Misabstraction in Sociotechnical Systems

De Troya, Íñigo; Kernahan, Jacqueline; Doorn, Neelke; Dignum, Virginia; Dobbe, Roel

DOI
10.1145/3715275.3732122

Publication date
2025

Document Version
Final published version

Published in
FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency

Misabstraction in Sociotechnical Systems

Íñigo de Troya
Delft University of Technology
Delft, Netherlands
I.M.D.R.deTroya@tudelft.nl

Jacqueline Kernahan
Delft University of Technology
Delft, Netherlands
J.A.Kernahan@tudelft.nl

Neelke Doorn
Delft University of Technology
Delft, Netherlands
N.Doorn@tudelft.nl

Virginia Dignum
Umeå University
Umeå, Sweden
virginia@cs.umu.se

Roel Dobbe
Delft University of Technology
Delft, Netherlands
R.I.J.Dobbe@tudelft.nl

Abstract

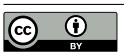
A sociotechnical systems lens on AI is often used to bring attention to the human factors and societal impacts that are often neglected through technical abstraction. However, abstraction is also a general principle of sociotechnical systems, where functional objectives (e.g. fair hiring decisions) are operationalised into low-level implementations (e.g. fair algorithms, recourse, legal basis). The trouble with abstraction arises when critical contextual factors are erroneously neglected, leading to an impoverished representation of the problem space. De-contextualisation can render the resulting solutions problematic when they are re-contextualised back into the site of use, where misabstractions may produce safety hazards, harms, moral wrongs, and context frictions. Despite growing recognition that context matters for how sociotechnical systems operate in practice, the normative implications of abstraction are still understudied. In this paper, we propose misabstraction as an analytic framework for thinking about the perils and challenges of sociotechnical abstraction. We use the framework to analyse the requirements specification outlined in the procurement tender of a recommender system for public employment services and show how misabstractions cascade through the sociotechnical stack, producing ripple effects that implicate *hidden* and *neglected* contextual factors across multiple frames (e.g. institutional, organisational, operational, and algorithmic). Misabstraction can help policymakers, system designers, critical scholars, and civil society alike to attend to the political conditions that shape design, and their implications for understanding and addressing systemic risk in sociotechnical AI systems.

CCS Concepts

• **Applied computing** → **Sociology**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

abstraction, context, sociotechnical systems, artificial intelligence



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732122

ACM Reference Format:
Íñigo de Troya, Jacqueline Kernahan, Neelke Doorn, Virginia Dignum, and Roel Dobbe. 2025. Misabstraction in Sociotechnical Systems. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715275.3732122

1 Introduction

Algorithmic and data-driven systems are increasingly being used to address complex societal and policy challenges in critical sectors such as education [67], healthcare [21], criminal justice [3, 55], and employment [32]. By virtue of their capacity to reveal patterns among vast volumes of data, these applications promise to remove human bias [59], increase service efficiency, augment existing capabilities, and alleviate the burdens of under-resourced public services [72]. At the same time, a growing body of evidence demonstrates how many of these deployments have resulted in undesirable outcomes such as entrenching societal biases [15, 30], eroding caseworker discretion [31, 43], supercharging discriminatory policies [80], and trapping citizens in Kafkaesque "digital cages" [84, 89]. These consequences are often a result of abstracting away contextual factors of the wider system in which they are embedded [5, 77, 106], and thus failing to adequately address the underlying issues they are supposed to solve [2, 95].

There is growing recognition that context matters for how sociotechnical systems operate in practice. Algorithmic systems are fundamentally sociotechnical in nature, requiring technical and social components to be jointly designed for [20, 24, 38]. As such, their potential negative impacts must be understood across the span of their technical and social components [108]. Even algorithmic harms that can be reasonably described within a technical frame can have ripple effects with consequences that propagate across various contexts [36, 90, 92]. There is a need to recognise the limitations of technical methods for ensuring AI safety [36], and to expand our toolkit to consider interventions which acknowledge the influence of social factors at operational, organisational, and institutional levels across the sociotechnical stack [18, 19, 48, 82]. We also need a systemic approach to determine appropriate safety interventions which can account for the ripple effects which give rise to harms [90, 106]. A central challenge to achieving this is interfacing between contexts which are siloed by disciplinary boundaries [14, 99, 121] constructed through different epistemic and ontological commitments [6, 7, 37, 49, 75]; divided by different technical

languages specific to their domain [58]; with different methodological approaches to understanding the negative outcomes of these systems, how to identify them, and how to address them.

Information transfer between such siloed disciplinary contexts necessarily results in abstraction. Abstraction is a general principle of software development, which allows simplifying complex problems into manageable sub-problems by hiding or neglecting low-level details [1, 56, 64]. However, abstraction can also remove nuance which is often important for understanding social problems. Selbst at al [106] described how algorithmic abstractions of social problems (e.g. fairness) are riddled with conceptual traps which frustrate algorithmic approaches to addressing fairness in sociotechnical systems. However, abstraction is not only confined to the technical components of a sociotechnical system. It is also endemic to complex multi-actor systems beyond their technological strata [70, 115]. Systems such as healthcare or public employment services, for example, require various forms of abstraction in order for workers across departments to carry out their own assigned functions. Systems engineering disciplines have long acknowledged the virtues and challenges of abstraction in sociotechnical systems [70, 115]. Abstractions help to simplify rich contextual information in order to relay the minimum viable information load to each subsequent actor or process in the system so that they can fulfill their own functional and safety objectives within the wider system [70]; and to do so in a safe manner without complicating or disrupting the objectives and operations of other actors within the system. However, despite the central role of abstractions in informing the design and analysis of algorithmic systems, a shared language and actionable understanding of its effects on algorithmic harm remains elusive. In this paper, we build on recent efforts in this direction with an emphasis on two foundational aims and contributions:

(1) Contributing to conceptual bases for *sociotechnical frames* for algorithms in their sociotechnical context to inform abstraction practices
(2) Developing a framework for *identifying shortcomings in abstractions* (i.e. "misabstractions") that are central in the design or (harm) analysis of sociotechnical systems

The rest of the paper is structured as follows. In Section 2, we discuss related work and open challenges in articulating harms in sociotechnical systems, and corresponding mitigation strategies. We cover core background on abstraction as a fundamental characteristic of systems-theoretic analyses of complex sociotechnical systems, and observe that the normative implications of sociotechnical abstraction are still understudied. In Section 4, we introduce an ontology for understanding algorithms as part of sociotechnical systems across four different frames; algorithmic, operational, organizational and institutional. In Section 5, we propose a framework for identifying forms of *misabstraction* across the various sociotechnical frames. In Section 6, we apply the framework to a case study of a recommender system for public employment services. In Section 7, we discuss how misabstraction can help system designers and scholars alike to think through the normative implications of abstraction, and we reflect on the limitations of the framework.

2 Core Challenges in Understanding and Addressing Algorithmic Harm

In this section, we cover relevant work on understanding and addressing algorithmic harm in the context of sociotechnical algorithmic systems. We do so along three inherent challenges that relate to the complex sociotechnical nature of harm.

2.1 Harm is an *emergent* system phenomenon

Alongside efforts to address the societal impacts of algorithmic systems through technical interventions (e.g. discrimination in resource allocation, fair ranking, etc. [15, 42, 45, 111]), there is a growing awareness of the ripple effects that technical safety hazards create in the wider sociotechnical systems in which they are embedded [90, 106]. For example, in job recommender systems, an algorithmic harm, such as being shown fewer relevant job opportunities due to predictive models reproducing biased hiring practices learned from training data, often triggers harms which can be better understood beyond the technical frame as sociotechnical harms, such as alienation, loss of opportunity, loss of agency, etc. [16, 45, 108]. These ripple effects implicate other actors, organisations, and institutions, thus broadening the contextual frame that is salient to harm prevention efforts. For example, applicants may demand explanations from frontline caseworkers [32], or seek recourse through the hiring platform [120], and will require an appropriate legal basis to plead their case [13, 93, 97]. Furthermore, hiring agents may feel alienated from their job as it becomes increasingly mediated through opaque decision support tools they feel unprepared to contest [32, 116, 123], caseworkers or hiring organisations may be burdened by addressing complaints [72], and employers may reach fewer qualified applicants [45], thus defeating the alleged virtues of these systems. Thus, these harms are *emergent* in sociotechnical systems resulting from technical, social and institutional components and factors and their interactions [68].

