conclusion

This paper examines the roles of different stakeholders engaged in the process that leads to algorithm procurement and development in the Dutch public sector. In particular, we focus on the analysis of communication patterns exhibited in fairness assessments. The field research we conducted points to inadequate or missing explicit guidance or rules regarding roles, skills, and information exchange throughout the algorithm's life-cycle which leads to misinterpretation of model outcomes and potentially to unfair outcomes. The results have been characterized using the Phase-Actor-Role-Task-Skill model and can be used to indicate a lack of involvement and feedback between developer, end-user, and policy-maker roles. End-users and policy-makers often lack the technical skills to interpret the system's output or to estimate potential fairness issues. Therefore, they rely on the technical skills of developers for making apparent technical decisions such as bias and fairness assessments, potentially influencing policy outcomes. As a result, developers take on extra roles such as advisor and decision-maker while potentially lacking the business logic and domain expertise required. This also results in developers being the most prominent role in most tasks and phases of the algorithm's lifecycle. Lastly, we observe that citizens are structurally absent throughout the algorithmic life-cycle, even though it is mentioned that their involvement is needed in the future for balanced fairness assessment. The PARTS model is built to describe these patterns and to be used and extended by domain experts: to audit and design frameworks to deal with communication issues, to support transparency and accountability amongst diverse actors, and to promote the creation of a common vocabulary on fairness assessments.

Appendix

See Table 4.

Table 4 Questionnaire

Questions	Notes
Institution/Department	Name of entity/department
What is your (team's) role?	Description team/staff involved brief
Who do you work with (directly)?	
Domain and topic of use case	
Start and end date	
What type of system is being developed for the use case?	Intended use/aim
What is the goal of the system?	
External partners developing technology for the use case (if any)	
Who are the (end) users—are they directly involved in the development process?	
II. Development process	
Could you guide us through the process of development by mapping out phases—and specific actions in each phase?	
Could you guide us through the decisions made about the system and by whom?	e.g., involvement in deciding on: the goal of the system | design of the system (metrics/labeling/test/training/error) | evaluation of system | monitoring | deploying system
What kind of decisions do you and your team make? (could you give an example)	
What input is needed /do you use to make decisions as a reference point	e.g., handbook, training, expert group?
What kinds of exchanges are needed in your decision process?	
Are there other teams or (external) stakeholders involved in the decision-making process?	
How do you support the decision process of your collaborators with your output?	
III. Considerations	
How can the development process be improved for the following, from your perspective?	e.g., information exchange, role division, handling error rates and biases, handling risks, responsibilities
What are the most difficult challenges and risks of failures for the system?	
How are these challenges and risks measured assessed and monitored?	
What information is needed (by whom) to handle these challenges and risks?	
Who is consulted for this information?	
What is your role in the process of addressing challenges and risks?	
Could any issues occur that might halt the development process, (if so could you give an example of how are these go/no-go decisions determined?)	
In real-life applications, could there be specific risks or negative impacts for individuals or social groups?	
Is error analysis / Bias analysis performed for negative impacts (if so how is this done and could you give an example)?	
Once the algorithm is deployed in practice, what kind of human oversight is available to control for error, bias, or negative impacts?	
What procedures and recourses, if any, are available for addressing the negative impacts of the system?	
Do you have access to explanations or training on the risks for individuals and social groups, e.g., from your colleagues or from external experts?	
IV. Follow-up questions	
Who is in charge / responsible for mitigating measures on respecting privacy and data protection? For instance, is there a valid legal basis for processing personal data?	

Table 4 (continued)

Questions	Notes
Are there cybersecurity or privacy-preserving measures deployed to preserve privacy and data security?	
If no challenges (or very few) concerns are mentioned in 4, provide a scenario?	e.g. complaints about the output; security breaches; what if the training set is not representative, high error rates

Curmudgeon corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Data availability We have no data available, due to confidentiality.

