Appendix SI Raw Data Extracted from the Reviewed Articles

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| P1 | The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making (de Bruijn, Warnier, and Janssen 2022) | XAI has fundamental limitations for use in complex situations….., hence we derive a series of strategies that might contribute to more legitimacy and trust.<br>(P1.1) Strategy 1: from explaining AI to explaining decisions produced using AI;<br>We could shift our attention from explaining AI to explaining the decision supported by AI. A decision might be fully or partially based on AI, but in any case, decision-makers should be able to explain why a decision has been made. When decision-makers have this burden of proof, there might be an incentive to scrutinize the algorithms used or deviate from AI-based decision-making critically. It might make them decide not to rely on AI or on a particular type of AI exclusively.<br>(P1.2) Strategy 2: from designing algorithms to negotiated algorithms;<br>In some cases, algorithms can be more authoritative if they are not designed by experts only. Instead, an approach of co-creation with the public and interested parties can be taken. The main choices algorithms are based upon can be discussed and published. The parties involved can try to find consensus about, for example, the variables that are taken into account by an algorithm or the scope of an algorithm – what decisions an algorithm should make and should not make and how humans should remain in control. A process like this results in *negotiated algorithms* in which every stakeholder has its say, and a consensus needs to be reached<br>(P1.3) Strategy 3: from explainable algorithms to explainable processes;<br>Closely related to the idea that algorithms can be explained is that the design process also can be explained. Transparency then refers to questions like who will be involved, who will have what role, what are the main issues that will be debated, how will parties deal with dissensus and uncertainties, how will they make their decisions. Not all discussions need to be documented in detail, but only the relevant processes that lead to decisions and the argumentation why decisions were taken. The attention shifts from making algorithms explainable to making the process of creating algorithms explainable.<br>(P1.4) Strategy 4: from an instrumental to an institutional approach;<br>Institution-building can comprise setting up organizational structures that facilitate the development of |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | these rules of the game. Examples are the establishment of regulators with authority to scrutinize and audit algorithms and to develop regulation or, within organizations, review committees that are positioned as countervailing powers of developers and users of algorithms. The higher the degree of institutionalization, the easier it is to design an institutional structure that can be used as a countervailing power in the design and use of algorithms. A continuous critical look from this countervailing power can be conducive to the right use of algorithms.<br>(P1.5) Strategy 5: from monopolistic algorithms and datasets to competing algorithms and datasets;<br>Using a metaphor from the world of economics, organizations often employ monopolistic algorithms and monopolistic datasets to a lesser extent. They develop one algorithm or one family of algorithms and use these algorithms to base their decisions upon. The transparency of AI-based decision-making can be enhanced by deliberately using competing algorithms and datasets. Only if competing algorithms that are trained on independently collected datasets result in more or less the same decision, it might be reasonable to assume that this is a correct decision. If competing algorithms provide different decisions, a human decision-maker should take over.<br>(P1.6) Strategy 6: from algorithms to value-sensitive algorithms;<br>AI-based decision-making can reinforce deeply rooted biases, and therefore result in morally wrong decisions. When designing algorithms, the parties involved can take certain key values into account. One should aim to design the algorithm in such a way that data that might result in biases or discrimination (e.g., age, gender, race) is ignored and verify whether these undesired variables have an impact on the proposed decisions of the algorithm (Du et al., 2019). Furthermore, tests can be conducted if humans are treated in the same manner. This 'value sensitive' design (Friedman, 1996) of algorithms incentivizes the parties involved to be transparent about what values they want to safeguard and how these values are guaranteed.<br>(P1.7) Strategy 7: from algorithms replacing professional decision-making to professionals challenging algorithmic decision-making.<br>There is a classic tension between analytical decision-making based on facts and figures, and intuitive decision-making of professionals based on their tacit knowledge. Both types of decision-making have their strengths and weaknesses. There is the risk that with the emergence of AI, intuitive decision-making will be replaced by predominantly analytical decision-making. Also, if professionals are |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | replaced, then their tacit knowledge will be lost. They often have deep insight into the nature of societal problems and what should be taken into account. It might be a strategy to make AI-based decisions *and* ask professionals to make decisions based on their tacit knowledge. |