Furthermore, beyond just harms, AI systems may produce other undesirable outcomes such as safety hazards [39], moral wrongs [35], and tensions [118] that do not necessarily produce harms or result from a moral deficiency, but are nevertheless undesirable. Misunderstandings, misalignment of mental models, vague requirements specifications all contribute to sociotechnical systems simply not working as intended [95].

2.2 Addressing sociotechnical harms requires distributed efforts and *coordination*

Ensuring AI safety requires preventative and reactive interventions that span across technical, operational, organisational, and institutional contexts [18, 19, 71, 82], including e.g. legal reform [117], regulation [85], oversight [9, 50, 96], enforcement [54, 85, 114], civic advocacy [91], shifting organisational culture [33, 39], affording users meaningful discretionary power to correct erroneous outputs [123], providing organisational mechanisms for recourse [22, 23, 46, 63], facilitating dissent [11, 37, 52, 119] and more. For example, to ensure that decision-support tools in public services don't produce undesirable outcomes, not only do they need to be algorithmically 'fair', but they also need to integrate well into existing service design workflows [74, 123]; responsive channels for complaint should be in place in case something does go wrong [37];

and we require adequate laws and regulation [93, 117], along with means of enforcement and the power to penalise infractions [54]. A narrow view on the technical safety of the artifact overlooks the various operational, organisational, and institutional mechanisms that need to be in place for the system to work [94].

The harms emerging from sociotechnical systems can materialise at different sites, depending on what contextual factors are brought into focus [90, 108]. Sociotechnical harms can emerge due to unsafe conditions and mechanisms that occur at different sites of the sociotechnical stack, which compound before eventually producing in an actual harm [69].

The interventions to eliminate or mitigate the risk of harm therefore require responsibility and accountability structures that are inherently *distributed* across different actors requiring *ongoing coordination and management* through both formal and informal institutional interventions.

2.3 Understanding harm requires *integration* of different forms of knowledge, expertise and experience

We still lack a means of understanding the ripple effects of systemic risks in AI. Thankfully, ripple effects have been widely studied in other fields such as disaster response, resilience engineering, and supply chains [60, 78, 102]. In order to deal with ripple effects, there needs to be coordination among actors to effectively address the emergent systemic issues. Coordination requires alignment of mental models [39], declaring assumptions [25, 62], clearly communicating requirements [58, 109], negotiating priorities to work towards common objectives [51, 98], and working with proxies in light of uncertainty [88].

However, intervention practices to mitigate harms are typically "siloed" [14, 99, 121], undertaken by different actors with different priorities [57], specialised technical language [58], and varying degrees of access to a system [51, 96, 110], while operating with different mental models of a system that reflect their situated awareness of salient contextual factors [26, 39, 110]. Crossing disciplinary boundaries may be complicated by a communication gap created by technical language (e.g. legalese, tech-speak) [58], different epistemic and ontological commitments [6, 7, 37, 49, 75], and erroneous assumptions about other actors [48, 110]. Given the complexity of many sociotechnical AI systems, most actors operate with varying granularity of focus into different aspects of the system, from low-level detail in their areas of expertise, to higher-level views of adjacent disciplines. Thus, before effective interventions for harms can be developed, an understanding of the system must be constructed through the *integration* of different forms of knowledge, expertise and experience across different actors. This necessitates a *shared language* for modeling and addressing harm. In order to develop this language, abstractions are required to negotiate, communicate, validate and enforce particular system design constraints. Good abstractions are crucial for coordinating across sociotechnical systems and ensuring the efficacy of safety interventions.

3 Abstraction in Sociotechnical Systems

In this section, we cover literature and core insights and aspects of abstraction in the analysis, design and governance of sociotechnical systems. We do so, covering two core purposes of abstraction as well as its connection to context.

3.1 Abstraction for simplification

The word 'abstraction' originates from the Latin *abstrahere*, literally, "to draw away". Abstraction is a general principle of software development [1, 56, 64], defined by Ousterhout as "a simplified view of an entity, which omits unimportant details" [87, p.33] – a phenomenon known as *information neglect* [27], whereby formal models are constructed by selecting and discarding features of interest to the modeller. Software abstractions allow low-level details to be encapsulated into a form that can be orchestrated to satisfy some high-level functionality – a phenomenon known as *information hiding* [27], which allows programmers to only think about a function's behaviour and ignore the details of its implementation. Similar to mathematical functions in the form $f(x)$, programming functions allow the mechanics of a routine to be obscured behind a mapping of inputs to outputs, which can be interfaced with other functions. Poor abstractions can lead to mismatched interfaces between functions, resulting in program failure and undesired outcomes [28]. Information hiding and information neglect are fundamentally political choices about what contextual factors are considered important or not, and how to represent them. As such, abstractions may have significant social implications.

While abstracting away social context can make it easier to render social problems calculable, they may also fundamentally reconfigure the context of use in the process [5, 10], even redefining the original problem to suit the proposed (techno)solution [61, 109]. This can also create impoverished representations of the problem that we are trying to solve. Technical abstractions of social problems can create conceptual traps that make seemingly sound technical solutions fall apart in practice [2, 17, 71, 106]. A narrow technical frame also risks losing sight of the downstream sociotechnical harms that emerge in the wider context in which these technical solutions are embedded [108]. However, abstractions are not only produced in the technical system, but are also an intrinsic aspect of sociotechnical systems.

3.2 Abstraction for information sharing

A growing community of scholars is drawing on the discipline of system safety engineering to bring a sociotechnical systems lens to the study of AI safety [34, 39, 94, 100] System safety is a discipline that has grappled with sociotechnical systems by accounting for technology, institutions, and processes across multiple contexts including technical, operational, organisational, and institutional domains. Leveson [69] frames system safety through the lens of control theory, whereby a system is kept safe through monitoring feedback signals and responding with control signals in order to keep the system within a safe operating state. Leveson [69]'s systems ontology for safety is built upon Vicente & Rasmussen's Abstraction Hierarchy in systems engineering [115] which describes how high-level functional purpose (e.g. ensuring safety) is realised at lower-level system operations (e.g. a safety mechanisms such as debiasing techniques or procedures for filing complaints) through the propagation of design specifications transacted across stakeholders including management, customers, systems engineers, component

engineers, and system operators [115]. Abstractions help to simplify rich contextual information in order to relay the minimum viable information load to each subsequent actor or process in the system so that it can fulfill its own functional objectives within the wider system [70]. Abstraction can thus also be understood as a fundamental building block of sociotechnical systems, extending beyond their technical strata.

An actor's positionality with respect to the broader system informs their visibility and understanding of it. This informs their *mental model* of the system. Different actors will have different mental models of the same system. For example, consider a data scientist and a caseworker, both working in public employment services. Each will have differing levels of visibility, exposure, and understanding of different components of the system. These will inform their own situated knowledge and thus, their distinct mental models. Effective abstractions allow socially situated actors to establish sufficient shared understanding for coordination, so that they can fulfill their own functional objectives within the wider system. For example, model explanations are abstractions which can create a shared understanding of model logic that both data scientists and caseworkers can use to inform their own decisions. Conversely, poor abstractions can lead to misaligned mental models (e.g. false assumptions, misunderstandings, information asymmetry, etc.) and frustrate coordination.

3.3 Context is constructed, dynamic and negotiated

Determining what to account for in abstractions requires understanding what is relevant contextual information, and what can be left out. Efforts at embedding some notion of "context-awareness" in technical systems often adopt a *representational view* of context, whereby contextual factors can be rendered measurable and calculable [4]. However, context is constantly shifting [46, 66] and cannot simply be captured in a static representation [40, 105]. In line with the core challenges identified in Section 2, that harm is emergent, requires distributed and coordinated efforts across various actors and the integration and negotiation of stakes and expertise, Dourish argues that we need an *interactional view* which acknowledges that context is negotiated, contested, continually (re)constructed, and subject to continual processes of interpretation and reinterpretation. Context is therefore politically contingent and relative to the situated perspective of whoever is defining the frame of relevance. As such, it is crucial to critically reflect on where that system boundary is drawn, how the notion of what is deemed 'relevant' context is defined, and by who.

In the next section, we introduce *canonical frames* for representing algorithmic systems in an *increasingly interactional fashion*, moving from the algorithmic frame to a rich sociotechnical frame. These then serve as a basis to develop a framework for identifying misabstraction in sociotechnical systems in Section 5.