References

Amarasinghe K, Rodolfa KT, Lamba H, Ghani R (2020) Explainable machine learning for public policy: use cases, gaps, and research directions. CoRR. abs/2010.14374. arXiv:2010.14374

Amershi S, Begel A, Bird C, DeLine R, Gall H, Kamar E, Nagappan N, Nushi B, Zimmermann T (2019) Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, pp 291–300

Antoniou G, Franconi E, Van Harmelen F (2005) Introduction to semantic web ontology languages. Reason web 3564(2005):1–21

Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning: Limitations and Opportunities. http://www.fairmlbook.org

Barocas S, Guo A, Kamar E, Krones J, Morris MR, Vaughan JW, Wadsworth WD, Wallach H (2021) Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 368–378

Bovens M (2007) Public accountability. The Oxford Handbook of Public Management. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780199226443.003.0009

Bovens M, Zouridis S (2002) From street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control. Public Admin Rev 62(2):174–184. https://doi.org/10.1111/0033-3352.00168

Chopra AK, Singh MP (2018) Sociotechnical systems and ethics in the large. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (New Orleans, LA, USA) (AIES '18). Association for Computing Machinery, New York, pp 48–53. https://doi.org/10.1145/3278721.3278740

Cobbe J, Veale M, Singh J (2023) Understanding aschccountability in algorithmic supply chains. arXiv preprint arXiv:2304.14749

Commission European (2021) Proposal for a regulation of the European parliament and of the council: laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Content European Commission, Directorate-General for Communications Networks and Technology (2019) Ethics Guidelines for Trustworthy Artificial Intelligence. https://doi.org/10.2759/346720

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp 797–806

Danaher J (2016) The threat of algocracy: reality, resistance and accommodation. Philos Technol 29(3):245–268

De Goede B, Pallister W (2019) Secrecy and methods in security research a guide to qualitative fieldwork. Routledge, London

Dolata M, Feuerriegel S, Schwabe G (2022) A sociotechnical view of algorithmic fairness. Inform Syst J 32(4):754–818

European Commission, Content Directorate-General for Communications Networks, and Technology (2020) The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. https://doi.org/10.2759/002360

Fass TL, Heilbrun K, DeMatteo D, Fretz R (2008) The LSI-R and the COMPAS: validation data on two risk-needs tools. Crim Justice Behav 35(9):1095–1108

Fest I, Wieringa M, Wagner B (2022) Paper vs practice: how legal and ethical frameworks influence public sector data professionals in the Netherlands. Patterns 3(10):100604

Floridi Luciano, Cowls Josh, Beltrametti Monica, Chatila Raja, Chazerand Patrice, Dignum Virginia, Christoph L, Robert M, Ugo P et al (2018) AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Franklin JS, Bhanot K, Ghalwash M, Bennett KP, McCusker J, McGuinness DL (2022) An ontology for fairness metrics. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp 265–275

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 329–338

Fujii LA (2018) Interviewing in social science research. A relational approach. Routledge