| P2 | Accountable Artificial Intelligence: Holding Algorithms to Account (Busuioc 2021) | (P2.1) Unless models are rendered interpretable and comprehensible by system designers to facilitate user understanding, their inscrutability will invariably entail that such scores are taken at face value;<br>(P2.2) Model transparency additionally necessitates concerted efforts of "system architects" to explain their models, of the computer science community more broadly to develop models that are understandable and interpretable in the first place, of the industry to systematically adopt such practices, and of public sector purchasers and regulators to demand them;<br>(P2.3a) It will also entail critical and ongoing cooperation between system designers and domain experts, and one that spans from the early stages of system design to real-world implementation (production) and monitoring of system functioning,<br>(P2.3b) it also underscores the responsibility of developers, in collaboration with domain experts, to specifically test and interrogate not only technical but also governance implications (broader considerations of fairness, bias, and transparency) at the model testing and validation phases<br>(P2.4a) This responsibility extends to ensuring that such systems are properly and independently vetted, that their functioning is continuously monitored, and that<br>(P2.4b) public sector staff are adequately trained to understand the tools they are to rely on,<br>(P2.4c) as well as that domain experts and affected citizens are allowed to diagnose and challenge unanticipated failures during use;<br>(P2.5) Public sector use of AI tools—where the stakes can be the likes of liberty deprivation, use of force, and welfare or healthcare denial—requires explanation of the *rationale* of individual decision-making;<br>(P2.6) Developers are increasingly coming to the realization that the continued reliance on—and proliferation of—such systems will come down to user trust in their outputs, with algorithm |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | understandability as an important ingredient thereof. For algorithmic models that are to be deployed in high-stakes contexts, it is recommended that models are a priori well understood, that their predictive features are understandable (importantly also to users and domain experts);<br>(P2.7) *From implicit to explicit value trade-offs*. Model explanation or justification necessarily also extends to an oft-forgotten aspect of model functioning: value trade-offs inherent in model design. While AI algorithms are often presented as "neutral devices," algorithmic systems necessarily encode important value trade-offs (e.g., recall versus accuracy, precision versus fairness). Deciding how to strike the balance among these is necessarily a *political* and not a purely technical act. This necessarily again highlights that prior deliberation—of the standards that are to be prioritized and inform specific models in different sectors—are crucial steps to adequate model design;<br>(P2.8) Regulatory efforts are thus vitally needed to ensure that AI tools are brought to bear in a thoughtful and effective manner. The wide adoption of a variety of regulatory tools, for the most part currently lacking, and being proposed by academic and some industry actors, such as: **model certification, third-party independent ongoing testing of model performance, the use of models that provide audit trails** of their decision-making, **algorithmic impact assessments, pre-public sector deployment**, etc. would further help to bolster algorithmic accountability. |
| P3 | In AI we trust? Citizen perceptions of AI in government decision making (Ingrams, Kaufmann, and Jacobs 2022) | (P3.1) We argue that researchers and practitioners give more attention to the balance of instrumental and value-based qualities in the design and implementation of AI applications<br>(P3.2) On the matter of our null finding regarding the interaction effect of task complexity on the relationship between AI decision making and the outcomes of red tape and trust, we followed a small body of empirical research that has shown how value-based concerns about AI are attenuated when the tasks become more complex. Theoretically, the potential trust problem here in the eyes of citizen users of public services is that a computer should not be trusted to do something that is complex in terms of making value-based decisions about individual cases. It is difficult to explain in retrospect why our reasoning did not hold true. The perceived problems of complex tasks being administered by AI may be more powerfully triggered by other types of value-based concerns under conditions that raise clearer ethical questions. Our tax auditing scenario was based on a real-world case that has raised some ethical concerns about accountability, but other cases have been criticized more strongly for privacy invasions and discrimination (Angwin et al., 2016; Lavertu, 2016). It may be in these more ethically salient cases that the matter of complexity becomes important. Citizens are likely to have a more positive disposition towards the use of AI in the context of simple tasks because the risky character of complex tasks would appear to really need human intelligence to manage appropriately (Nagtegaal, 2020). According to |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | Jansson and Erlingsson (2014), citizens often question the legitimacy of AI being used in complex situations. Similarly, simple tasks performed by an AI will most likely not trigger the aforementioned cognitive shortcuts that lower citizen trust. But, in contrast, when the AI is tasked to do something more complex, such as deciding how many police officers to assign to a neighborhood for its safety (Meijer & Wessels, 2019), citizens are likely going to be even more skeptical in their assessment of trustworthiness.