4 Framing the sociotechnical stack

In this section, we draw on the history of *sociotechnical systems theory* to situate AI systems within their broader social, organisational and institutional context. Taking a systems view allows us to examine abstraction across three forms of design, namely technological design, institutional design, and process design [65]. Below, we propose a lexicon for framing the different contextual layers that make up the *sociotechnical stack*. We begin with (1) the *algorithmic frame*, and then expand Selbst et al's notion of the sociotechnical frame [106] to include (2) the *operational frame*, (3) the *organisational frame*, and (4) the *institutional frame*. We develop these frames by drawing on the sociotechnical systems ontology of software-based automation developed in the system safety literature [39, 69]. At each frame, we introduce some of the core artifacts that contribute to the broader *operational process* of the system.

4.1 The algorithmic frame

The algorithmic frame concerns the technical artifacts and subsystems such as the algorithmic model, input data, target labels, model outputs; as well as the software that constructs, contains, and interfaces with the model; and the hardware and infrastructure necessary for development and deployment. This frame also concerns the measures of performance of these components, such as evaluation benchmarks, performance metrics, fairness criteria, etc.

While certain aspects of harm can be related to model errors or bias or other technical malfunctioning in the algorithmic frame, actual harms *occur and are experienced* in the actual use and operation of the algorithmic model and, as such, cannot be fully understood in the algorithmic frame [39]. As Selbst et al. argued, the algorithmic frame was the basis for most early scholarship on fairness in algorithmic systems, and richer *sociotechnical frames* were lacking [106]. In the following, we build on lessons in sociotechnical systems theory to build out the sociotechnical frame incrementally across three subframes: the operational, the organisational, and the institutional. [20, 65, 115].

4.2 The operational frame

To describe the *operational dynamics and factors* that contribute to algorithmic harm, we need to extend our frame to consider the *core operational process* and the associated *operational control and decision-making functions* [39, 69]. Within this frame lies the operational system, which can be described as the *controlled process*, the *controller*, and their coupling.

The *controlled process* refers to the operational process which is to be maintained within a safe operating margin through the application of control actions by the *controller*. For example, in a hiring context, the controlled process may be a decision-support tool that makes hiring recommendations, while the frontline caseworker acts as a controller in using the tool and overseeing its operation. In the system safety literature, the coupling between the two is expressed through complementary *reference and control channels*. The *reference channel* represents the relay of information *from* the controlled process *to* the controller, describing its operating state. The *control channel* represents the means by which the controller maintains the underlying controlled process within a safe operating margin (e.g. that the recommendations or predictions made by the decision-support tool are fair, seem reasonable, etc.).

This frame allows us to see the *direct factors* which contribute to the operation of the controlled process, namely the human-machine interactions which govern the technical artifacts in the algorithmic frame. However, this frame *hides* the *indirect factors* which create

the context in which these interactions are orchestrated. In order to understand the conditions that shape the operational frame, we need to expand the sociotechnical frame further to capture the organisational and institutional factors that constitute the broader sociotechnical system [65, 69].

4.3 The organisational frame

The organisational frame describes the broader set of processes surrounding the operation and maintenance of an algorithmic system. These processes include, for example, design activities, management of operators, best practices, and the provision of complaint procedures [68]. These indirect factors provide the context in which operational and algorithmic processes in the preceding frames unfold.

Organisations are typically composed of formal hierarchical structures, including vertical lines of command and horizontal relations across teams [70]. Bringing these factors into view allows us to examine how functional objectives can be operationalised in practice [115]. This requires understanding what objectives are prioritised, how mental models are formed [39], what assumptions are held by different actors [25, 62], how requirements are communicated [58, 109], how design processes are documented, and so on. These relations and practices are defined by institutions: the written and unwritten social rules and norms that structure the behaviour and interaction of actors, processes, and systems.

The organisational frame can capture multiple interdependent organisations, such as the organisation that publishes a public procurement tender, and the organisation that is awarded the project. As such, the organisational frame allows us to map both intra-organisational relations (such as between business units or individual workers), as well as inter-organisational relations (such as contracts for the procurement and maintenance of systems).

4.4 The institutional frame

Institutions are rules that structure and constrain human behaviour. These can be both formal (e.g. laws, standards, rights) and informal (e.g. codes of conduct, norms) [81]. Institutions shape and inform the design and operation of both the technical and the social components of sociotechnical systems [65, 83].

Formal institutions inform the design of processes such as the operational process (e.g. by mandating some norm on human oversight), and oversight and maintenance procedures (e.g. by setting requirements on how a risk assessment should be done). Informal institutions can also determine the shape and efficacy of the core processes, both within organisational control mechanisms (e.g. mobilising informal leadership to enforce a norm) as well as in the design of processes itself (e.g. practising one's discretionary power to ignore a particular norm and await enforcement or legal procedures).

Expanding the sociotechnical frame to account for institutions allows us to develop a richer understanding for both the interrelations between system components, as well as the underlying conditions which give rise to certain system configurations (and not others). Furthermore, institutions allow us to articulate the normative dimensions of sociotechnical systems by probing design choices, social relations, and governance mechanisms.

5 Misabstraction: a reflexive analytic framework for system-theoretic abstraction

In Section 3 we presented abstractions as an intrinsic property of complex sociotechnical systems. We now focus on abstractions that are problematic in reference to the sociotechnical subframes introduced in Section 4. In this section we introduce the concept of a misabstraction.

We define a misabstraction (noun) as *a representation of an entity, phenomenon, or procedure that omits critical contextual information and renders that representation problematic when it is reintegrated into the context of the sociotechnical system for which it has been made.* To misabstract (verb) is then the process(es) by which such misabstractions are produced.

Misabstraction occurs through a two-stage process of (i) decontextualising information from the target site, and (ii) recontextualising an intervention back into the site. During de-contextualisation, salient contextual factors from the target site are *neglected* or *hidden* in order to simplify the problem space into a design task *"in the lab"* [48]. Following the design of the artifact or intervention at the site of development. it is re-contextualised back into the target site, where it is exposed to those contextual factors that were *neglected* or *hidden*. The process through which misabstraction takes place is illustrated in figure 1.

Contextual fracture takes place when the designed intervention is confronted with contextual factors that were not accounted for during design. As a result, prior linkages between contextual factors may be broken. For example, introducing a decision-support tool to mediate interactions between caseworkers and their clients may interrupt the face-to-face communication they would have had prior, through which they may have developed a better understanding of the client's particular situation or circumstances. Here, the importance of the interpersonal rapport between the caseworker and the jobseeker is the social factor that was misabstracted, resulting in a fractured context in which their communication is impoverished, thus potentially frustrating the functional objective of providing jobseekers support in navigating their options, given their personal circumstances.

5.1 A recipe for misabstraction analysis

A misabstraction analysis can be performed by following the schema outlined in Figure 1 and specifying the following elements:

(1) a *contextual frame* within which misabstraction is taking place
(2) a *contextual factor* that is abstracted away in the transfer of information to the **site of design**
(3) the resulting *misalignment* between the site of design and the site of use
(4) a *limitation* of the intervention that causes contextual fracture when it is *re-contextualised* back into the **site of use**
(5) a *consequence* of the limitation (e.g. a safety hazard, harm, moral hazard, or friction), including ripple effects across other contextual frames.

We apply this recipe to a case study below.