Golpayegani, Harshvardhan, Lewis (2022) AIRO: an ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards. ResearchGate
Guarino N, Oberle D, Staab S (2009) What is an ontology? Handbook on ontologies. Springer, Cham, pp 1–17
Haakman M, Cruz L, Huijgens H, van Deursen A (2020) AI lifecycle models need to be revised. an exploratory study in fintech. arXiv preprint arXiv:2010.02716
Hoekstra, Chideock, Van Veenstra (2021) TNO Rapportage Quickscan AI in the Publieke sector II. https://www.rijksoverheid.nl/documenten/rapporten/2021/05/20/quickscan-ai-in-publieke-dienstverlening-ii
Holstein K, Wortman Vaughan J, Daumé III H Dudik M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–16
Holten Møller N, Shklovski Irina, Hildebrandt Thomas T (2020) Shifting concepts of value: designing algorithmic decision-support systems for public services. In: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, pp 1–12
Jonk E, Iren D (2021) Governance and communication of algorithmic decision making: a case study on public sector. In: 2021 IEEE 23rd Conference on Business Informatics (CBI), Vol. 1. IEEE, pp 151–160
Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature 583(7815):169–169. https://doi.org/10.1038/d41586-020-02003-2
Latour Bruno (1992) Where are the missing masses? The sociology of a few mundane artifacts. Shaping Technol/Build Soc: Stud Sociotech Change 1(1992):225–258
Latour B (1994) On technical mediation. Common Knowl 3(2):29–64
Latour B (1999) On recalling ANT. Sociol Rev 47(1 suppl):15–25
Lee MK, Kusbit D, Kahng A, Kim JT, Yuan X, Chan A, See D, Noothigattu R, Lee S, Psomas A et al (2019) WeBuildAI: Participatory framework for algorithmic governance. In: Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), pp 1–35
Madaio M, Egede L, Subramonyam H, Wortman Vaughan J, Wallach H (2022) Assessing the Fairness of AI systems: AI Practitioners' Processes, Challenges, and Needs for Support. In: Proceedings of the ACM on Human-Computer Interaction 6, CSCW1 (2022), 1–26
Ninareh M, Fred M, Nripsuta S, Kristina L, Aram G (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6):1–35
Noy N, McGuinness B (2001) Ontology development 101: a guide to creating your first ontology. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf
Rodolfa KT, Salomon E, Haynes L, Mendieta IH, Larson J, Ghani R (2020) Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In: FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020
Rodolfa KT, Lamba H, Ghani R (2021) Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. Nat Mach Intell 3(10):896–904
Ropohl G (1999) Philosophy of socio-technical systems. Soc Philos Technol Quart Electron J 4(3):186–194
Saldaña J (2013) The coding manual for qualitative researchers. SAGE
Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2018) Aequitas: a bias and fairness audit toolkit. https://doi.org/10.48550/ARXIV.1811.05577
Saxena D, Badillo-Urquiola K, Wisniewski PJ, Guha S (2021) A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. In: Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–41
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency, pp 59–68
Siffels L, van den Berg D, Schäfer MT, Muis I (2022) Public values and technological change: mapping how municipalities grapple with data ethics. New Perspect Crit Data Stud 2022:243
Spierings J, van der Waal S (2020) Algoritme: de mens in de machine - Casusonderzoek naar de toepasbaarheid van richtlijnen voor algoritmen. https://waag.org/sites/waag/files/2020-05/Casusonderzoek_Richtlijnen_Algoritme_de_mens_in_de_machine.pdf
Stapleton L, Saxena D, Kawakami A, Nguyen T, Ammitzbøll Flügge A, Eslami M, Holten Møller N, Lee MK, Guha S, Holstein K et al (2022) Who has an interest in "public interest technology"? Critical questions for working with local governments & impacted communities. In: Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing, pp 282–286
Strauss AL (1987) Qualitative analysis for social scientists. Cambridge University Press, Cambridge
Suresh H, Guttag JV (2021) A framework for understanding sources of harm throughout the machine learning life cycle. In: EAAMO 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtual Event, USA, October 5–9, 2021. ACM, 17:1–17:9. https://doi.org/10.1145/3465416.3483305
Tamburri DA, Van Den Heuvel W-J, Garriga M (2020) Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching. In: 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, pp 391–394
van de Poel I, Royakkers L (2011) Ethics, technology, and engineering: an introduction. Wiley-Blackwell, Cham
van Hage W, Malaisé V, Segers R, Hollink L, Schreiber G (2011) Design and use of the Simple Event Model (SEM). Web Semantics 9(2011):128–136. https://doi.org/10.1093/oxfordhb/9780199226443.003.0009
Van Veenstra AFE, Djafari S, Grommé F, Kotterink B, Baartmans RFW (2019) Quickscan AI in the Publieke dienstverlening. http://resolver.tudelft.nl/uuid:be7417ac-7829-454c-9eb8-687d89c92dce
Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833
Williamson B (2016) Digital education governance: data visualization, predictive analytics, and 'real-time' policy instruments. J Educ Policy 31(2):123–141