<br><br>(P3.3) The effect of the whether the decision maker is a human or AI on trust is significant across all dimensions of trust, which shows that a human decision maker is perceived as more trustworthy for carrying out of such a decision in terms of competence, honesty, and benevolence.<br>(P3.4) Research has shown that design of AI applications in government happens at several organizational levels with both technical and managerial staff (Kitchin, 2014). It will be necessary for the designers at those levels to develop better socio-technical understanding of their applications—including how perceptions of citizens are affected—to integrate those considerations into the design process rather than leaving it as an afterthought to administrators who are responsible for their implementation and monitoring (Kattel et al., 2019).<br>(P3.5) We recommend further research into citizen perceptions of AI decision making in government to understand these variables and provide empirical evidence that policymakers can use for their decisions about the degrees of automation and human accountability that should be integrated into future AI applications |
| P4 | AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings (Kuziemski and Misuraca 2020) | *(P4.1) Both Canadian and Polish cases have underlined the strong role civil society and academia can play in scrutinizing automated decision making systems – both at the stage of goal-setting, procurement and implementation. At the same time, it is becoming evident that the role of the state in AI policymaking is not to be downplayed – even if it takes a form of governing through adopting technological solutions at the center of its operations, and not writing laws;*<br><br>(P4.2) That being said, current AI policy debate is heavily skewed towards voluntary standards and self-governance, somehow disregarding power-related considerations. Introduction of digital technologies in general, and in the public sector in particular is often portrayed as beneficial to the end users. Yet, are the processes happening under the banners of 'democratization', 'convenience' and 'choice' serving its advertised purposes? Or are these disguised attempts to strengthen the grip of control over the citizens? In other words - is AI facilitating the power shift between the public sector and citizens or merely intensifying existing distribution? Is the use of AI in the processes of governance changing the way power is exercised? Whatever the conclusions, these issues are not neutral to the general public; |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P4.3) As governments, municipalities and public agencies around the world resort to automation in as diverse sectors as healthcare, law enforcement, and social services - sometimes with suboptimal, or downright unfair results - it is key to consider desired directions of the development of the field, and scrutinize existing algorithmic practices. What goals should public sector organisations pursue when commissioning automated decision systems? Whose benefits should be prioritized? Finally, it is important to ask ourselves: "For what are we optimising?" (Bavitz & Hessekiel, 2018). No proper guidance for the public sector use of automated decision systems can fail to imagine the states of the world it envisions, and the values that it wants to support;<br><br>(P4.4) As a result, a guideline for the *Responsible use of AI* has been created. Its guiding principles include the government's commitment to (Government of Canada, 2018):<br>(P4.4a) understanding and measuring the impact of using AI by developing and sharing tools and approaches;<br>(P4.4b) being transparent about how and when using AI, starting with a clear user need and public benefit analysis;<br>(P4.4c) providing meaningful explanations about AI decision-making,<br>(P4.4d) while also offering opportunities to review results and challenge these decisions;<br><br>(P4.5a) This implies being as open as possible by sharing source code, training data, and other relevant information,<br><br>(P4.5b) all while protecting personal information, system integration, and national security and defence,<br><br>(P4.5c) as well as providing sufficient training so that government employees developing and using AI solutions have the responsible design, function, and implementation skills needed to make AI-based public services better |