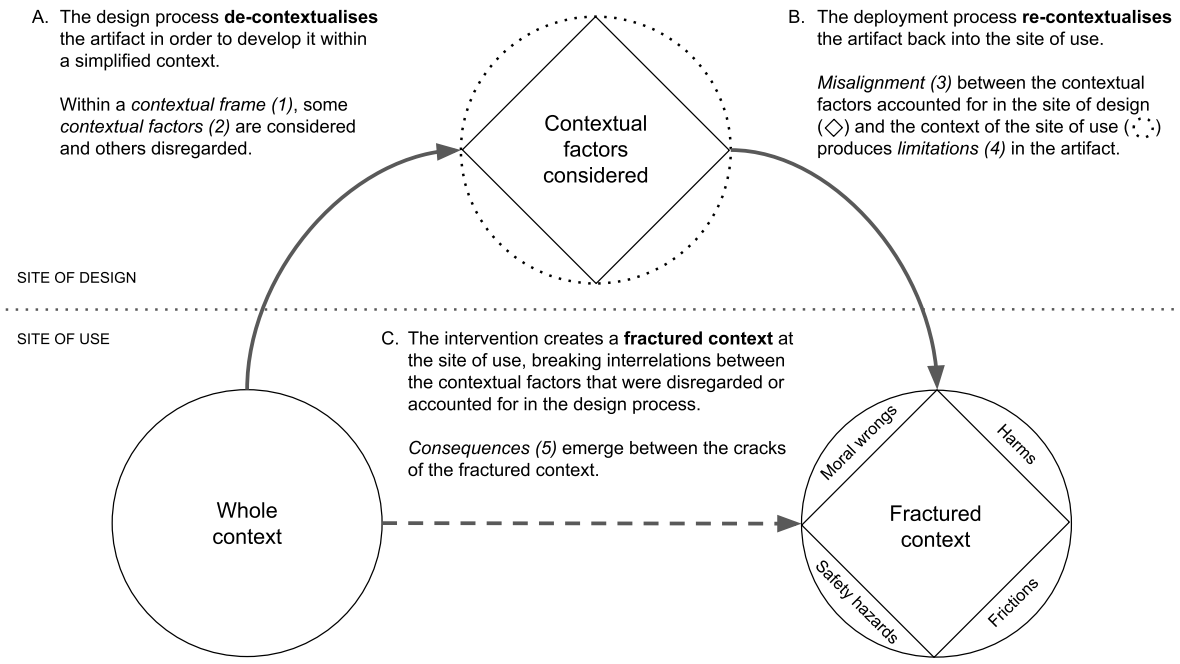


Figure 1: Misabstractions are abstractions that are made about contextual factors in the site of use during the design process (e.g. simplified representations of entities, phenomena, or procedures which rely on incorrect or incomplete assumptions), which have negative consequences when redeployed back into those sites (e.g. sociotechnical harms, safety hazards, moral wrongs, and frictions). In the figure, the circles represent the whole context of the real world; the squares represent the contextual factors that are taken into account at design time; while the discrepancy between the two may give rise to negative consequences. Steps A, B, C outline the process of de-contextualisation, re-contexualisation, and context fracture. Numbers (1) - (5) correspond to the elements of the recipe for misabstraction analysis in section 5.1.

6 Case study: a public procurement tender for a Skills-Matching System at a Public Employment Service

In this section, we apply the misabstraction framework to a particular context of coordination between different actors working towards the design of an algorithmic system for a Public Employment Service (PES) in The Netherlands (Uitvoeringsinstituut Werknemersverzekeringen or UWV). We look at a tender process between the PES and possible external developers. We first present our dataset (tender documentation giving a specification of the intended system), and review known harms and factors for similar algorithmic systems. We then apply the misabstraction framework relative to the different sociotechnical frames from Section 4.

6.1 Tender documentation

We consider misabstraction in the design specification of a skills-matching system (SMS) at a PES. To do this, we obtained publicly available documentation through an online portal for public procurement tenders in the Netherlands, https://www.tenderned.nl/. These materials reveal what aspects of system design are considered relevant to the PES. The public procurement tender documentation for the SMS [113] (hereafter, the tender) was made up of 17 documents, including system requirements, process requirements (including labour standards, security guidelines, etc.), and administrative paperwork (including application forms, financial declarations, etc.). In our analysis we focus on the documentation related to system requirements, which amounted to 6 separate documents, the contents of which are summarised in Table 1 (see Appendix A), annotated as [D1] - [D6] in the remainder of the paper.

6.2 Description of the system specification

The tender describes a skills-matching system which is meant to support jobseekers in reintegrating into the labour market. The stated functional requirements [D1. pp. 6-9] include "search and find", "matching", "knowledge management", "analysis of distance to the labour market", "reducing mismatch", "profiling and segmentation", "data enrichment and analysis". The system achieves this by profiling and segmenting job candidates based on a skills taxonomy, CompetentNL (similar to existing taxonomies such as ESCO). The taxonomy serves as an interoperable standard which can be used by different public and private parties, including software developers and the national Public Employment Service.

The tender includes procedures for reporting security incidents, including breaches of personal information, with reference to GDPR Articles 4(12), 33, and 34 [44] [D4]. While the tender has a strong emphasis on security and privacy, as we note in our analysis below, there is a distinct lack of attention to algorithmic discrimination.

6.3 AI in hiring

AI tools have been used to address a range of challenges in hiring, such as candidate sourcing, screening, selection, and evaluation for employers [45]; unemployment risk prediction for caseworkers [103, 124]; or job and training recommendation for jobseekers [53]. While some champion AI as a means of eliminating human bias in hiring decisions [59], it is no surprise that these systems machine learn societal biases and reproduce them at scale [16, 45, 47, 107]. Beyond algorithmic biases, these systems produce other forms of sociotechnical harms and moral wrongs, such as loss of agency in representation [8, 104]; performing better for individuals that conform to stereotypes well-represented in sample populations [86, 122]; obfuscating decision-making logic such that caseworkers' discretionary power is undermined by their reluctance to disagree with the computer [112]; alienating caseworkers when AI tools don't fit into existing workflows [116, 123]; operationalising austerity politics that marginalise people predicted to be at risk of unemployment [80]; codifying the notion of an ideal candidate through prioritisation regimes [41]; unhelpfully reducing individuals' employment needs to risk statistics [12, 123]; and more.

While in this paper we will focus on a case study public employment services, we will also draw on literature on private sector applications as those practices tend to be eventually inherited by the public sector. For example, while recommender systems are commonplace in commercial hiring platforms, their adoption in Public Employment Services is still in its infancy [53, 79].

6.4 Results: applying the misabstraction framework to the case study

We now apply the misabstraction framework to the case study. We consider how misabstractions of critical contextual factors occur across each of the four contextual frames (algorithmic, operational, organisational, institutional). We show how (i) these misabstractions ripple across (are inherited by) the other frames , and (ii) how the nature of their problematic nature may shift in the process, for instance, first producing a safety hazard in one frame before materialising as a sociotechnical harm in another frame. We follow the recipe for misabstraction analysis outlined in section 5.1 and enumerate the components of the recipe accordingly for the misabstractions we have identified.

6.4.1 Misabstraction in the algorithmic frame: evaluation criteria missing. The tender makes no reference to evaluating the performance of the recommender system. The text suggests that "the quality of the results [...] is increased", but provides no basis on which to demonstrate that claim:

> *"UWV collects data about vacancies and profiles to enable searching, finding and matching.* ***The quality of the results of searching, finding and matching is increased*** *by standardizing and using data based on taxonomies or ontologies [. . .]"* [D1, p.1]

This claim is not substantiated by any choice of metric or declaration of a minimum threshold of acceptable performance. For example, there is no mention about whether the system should be able to provide accuracy comparable to or better than a human caseworker. Phrased differently, the tender provides no guideline regarding the acceptable operating margins for reliability. As such, the algorithmic system may be delivered while giving altogether unhelpful recommendations. There is essentially no requirement that determines an acceptable quality of system performance.

Furthermore, this exclusion also means that there is no requirement to provide a stratified evaluation of model performance based on different candidate profile segments, e.g. demographic characteristics or job sector (which may be a proxy for demographics [30]). This would be a first step in auditing for bias, but is missing here. As a result, the system may work better for some users than for others. In this sense, the omission of evaluation criteria is a normative blind spot, because it fails to address the inequitable distribution of system errors and their subsequent social impacts.

Following the recipe outlined in 5.1, the contextual factor being abstracted from the algorithmic frame <1> is *quality assurance* <2>. This requirement is *de-contextualised* from the problem site by omission from the tender <3>. Since this requirement is not communicated to the supplier (the 'solution owner'), their *mental models will be misaligned* with the procurer's mental model of the problem context (the 'problem owner'). Since the tender does not specify this as a design requirement, the supplier does not need to take this factor into account when they design the system in order to be compliant with the tender. As a result, when the system is *re-contextualised* back into the site of use, the system will *fracture* the context by not meeting normative expectations of quality assurance which, while not communicated in the tender, will nonetheless be present <4>. As a consequence <5>, it is possible that the system will perform inequitably for different populations, which may result in sociotechnical harms such as opportunity loss that are distributed in a discriminatory manner.

6.4.2 Misabstraction in the operational frame: unclear if sufficient interpretability given to provide meaningful feedback. In the operational frame <1>, the tender specifies that caseworkers can suggest *"structural adjustment of match results by providing feedback to the supplier about the search results"* [D1, p.8] [113] However, it's not clear whether there will be any facility provided to caseworkers that enables them to interpret why certain matches were given. The contextual factor that is abstracted away is the fact that caseworkers need to be able to interpret the model in order to provide meaningful feedback <2>. System developers are not made aware of caseworkers' information needs and situated expertise (*information hiding*), and thus may not account for these at the site of design <3>. As such, caseworkers may be limited in their analytic capacity to identify and contest problematic model outputs <4>. This limitation may prevent them from surfacing patterns of algorithmic discrimination <5>.

6.4.3 Misabstraction in the organisational frame: complaints procedure missing. In the organisational frame <1>, the tender does not specify any mechanisms through which jobseekers may file a complaint or provide feedback about their experience with the system. Jobseekers' potential need to contest perceived unfair treatment is thus abstracted away from the design specification <2>. While it may be possible that such a complaints mechanism is already in place at UWV, it is not integrated into the tender's system design specification such that complaints may be more directly

coupled with the actual system functionality <3>. This also suggests that any staff currently responding to complaints may not be knowledgeable enough about the system to provide an adequate response to emergent issues that may be specific to the nature of the algorithmic system. Furthermore, there is no mention that staff at the existing complaints department should undergo training to field complaints about this new system. If jobseekers need to file a complaint through existing channels that are not designed into the system, it may create friction when trying to describe the situation they are experiencing, without support from the system to help situate the source of a complaint. This may make it more difficult for jobseekers to file their complaint or may discourage them from raising an issue altogether <4>. As a result, sociotechnical harms, e.g. opportunity loss due to poor job recommendations, may emerge and lack adequate response mechanisms <5>.

6.4.4 Misabstraction in the institutional frame: omission of anti-discrimination regulation. In the institutional frame <1>, we find that while the tender makes note of appropriate regulation pertaining to the processing of personal information [D3, D4], it does not make reference to non-discrimination law <2>. This omission is particularly salient given the known potential for discrimination in both algorithmic systems such as recommender systems, as well as in the hiring context more generally (with or without algorithmic intermediation). Algorithmic discrimination in the hiring context has been widely documented, both directly through protected attributes such as age and gender, as well as through proxies such as occupational history (see section 6.3).

While the organisation may claim that they generally abide by extant employment regulation, including non-discrimination law, it is nevertheless not made an explicit design requirement in the tender. As such, while UWV may claim to abide by such laws, this obligation is not communicated to the suppler in the system specification <3>. This institutional neglect further manifests itself in the marked absence of any requirements for the recommender system to be audited for potential biases at any point in its pipeline. As a consequence of this lack of bias measurement, there is evidently no mention of bias mitigation measures in place in the event that such biases may emerge. This omission leaves the organisation altogether unprepared for the eventuality of algorithmic discrimination produced via the recommender system <4>, as the supplier is not required to ensure that be appropriate response mechanisms are in place to resolve these challenges when they do arise <5>.

6.5 Sociotechnical stack trace: context fracture beyond harms

Given the nature of the case study – a tender for a system that has yet to be built –, it can help us to appreciate the fact that no material (or immaterial) harms have yet to be produced. Nevertheless, it is evident that the resulting consequences of these misabstractions are still undesirable. Our analysis allows us to probe how misabstractions give rise to conditions that may result in harms downstream. Understanding these conditions also invites us to expand the scope of inquiry to other undesirable consequences beyond harms, such as safety hazards, moral wrongs, and frictions. We here introduce the concept of a *sociotechnical stack trace* to follow how a misbastraction's ripple effects propagate across contextual frames before their consequence finally emerge.

In the algorithmic frame, the exclusion of evaluation criteria presents a *safety hazard* by not providing a measure through which harmful outcomes such as algorithmic discrimination may be identified and mitigated prior to deployment. While the *hazard* originates in the algorithmic frame, the consequent *harms* of the potential algorithmic discrimination will be experienced by the jobseeker in the operational frame.

In the operational frame, the lack of attention to caseworkers' information needs may cause *friction* which may make it difficult for caseworkers to provide meaningful feedback to the system developers. As such, system developers may have a false impression that their system works well simply because caseworkers are not submitting feedback, whereas in reality, they may simply not feel empowered to do so. These *frictions* may cause *safety hazards* to go undetected, eventually producing tangible *harms*.

In the organisational frame, the lack of a complaints procedure through which jobseekers may raise concerns may similarly result in *friction* by which jobseekers will struggle to voice their concerns with the system. This misabstraction in the organisational frame may lead jobseekers to raise their frustrations directly with the caseworker in the operational frame.

In the institutional frame, the lack of anti-discrimination regulation is a neglect of a *moral wrong*. This neglect fails to draw system developers' attention to the potential for algorithmic discrimination, and what legal bounds they must ensure the system operates within. As before, the consequence of this *wrong* will eventually be felt as a *harm* to the jobseeker in the operational frame.

By expanding our scope of interest beyond harms to also include safety hazards, moral wrongs, and frictions, we can better appreciate how undesirable consequences of abstraction have ripple effects throughout the sociotechnical stack. The four sociotechnical frames help us to trace the normative dimensions of abstraction as they traverse the sociotechnical stack.

7 Discussion

7.1 Surfacing unknown knowns

In the case study examined above, the tender serves as a vehicle for communicating the social needs that the resulting systems should address. Omitting critical contextual factors (e.g. anti-discrimination law, complaints procedures, caseworkers' information needs, evaluation criteria) precludes the downstream development teams from incorporating them into their own mental models of the site of use, thus preventing them from considering those factors at the design site. The specification is itself an abstraction which should *hide* irrelevant details but not *neglect* critical ones. The actors involved in constructing and issuing the tender therefore have a responsibility to ensure that abstraction contains all the critical contextual factors for any resulting intervention to correctly fit into the target site.

Consideration of the sociotechnical context is often proposed as a solution to avoiding harms created by technical abstractions [106]. The tender analysed in our use case lays out what may be described as a sociotechnical specification, in the sense that it accounts for the

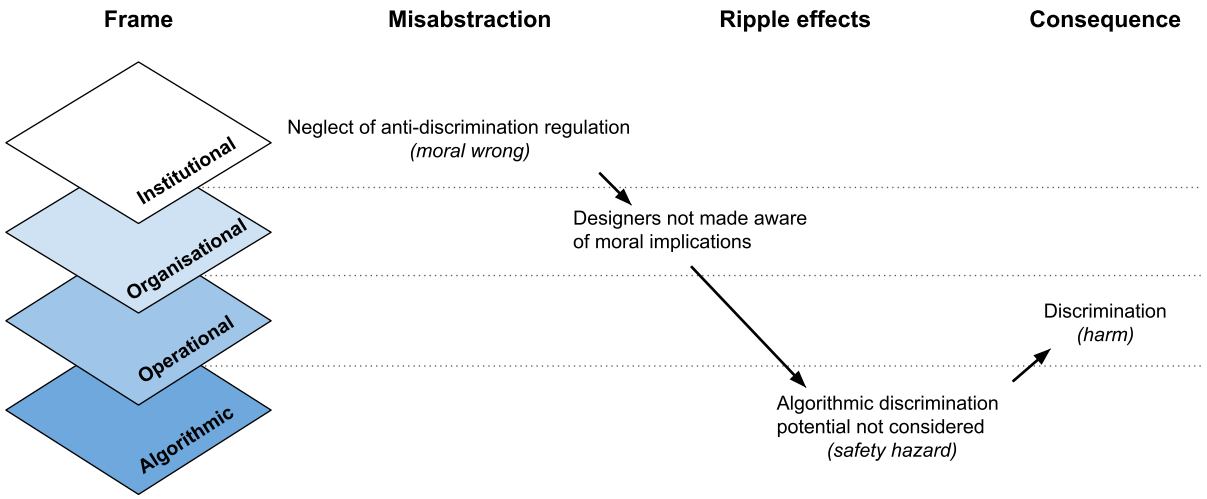


Figure 2: A sociotechnical stack trace can help to show how misabstractions ripple across contextual frames before their consequences emerge. In this figure, we map one of the misabstractions identified in section 6.4.4 and see how a misabstraction can begin as a moral wrong before becoming a safety hazard, thus creating the conditions for a harm to emerge.

joint design of technical and social components of the system [20]. However, as we have shown, critical contextual factors are missing across each of the algorithmic, operational, organizational, and institutional frames. This illustrates the challenge of considering the sociotechnical context. It is unknown by each actor exactly what is known by other actors. A directed approach is needed to identify the contextual factors that are important to facilitate sufficient sharing of information.