| P5 | Cultivating Trustworthy Artificial Intelligence in Digital Government (Harrison and Luna-Reyes 2022) | *(P5.1) The more information available to AI, the more systems can learn and improve accuracy (Marr, 2017). On the other hand, without AI, Big Data loses some of its potential for innovation. However, problematic data introduce risks that can lead to economic devastation for industry and that may erode trust in and the legitimacy of government. Thus, trustworthy AI rests on the quality of this fuel. Although much attention is now devoted to improving data analytics, much less attention is devoted to improving data management, which is essential to data analysis of any kind, but critical to trustworthiness within the AI context. Unfortunately, providing a sound foundation for building AI systems is no easy accomplishment, since government agencies do not typically possess appropriately curated data resources (Mehr, 2017). Moreover, critics remind us* |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | *that Big Data is not necessarily better data and that data taken from its original contexts threatens to lose its original Meaning (Boyd & Crawford, 2011);*<br>*(P5.2) Government agencies are required by statute and regulation to safeguard access to personally identifiable data, although as a recent data breach at the Office of Personnel Management illustrates, they do not always measure up to these challenges (Marks, 2018). Privacy is even more relevant recently because the greatest value obtained from AI comes from integrating volumes of data from multiple sources that, at the same time, can readily generate trails of potentially identifiable data (Johnson, 2017; Kitchin, 2016). When joined and analyzed together, integrated data sets provide the potential for "the ability to assemble multiple views of the customer [that] may provide inappropriate insights" (O'Leary, 2014, p. 72). Such integration is what Big Data and AI make possible and valuable. There is an inherent tension here because "the utility and privacy of data are intrinsically connected, no regulation can increase data privacy without also decreasing data utility" (Ohm, 2010)*<br>(P5.3) We propose two recommendations related to data used in AI:<br><br>(P5.3a) First, government needs to implement enterprise data management as an essential foundation for trustworthy AI. Data management is relevant at two points in the system. At the operational level, data management improves the general fidelity of the information systems and ensures the existence of metadata documenting data's "origin, format, lineage, and how it is organized, classified, and connected;" metadata are vital for determining the uses to which the data can be applied (Brunet, 2018). In the AI Pre-processing stage, data management enables reliability of data acquisition and replicability of the analysis, both of which are essential to preserving the transparency of any analysis;<br><br>(P5.3b) Second, governments seeking to deploy AI should cultivate data literacy. In the new digital government, the integrity, security, and appropriateness of data is a prerequisite for trustworthy decision making. Data literacy must be seen as an imperative for any government employee whose responsibilities bear on data collection, manipulation, and use. In addition, literacy for AI practitioners requires careful documentation of judgment calls in the processes of data cleansing and preparation as well as continuous reflection on the need to understand limitations of the data set and what questions can be asked of it (Boyd & Crawford, 2011). Such practices enable policy and decision makers to assess and critique the grounds upon which AI development takes place, enabling transparency and providing a basis for accountability;<br><br>(P5.4) To address the problem of algorithm-related privacy problems, the developing field of machine ethics (Allen et al., 2006) has undertaken initial steps using the concept of differential privacy (Han et |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | al., 2017; Kearns & Roth, 2019). The approach involves the use of algorithms that produce randomized data following known probability distributions. Given that the algorithm developer knows the way in which data are scrambled, it is possible to extract main patterns and models while protecting individual's privacy;<br>(P5.5) In the area of fair models (algorithms), some initial steps are also being taken. Although common approaches to fairness conceptually involve the idea of being "blind" to individual characteristics (gender, race, and religion) to avoid discrimination on the basis of such characteristics, current efforts in producing fair models show that it is necessary to adopt an open and transparent definition of the populations that are to be protected as well as definitions of accuracy of the model (Kearns & Roth, 2019);<br>*(P5.6) In the meantime, the problems with issues of algorithmic opacity will continue to jeopardize government's traditional values for transparency, explainability, and accountability. These conditions argue for conservative use of AI support in decisions with consequences for individual citizen welfare (Mehr, 2017). Instead, substantial human oversight and domain expertise would seem to be minimal requirements for AI system development;*<br>(P5.7) The goal should be an accumulated track record of successful AI development in low-risk applications that provide test beds for experimentation, along with other strategies discussed below. As Burrell (2016) suggests, experimenting with simplified forms of machine learning enables "feature extraction," which identifies