Misabstraction analysis can serve as a directed reflexive device for domain experts to surface and attend to those factors that should be accounted for in order to prevent *context fracture*. It can be used internally to identify problematic exclusions or mischaracterisations from a design specification based on existing norms and best practices. While the framework won't necessarily surface all the contextual factors that need to be included, it facilitates a reflective design process for tracing the ripple effects and consequences of abstracting contextual factors.

7.2 Do abstractions have politics?

Cause and effect of misabstractions across sociotechnical frames. In section 6 we show how abstractions in a single frame have ripple effects through other frames. Abstractions in one frame may have consequences that emerge in other frames. However, misabstractions are not the beginning of these causal chains. They themselves are the consequences of prior conditions across the four frames. This is observed in the results of our analysis. For example, in the algorithmic frame, the omission of evaluation criteria is a normative blind spot or *moral hazard*, because it fails to address the inequitable distribution of system errors that emerge due to potential algorithmic discrimination in the algorithmic frame. The consequences of this omission may be felt in the operational and organisational frames where actors are not familiar with the limitations and vulnerabilities of algorithmic tools (designers, caseworkers, jobseekers, and others alike).

In contrast, the tender considers security and privacy concerns in detail. This reveals an organisational blind spot for institutionalised discrimination, that is not present for security and privacy. This is caused by underlying conditions in the organisational or institutional frames. For example, there could be a culture of hesitation to address issues of discrimination for fear of judgment or reprimand for failing to solve issues that are only partly caused by these design choices [33]. This demonstrates that it is not sufficient to analyse misabstraction within a single frame. Instead, the ripple effects of misabstractions suggests that we need an integral view of the system to identify the causes and consequences of misabstractions.

Misabstractions reveal and circumscribe power. Identifying the causes of misabstractions requires an understanding of the power dynamics within a system. What is (mis)abstracted is determined by who has the power to make abstractions and to delineate system boundaries within which to do so. Abstraction can be a way of capturing (in Agre's sense [5]) a system according to one's specific agenda or priorities.

Abstractions are political; a product of choices made by socially situated actors [29, 110]. Those who hold power – to make decisions, to set agendas, to raise awareness [73] – can impose their priorities onto a system through abstractions informed by their own mental models, whether intentionally or not. These abstractions make choices about possible worlds, and close off other possible systems [5]. The resulting system configurations then enable and constrain the behaviours and actions of its users and subjects according to the consequences of the abstraction. As such, abstractions form the grammar of the 'scripts' [10] which actors must follow in order to engage with an artifact or system. In this way, power is enacted and translated through abstractions and their ripple effects. It is important to note that this power may be wielded unknowingly, and not necessarily with malicious intent. Nevertheless, the consequences of how decision-making, agenda-setting, and awareness-raising

produce abstractions beckons us to understand the conditions that give rise to certain representations and not others.

In addition to minding the politics of those who make abstractions, we must also attend to the experiences of those who bear their burdens. In the case study we examined, we focused less on who made the misabstractions in the tender, and more so on those actors downstream, such as design teams receiving the tender as a system specification; caseworkers using the system in practice; and jobseekers subject to the system. Identifying the actors implicated across the different contextual frames allows us to ask questions about how misabstractions are experienced, identified, and addressed. Whose voices are heard? What response do they get (if any)? Who is responsible for resolving misabstractions when they are made known? How do power relations across the sociotechnical system facilitate or obstruct the identification and resolution of misabstractions? Epistemic diversity and ontological reflection are critical in order to understand how to determine which contextual factors are critical at specific junctures. The sociotechnical stack trace (Section 6.5 and Fig. 2) allows us to begin to unpack the origins, ripple effects, and consequences of misabstraction, and to identify which actors we may engage with in order to address these issues.

Our analysis has made evident the need for a more approach to understand and analyse systems across the various sociotechnical frames, in order to better track how conditions created by misabstractions ripple throughout the stack. Lessons from system safety make evident the need for understanding harms as an emergent phenomenon that requires a systemic approach, including establishing a safety culture within the organisation, which promotes better understanding, coordination, and resolution of misabstractions and their consequence. This includes considering how power relations within the problem specification and design process arenas privilege certain mental models over others, at the expense of misasligning situated perspectives that could be more integrally coordinated.

7.3 Future work

In this paper, we focused on a tender for a sociotechnical system for introducing a technical artifact into an organisation. While building such a system requires interventions beyond the algorithmic frame, in the case study we examined, those are ultimately intended to support the introduction of the algorithmic system itself. We invite scholars, practitioners, policymakers, and civil advocates alike to reflect on misabstractions in other kinds of sociotechnical interventions, such as in the development and implementation of governance instruments.

At a political level, the misabstraction framework may help to articulate the deficiencies introduced in AI regulation by the influence of private interests, as in the European Commission's General-Purpose AI Code of Practice for determining the obligations of model providers for identifying and addressing systemic risks [76]. For practitioners, the misabstraction framework may help to understand how to better account for the perspectives of impacted stakeholders in the design of governance interventions. For example, algorithm registers still struggle to meet the transparency needs of socially situated actors [101]. Misabstraction can help to highlight the implications of hiding or neglecting certain forms of information from such registers.

Lastly, we also invite further conceptual and empirical research on how to determine appropriate degrees of abstraction in design, development, and deployment alike. In addition to exploring these questions in site-specific studies, it may also be interesting to explore more general design patterns, similar to the recipe and sociotechnical stack trace we provided here.

8 Conclusion

In this paper, we advanced our understanding of the normative implications of abstraction in sociotechnical systems. We introduced the concept of misabstraction as a means of understanding how the exclusion of critical contextual factors from the design process can render the resulting interventions problematic when they are introduced into the site of use. We provided a conceptual framework, made actionable through an analytic procedure, and applied it to a case study of algorithmic systems in public employment services (PES). Our treatment of the case study revealed a set of systemic misabstractions that may inform a more rigorous tender coordination through critical reflection on the emergence and implication of abstraction practices. To achieve this, we built on core challenges in understanding and addressing harm in such systems, developing a conceptual basis of sociotechnical frames for situating algorithmic systems within their sociotechnical context, accounting for technical, operational, organisational, and institutional factors.

Awareness of misabstraction can support the specification and implementation of sociotechnical systems by shedding light on the normative dimensions of design choices and their consequences. Misabstractions contribute not only to the production of harms, safety hazards, moral wrongs, and frictions, but also to shortcomings in harm prevention and mitigation efforts. Understanding systemic risks requires understanding systemic misabstractions. The concept of misabstraction helps us to understand how sociotechnical harms are a product of factors and dynamics that accumulate and cascade throughout a broader system, and, as a result, allows us to identify what interventions, practices, resources, and capacities are needed to anticipate, prevent, and adequately address harm in sociotechnical systems.

Acknowledgments

We thank the participants of the Netherlands Institute of Governance 2023 panel on Algorithms & Digital Government for their insightful feedback on an early version of this manuscript, and to the FAccT reviewers for their inspiring comments and suggestions. This research was conducted as part of the Gravitation research program Hybrid Intelligence, funded by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) under file number 024.004.022.

References

[1] Harold Abelson and Gerald Jay Sussman. 1984. *Structure and interpretation of computer programs.* The MIT Press.
[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.
[3] Klaus Ackermann, Joe Walsh, Adolfo De Unánue, Hareem Naveed, Andrea Navarrete Rivera, Sun-Joo Lee, Jason Bennett, Michael Defoe, Crystal Cody,