features critical to classification may provide an initial strategy for experimentation;<br>(P5.8) Although AI decision making has been celebrated for its ostensible ability to operate without human intervention (Henman, 2019), we view this as an untrustworthy strategy for government decision making. A more useful approach may be to view AI-driven systems as composed of computational components, human actors, and institutional arrangements that guide the use of AI (Johnson & Verdicchio, 2017) through the creation of AI governance structures. Such an approach would highlight questions relevant to trustworthiness such as: Who decides which AI systems are designed? Who makes decisions about their design and implementation? Which tasks are delegated to machines and which to humans? How are humans who work with AI systems trained? and How can algorithmic decisions be appealed or challenged within legal and regulatory processes? AI governance structures must be populated by the widest range of stakeholders including policy makers, government domain experts, and AI systems developers inside and external to government, along with individuals who will be affected by AI decision making (Dhasarathy et al., 2019; Gasser & Almeida, 2017). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P5.9) We call attention to three further strategies likely to improve trustworthy AI development:<br>(P5.9a) algorithmic accountability – Algorithms will only be sufficiently transparent if government creates and maintains records that document their objectives for algorithms and vendors disclose sufficient information describing how algorithms are developed, making it possible to trace the decisions taken in the construction of AI algorithms (Brauneis & Goodman, 2018);<br>(P5.9b) algorithmic audits - Algorithmic audits using a variety of research and statistical methods can be used to assess the extent and type of harmful bias that may inhere in the algorithms used by publicly available online service providers such as Google and Netflix (Sandvig et al., 2014); public managers should insist on analogous assessments related to government AI products. A more pointed approach consists of algorithmic impact assessments, such as those recommended by AI Now (Reisman et al., 2018); and<br>(P5.9c) Participatory AI development and testing - It is essential that a wide variety of AI stakeholders be involved in the processes of participatory AI development and testing. Such involvement should take place in contexts in which AI system designs, functions, and operations are "contestable" (Mulligan et al., 2020). Possible scenarios include interaction between software developers and domain experts at the point of development or interactions between experts and users with systems already developed, empowering domain experts to use their knowledge about training data and decision rules to shape how system decision making takes place and to play a governance and oversight role. It should also be possible to challenge particular outcomes through the use of feedback from users and others affected by system outcomes (Dhasarathy et al., 2019). |
| P6 | Governmental transparency in the era of artificial intelligence (van Engers and de Vries 2019) | (P6.1) Therefore, it is of great importance that those (partly) automated decision systems are transparent on their reasoning mechanisms and carefully explain their decisions. In order to achieve a better understanding of the citizen, a digital letter should be compatible with existing knowledge of the citizen; the parts of the letter have to fit together and use as few causes possible; and the letter should be written clearly, provide contrastive information and offer the opportunity to interact.<br>The major requirement proposed by the paper is *explanation of AI decisions*, which authors argued, based on empirical data can be enhanced, by making sure it satisfies the 6 primary quality criteria highlighted in the conceptual framework: (i) external coherence; (ii) internal coherence; (iii) simplicity; (iv) articulation; (v); contrastiveness; (vi) interaction. Therefore, governments can increase the acceptance rate of citizens by improving the clarity of their explanations, and this can create a new field |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | of interest in *explanation optimisation*.<br>(P6.2) In order to try to protect some essential social fundamental values, The Dutch Council of State (advisory body to the government) published a report on the influence of new technologies on constitutional relations [22]. The Council advises the government to pay closer attention to the motivation of their automated decisions. They demand that it should be clear which decision-rules (algorithms) and data the governmental authority used for a specific decision. Furthermore, it should be made clear which data is taken from other governmental authorities. Citizens could be offered more insights into the specific components of the decision made, the calculations applied and sources of law that contain the rules underlying the decision-making process. |
| P7 | Beyond State v. Loomis: Artificial intelligence, government algorithmization and accountability (Liu, Lin, and Chen 2019) | (P7.1) Risk assessment tools (like COMPAS) must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations. In contrast to the Loomis decision, the Supreme Court of Western Australia underscored the fact that the tools in dispute 'were not devised for and do not necessarily take account of the social circumstances of indigenous Australians in remote communities,' thus concluding to have 'grave reservations as to whether a person of the respondent's background can be easily fitted within the categories of appraisal presently allowed for by the assessment tools';<br>(P7.2) The crux of the present case is, in our view, how to ensure that the defendant has the means to challenge information before the court, namely, the COMPAS risk score. The defendant's right to interrogate the algorithm, therefore, is crucial, especially against the backdrop of potential discrimination hidden in the data<br>(P7.3) Non-transparency is yet another flaw. While COMPAS algorithms drew on public data and information provided by Loomis, it did not explain the breakdown of each variable, relevant weighting, and their correlation. The defendant did not know how the algorithm was designed or how the decision was made; nor could he challenge it;<br>(P7.4) The black box problem should be further disentangled into: legal black box and technical black box.<br>(P7.4a) In terms of the 'legal black box' — i.e. opacity arising from the propriety characteristics of statistical models or source codes, which are legally protected by relevant trade secret statutes. One way to address this is to unpack the legal black box to the public upon specific conditions to secure a certain level of transparency and accountability. As Public disclosure would likely be objected to by companies keen to protect the secrecy, an alternative would be disclosure only to interested parties or to an expert |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | committee in a confidential manner. For instance, the relevant laws could be amended to compel private firms earning profits while *performing essential public services to* disclose their algorithmic processes to court-approved parties or expert committees or at least to limit trade secret protection in such circumstances. To directly tackle the legal black box, free information laws must change to require disclosure by private vendors when their software is used for governance<br>(P7.4b) More problematic is what we called a '*technical black box,*' which occurs when AI techniques including machine learning and deep learning are involved. This '*technical black box*' problem may significantly frustrate governance efforts to foster transparency and accountability in the government's use of AI-based systems. A potential solution is to consider certain opt-in or opt-out mechanisms, allowing the defendant or those subject to computational decision-making, a certain degree of autonomy regarding the use of such technology in his or her case. The subject should be fully informed of the potential risks as well as the benefits of the use of the software. He or she should also be informed of the limits on understanding how the software makes projections the way it does.<br>(P7.4c) a useful starting point is to rethink the question: who is the decision maker? In a world that often blindly portrays numbers to be scientific, neutral and objective, human decision-makers are likely to surrender their powers to data. As seen in the *Loomis* decision, ill-informed deference to the privately made machines marginalizes the role of public authority and public scrutiny in governance. Conceivably, one way to redress this may be to draw a line between nondiscretionary and discretionary decision-making process in relation to public officials' use of automated machines. Discretionary processes must require human intervention, although the fashion and level of such intervention is contingent upon factors like the nature of the decision, issue area, interests involved, available remedy, and resource distribution, etc. An automated system, moreover, 'must be designed in a way that accurately reflects the government policy it models and agencies should be careful that the system does not fetter the decision-maker in exercising any discretion';<br>(P7.5) A more thorough solution to ensure the ethics of algorithm designers and to hold them accountable involve lengthy multi-stakeholder deliberation. Searching for a technologically-informed and socially-apt governance model in the age of big data can only be fruitful by engaging multi-stakeholders through constructive debate and dialogue. |
| P8 | How can we open the black box of public administration? Transparency and accountability in the use of algorithms (Martínez | (P8.1) Ultimately, we must remember that the use of algorithms in public decision-making changes how decisions are made and, therefore, the motives behind any administrative act. Accordingly, public administrations will also have to explain how they have reached a certain decision, what the process was, what data were taken into account and what the objectives behind it were. However, there is not yet |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | 2019) | any explicit obligation in this regard. For this reason, we should move towards defining a government obligation to explain the algorithms they use, their purposes and operation, as well as the data used so that public employees are aware of all these factors, and so that they can, ultimately, be monitored by citizens. Meanwhile, we should recognise the right to receive an explanation of how algorithms are used by the public administration;<br>(P8.2) Access to algorithms is necessary, but not sufficient to ensure transparency. Access to the source code of the algorithm does not necessarily make it more transparent. It is also essential to understand how an algorithm works. To ensure that algorithms can be understood, it would be useful for public administrations to provide a description of them in natural language. In this regard, we must remember that transparency legislation refers to the fact that public administrations should share information in a way that can be clearly understood by the public. Public information must also be intelligible.