Lauren Haynes, et al. 2018. Deploying machine learning models for public policy: A framework. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 15–22.
[4] Gediminas Adomavicius and Alexander Tuzhilin. 2010. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.
[5] Philip E Agre. 1994. Surveillance and capture: Two models of privacy. *The information society* 10, 2 (1994), 101–127.
[6] Philip E Agre. 1997. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. Erlbaum.
[7] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2024. Field-building and the epistemic culture of AI safety. *First Monday* (2024).
[8] Evgeni Aizenberg, Matthew J Dennis, and Jeroen van den Hoven. 2023. Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation. *AI & society* (2023), 1–9.
[9] Ifeoma Ajunwa. 2020. An auditing imperative for automated hiring systems. *Harv. JL & Tech.* 34 (2020), 621.
[10] M Akrich. 1992. The De-Scription of Technical Objects. In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, Wiebe E. Bijker and John Law (Eds.). Cambridge, MA, The MIT Press.
[11] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.
[12] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. *Frontiers in Big Data* (2020), 5.
[13] Lori Andrews and Hannah Bucher. 2022. Automating Discrimination: AI Hiring Practices and Gender Inequality. *Cardozo L. Rev.* 44 (2022), 145.
[14] McKane Andrus, Sarah Dean, Thomas Krendl Gilbert, Nathan Lambert, and Tom Zick. 2020. AI development for the public interest: From abstraction traps to sociotechnical risks. In *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 72–79.
[15] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.
[16] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–18.
[17] Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. Nothing Comes Without Its World–Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 61–73.
[18] Agathe Balayn and Seda Gürses. 2021. Beyond Debiasing: Regulating AI and its Inequalities. https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/. Published: 2021-09-21.
[19] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.
[20] Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From design methods to systems engineering. *Interacting with computers* 23, 1 (2011), 4–17.
[21] Andrew Bell, Alexander Rich, Melisande Teng, Tin Orešković, Nuno B Bras, Lénia Mestrinho, Srdan Golubovic, Ivan Pristas, and Leid Zejnilovic. 2019. Proactive advising: a machine learning driven approach to vaccine hesitancy. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–6.
[22] Aleksander Buszydlik, Patrick Altmeyer, Cynthia CS Liem, and Roel Dobbe. 2024. Grounding and Validation of Algorithmic Recourse in Real-World Contexts: A Systematized Literature Review. (2024).
[23] Lucius EJ Bynum, Joshua R Loftus, and Julia Stoyanovich. 2024. A New Paradigm for Counterfactual Reasoning in Fairness and Recourse. *arXiv preprint arXiv:2401.13935* (2024).
[24] Brian J Chen and Jacob Metcalf. 2024. Explainer: A sociotechnical approach to AI policy. *Data & Society* (2024).
[25] Alexandra Chouldechova, Chad Atalla, Solon Barocas, A Feder Cooper, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, et al. 2024. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. *arXiv preprint arXiv:2412.01934* (2024).
[26] William J Clancey. 1993. The knowledge level reinterpreted: Modeling socio-technical systems. *International journal of intelligent systems* 8, 1 (1993), 33–49.
[27] Timothy Colburn and Gary Shute. 2007. Abstraction in computer science. *Minds and Machines* 17 (2007), 169–184.
[28] Jon Crowcroft, Ian Wakeman, Zheng Wang, and Dejan Sirovica. 1992. Is layering harmful? *IEEE Network* 6, 1 (1992), 20–24.
[29] Jenny L Davis. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. The MIT Press.
[30] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.
[31] Noortje De Boer and Nadine Raaphorst. 2023. Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Management Review* 25, 1 (2023), 42–62.
[32] Íñigo M. d. R. de Troya, Ruqian Chen, Laura O Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B Brás, and Leid Zejnilovic. 2018. Predicting, explaining, and understanding risk of long-term unemployment. In *NeurIPS Workshop on AI for Social Good*.
[33] Sidney Dekker. 2016. *Just culture: Balancing safety and accountability*. CRC Press.
[34] Jeroen Delfos, Anneke MG Zuiderwijk, Sander van Cranenburgh, Caspar G Chorus, and Roel IJ Dobbe. 2024. Integral system safety for machine learning in the public sector: An empirical account. *Government Information Quarterly* 41, 3 (2024), 101963.
[35] Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1725–1732.
[36] Roel Dobbe. 2025. AI Safety is Stuck in Technical Terms – A System Safety Response to the International AI Safety Report. doi:10.48550/arXiv.2503.04743 arXiv:2503.04743 [cs].
[37] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artificial Intelligence* 300 (2021), 103555.
[38] Roel Dobbe and Anouk Wolters. 2024. Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds and Machines* 34, 2 (2024), 1–51.
[39] Roel I. J. Dobbe. 2024. System Safety and Artificial Intelligence. In *The Oxford Handbook of AI Governance*. Oxford University Press. doi:10.1093/oxfordhb/9780197579329.013.67
[40] Paul Dourish. 2004. What we talk about when we talk about context. *Personal and ubiquitous computing* 8 (2004), 19–30.
[41] Eleanor Drage and Kerry Mackereth. 2022. Does AI debias recruitment? Race, gender, and AI's "eradication of difference". *Philosophy & technology* 35, 4 (2022), 89.
[42] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*. PMLR, 172–186.
[43] Doaa Abu Elyounes. 2020. " Computer Says No!": The Impact of Automation on the Discretionary Power of Public Officers. *Vand. J. Ent. & Tech. L.* 23 (2020), 451.
[44] European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. https://data.europa.eu/eli/reg/2016/679/oj
[45] Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. 2024. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology* (2024).
[46] João Fonseca, Andrew Bell, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. 2023. Setting the right expectations: Algorithmic recourse over time. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.
[47] Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the Use of Language Models to Guide Hiring Decisions. *arXiv preprint arXiv:2404.03086* (2024).
[48] Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT undermines its organizing principles. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1982–1992.
[49] Ben Green. 2021. Data science as political action: Grounding data science in a politics of justice. *Journal of Social Computing* 2, 3 (2021), 249–265.
[50] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.
[51] Joan Greenbaum and Morten Kyng. 1991. Introduction: situated design. In *Design at work: cooperative design of computer systems*. CRC Press, 1–24.
[52] Tricia A Griffin, Brian P Green, and Jos VM Welie. 2024. The ethical agency of AI developers. *AI and Ethics* 4 (2024), 179–188.
[53] Francisco Gutiérrez, Sven Charleer, Robin De Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. 2019. Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 60–68.
[54] Gabriel Hallevy. 2015. *Liability for crimes involving artificial intelligence systems*. Springer. doi:10.1007/978-3-319-10124-8

[55] Jennifer Helsby, Samuel Carton, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Andrea Navarrete, Klaus Ackermann, Joe Walsh, Lauren Haynes, Crystal Cody, et al. 2018. Early intervention systems: Predicting adverse interactions between police and the public. *Criminal justice policy review* 29, 2 (2018), 190–209.

[56] Gregor Hohpe. 2023. Programming without a stack trace: When abstractions become illusions. https://architectelevator.com/architecture/stacktrace-abstraction/. Published: 2023-4-14.

[57] Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. 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*. 1–12.

[58] Stefan Albert Horstmann, Samuel Domiks, Marco Gutfleisch, Mindy Tran, Yasemin Acar, Veelasha Moonsamy, and Alena Naiakshina. 2024. "Those things are written by lawyers, and programmers are reading that." Mapping the Communication Gap Between Software Developers and Privacy Experts. *Proceedings on Privacy Enhancing Technologies* (2024).

[59] Kimberly A Houser. 2019. Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.* 22 (2019), 290.

[60] Dmitry Ivanov, Alexandre Dolgui, and Boris Sokolov. 2019. Ripple effect in the supply chain: Definitions, frameworks and future research perspectives. *Handbook of Ripple Effects in the Supply Chain* (2019), 1–33.

[61] J. 2025. Ghost in the shell: problem drift. In *forthcoming*.

[62] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[63] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 353–362.

[64] Donald E Knuth. 1997. *The Art of Computer Programming, volume 1 Fundamental Algorithms*. Addison Wesley Longman Publishing Co., Inc.

[65] Joop Koppenjan and John Groenewegen. 2005. Institutional design for complex technological systems. *International Journal of Technology, Policy and Management* 5, 3 (2005), 240–257.

[66] Kristian Kreiner. 1995. In search of relevance: Project management in drifting environments. *Scandinavian Journal of Management* 11, 4 (1995), 335–346.

[67] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L Addison. 2015. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1909–1918.

[68] Nancy Leveson. 2004. A new accident model for engineering safer systems. *Safety science* 42, 4 (2004), 237–270.

[69] Nancy G Leveson. 2012. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

[70] Nancy G Leveson. 2017. Rasmussen's legacy: A paradigm change in engineering for safety. *Applied ergonomics* 59 (2017), 581–591.

[71] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations. *arXiv preprint arXiv:2406.18346* (2024).

[72] Michael Lipsky. 1980. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

[73] Steven Lukes. 1974. *Power: A Radical View*. Macmillan.

[74] Erin E Makarius, Debmalya Mukherjee, Joseph D Fox, and Alexa K Fox. 2020. Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research* 120 (2020), 262–273.

[75] Maya Malik and Momin M Malik. 2021. Critical technical awakenings. *Journal of Social Computing* 2, 4 (2021), 365–384.

[76] Gaia Marcus. 2025. Ada Lovelace Institute responds to General-Purpose AI Code of Practice draft. https://www.adalovelaceinstitute.org/news/gpai-code-of-practice/.

[77] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Extending the machine learning abstraction boundary: A Complex systems approach to incorporate societal context. *arXiv preprint arXiv:2006.09663* (2020).

[78] National Academies of Sciences, Engineering, and Medicine and others. 2022. *Resilience for compounding and cascading events*. National Academies Press: Washington, DC.

[79] Victor Alfonso Naya, Guillaume Bied, Philippe Caillou, Bruno Crépon, Christophe Gaillac, Elia Pérennes, and Michèle Sebag. 2021. Designing labor market recommender systems: the importance of job seeker preferences and competition. In *4. IDSC of IZA Workshop: Matching Workers and Jobs Online-New Developments and Opportunities for Social Science and Practice*.

[80] Jędrzej Niklas, Karolina Sztandar-Sztanderska, Katarzyna Szymielewicz, A Baczko-Dombi, and A Walkowiak. 2015. Profiling the unemployed in Poland: social and political implications of algorithmic decision making. *Fundacja Panoptykon* (2015).

[81] Douglass C. North. 1991. Institutions. *Journal of Economic Perspectives* 5, 1 (March 1991), 97–112. doi:10.1257/jep.5.1.97

[82] Sem Nouws, Íñigo Martinez De Rituerto De Troya, Roel Dobbe, and Marijn Janssen. 2023. Diagnosing and Addressing Emergent Harms in the Design Process of Public AI and Algorithmic Systems. In *Proceedings of the 24th Annual International Conference on Digital Government Research*. 679–681.

[83] Sem Nouws and Roel Dobbe. 2024. The Rule of Law for Artificial Intelligence in Public Administration: A System Safety Perspective. In *Digital Governance: Confronting the Challenges Posed by Artificial Intelligence*. Springer, 183–208.

[84] Sem Nouws, Marijn Janssen, and Roel Dobbe. 2022. Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems. In *International Conference on Electronic Government*. Springer, 307–322.

[85] Claudio Novelli, Philipp Hacker, Jessica Morley, Jarle Trondal, and Luciano Floridi. 2024. A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. *European Journal of Risk Regulation* (2024), 1–25.

[86] Selin E Nugent and Susan Scott-Parker. 2022. Recruitment AI has a Disability Problem: anticipating and mitigating unfair automated hiring decisions. In *Towards Trustworthy Artificial Intelligent Systems*. Springer, 85–96.

[87] John K Ousterhout. 2018. *A philosophy of software design*. Vol. 98. Yaknyam Press: Palo Alto, CA, USA.

[88] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[89] Rik Peeters and Arjan Widlak. 2018. The digital cage: Administrative exclusion through information architecture – The case of the Dutch civil registry's master data management system. *Government Information Quarterly* 35, 2 (2018), 175–183.

[90] Laurie Pel. 2022. Ripple Effects of Law Execution Automation in Governmental Systems: The Wajong Case. *MSc thesis, Engineering and Policy Analysis, TU Delft*. (2022).

[91] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering technology in discourse on discrimination. *Information, Communication & Society* 22, 7 (2019), 882–899.

[92] Charles Perrow. 2011. *Normal accidents: living with high risk technologies*. Princeton University press.

[93] Manish Raghavan and Pauline T Kim. 2024. Limitations of the "Four-Fifths Rule" and Statistical Parity Tests for Measuring Fairness. *Geo. L. Tech. Rev.* 8 (2024), 93.

[94] Inioluwa Deborah Raji and Roel I J Dobbe. 2020. Concrete problems in AI safety, revisited. In *Workshop on Machine Learning In Real Life at the International Conference on Learning Representations. Addis Abeba*.

[95] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[96] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.

[97] McKenzie Raub. 2018. Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Ark. L. Rev.* 71 (2018), 529.

[98] David Ribes and Geoffrey C Bowker. 2009. Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization* 19, 4 (2009), 199–217.

[99] Shalaleh Rismani, Roel Dobbe, and AJung Moon. 2024. From Silos to Systems: Process-Oriented Hazard Analysis for AI Systems. *arXiv preprint arXiv:2410.22526* (2024).

[100] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[101] Rotterdam Court of Auditors. 2024. Kleur bekennen: vervolgonderzoek naar algoritmes. https://rekenkamer.rotterdam.nl/onderzoeken/kleur-bekennen/. Published: 2024-05-13.

[102] Kevin P Scheibe and Jennifer Blackhurst. 2019. Systemic risk and the ripple effect in the supply chain. *Handbook of Ripple Effects in the Supply Chain* (2019), 85–100.

[103] Anette Scoppetta and Arthur Buckenleib. 2018. Tackling Long-Term Unemployment through Risk Profiling and Outreach – A discussion paper from the employment thematic network. *Technical Dossier no. 6* (05 2018).

[104] Kristen M Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In *Proceedings*

of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2138–2148.

[105] Nick Seaver. 2015. The nice thing about context is that everyone has it. *Media, Culture & Society* 37, 7 (2015), 1101–1109.

[106] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[107] Päivi Seppälä and Magdalena Małecka. 2024. AI and discriminative decisions in recruitment: Challenging the core assumptions. *Big Data & Society* 11, 1 (2024), 20539517241235872.

[108] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[109] Lotje E Siffels and Tamar Sharon. 2024. Where Technology Leads, the Problems Follow. Technosolutionism and the Dutch Contact Tracing App. *Philosophy & Technology* 37, 4 (2024), 125.

[110] Lucy Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

[111] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[112] Karolina Sztandar-Sztanderska and Marianna Zieleńska. 2022. When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland. *Sozialer Fortschritt* 6-7 (2022), 465–487.

[113] TenderNed. 2021. PES-suite voor Bemiddelingsservice. Reference number CAA.ICT.2021.7.016. PB/S number 2021/S 143-379823. https://www.tenderned.nl/aankondigingen/overzicht/233904. Published: 2021-07-24.

[114] Rosamunde Van Brakel. 2021. How to watch the watchers? Democratic oversight of algorithmic police surveillance in Belgium. *Surveillance & Society* 19, 2 (2021), 228–240.

[115] Kim J Vicente and Jens Rasmussen. 1992. Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics* 22, 4 (1992), 589–606.

[116] Kate Vredenburgh. 2022. Freedom at work: Understanding, alienation, and the AI-driven workplace. *Canadian Journal of Philosophy* 52, 1 (2022), 78–92.

[117] Sandra Wachter. 2024. Limitations and loopholes in the EU AI Act and AI Liability Directives: what this means for the European Union, the United States, and beyond. *Yale Journal of Law and Technology* 26, 3 (2024).

[118] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.

[119] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.

[120] Paris Will, Dario Krpan, and Grace Lordan. 2023. People versus machines: introducing the HIRE framework. *Artificial Intelligence Review* 56, 2 (2023), 1071–1100.

[121] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.

[122] Thespina J Yamanis, M Giovanna Merli, William Whipple Neely, Felicia Feng Tian, James Moody, Xiaowen Tu, and Ersheng Gao. 2013. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociological methods & research* 42, 3 (2013), 392–425.

[123] Leid Zejnilović, Susana Lavado, Íñigo Martínez de Rituerto de Troya, Samantha Sim, and Andrew Bell. 2020. Algorithmic long-term unemployment risk assessment in use: counselors' perceptions and use practices. *Global Perspectives* 1, 1 (2020).

[124] Leid Zejnilovic, Susana Lavado, Carlos Soares, Íñigo Martínez De Rituerto De Troya, Andrew Bell, and Rayid Ghani. 2021. Machine learning informed decision-making with interpreted model's outputs: A field intervention. In *Academy of Management Proceedings*, Vol. 2021. Academy of Management Briarcliff Manor, NY 10510, 15424.

A Materials: procurement tender documentation for a skills-matching system at a Public Employment Service

Table 1: Tender documents used in the analysis (source: [113]).

Document	Description
D1. Program Requirements & Wishes	General specification & requirements
D2. Secure Software Development (SSD) requirements	System security specifications
D3. Security and Processing Agreement	Processing of personal information, incl. GDPR
D4. Procedure for reporting a Security Indicent by Processor	Procedure for reporting breaches of personal information; in reference to GDPR Art. 4, 33, 34
D5. General Purchase Conditions ICT 2019	Personal data processing, maintenance requirements, and other
D6. Applicant's questions, with PES's answers	Table of questions submitted by one tender applicant, including responses by the PES