<br>(P8.3)To facilitate the understanding of the information, the recipients will also need to have the skills and the time to be able to understand and analyse the algorithms;<br>(P8.4) Prior to the implementation of algorithms and their use by public administrations, the organisation would monitor them to try to avoid errors or biases and assess the impact they have on fundamental rights such as personal data protection. Once the algorithms are already in use, they would be monitored to try to assess their impact and effectiveness as well as to check the legality of any decisions taken by the government through the algorithms;<br>(P8.5) It should be remembered that access to information can be limited when it may cause harm to various assets and rights identified by law, such as national security and public safety, commercial interests, intellectual and industrial property and confidentiality or secrecy in decision-making processes. It should also be borne in mind that in cases where algorithms have been acquired by third parties, the legislation regulating public sector contracts may be applicable and, therefore, the contractor may declare it confidential. In any case, an assessment should be made of the damage that access to the source code or algorithm may entail or in relation to information that may have been generated on its design and operation |
| P9 | Algorithmic decision-making and system destructiveness: A case of automatic debt recovery (Rinta-Kahila et al. 2022) | (P9.1) The first instance of organisational limits that we observed in the timeline of our case study was top management's limited vision that made them unable or unwilling to foresee and critically evaluate the ADM programme's impact on providing social welfare services. Limited managerial vision was characterised by two aspects of the government's approach: welfare-critical ideology and economic imperative (see the data structure shown in Figure 2). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | *(P9.1a) First, the top management’s welfare-critical ideology exhibited an overall critical view towards welfare services. This attitude was evident in significant resource cuts that DHS, as the main welfare agency, experienced during the years preceding OCI: “Over the past five years, the professional and technical capacity of the department has been severely eroded. There has been a significant reduction of permanent staff . . . ” (CPSU, 2017, p. 14). The ideology viewed welfare non-compliance (framed by the government as “welfare fraud”) as a significant societal problem, Minister Alan Tudge stating: “we’ll find you, we’ll track you down and you will have to repay those debts and you may end up in prison” (ACOSS, 2017, p. 4)*<br>*(P9.1b) Second, the ADM programme was driven by an economic imperative: an algorithm was implemented to automatically calculate debts and dispatch debt notices to citizens primarily “in a bid to save money” (ABC News, 5/4/2017). This imperative resulted in tunnel vision (i.e., “the tendency to focus exclusively on a single or limited goal or point of view”).*<br>(P9.2) Top management’s limited vision initiated an ADM programme that can be characterised as having limited human agency, a limited ADM solution, and lack of best practices.<br><br>(P9.2a) the new, redesigned process shifted responsibilities for making decisions and performing work tasks from humans to an ADM artefact, resulting in a work system with notably limited human agency. The change was aligned with the tunnel-vision perspective described above: work was delegated to the cheapest labour as algorithms became responsible for calculating potential overpayments and citizens became responsible for validating the algorithm’s calculations. Human agency was limited in three ways: minimising human oversight, reversing the onus of proof, and requiring citizens to self-service.<br><br>*(P9.2b) Minimising human oversight entailed full automation of debt-collection processes and put the machine at the centre of the previously human-centred and largely manual process. Even before OCI, human case workers had leveraged a data-matching system to identify potential debtors, but they had but they had also “manually checked [the data-matching system’s] information for accuracy and contacted the recipient and/or their employer to clarify the information” (Senate Committee, 2017, p 15). The automated system independently estimated welfare overpayments and sent debt notification letters to citizens without human scrutiny. There were no longer any checks of accuracy with the recipient or employer.*<br><br>*(P9.2c) Reversing the onus of proof of debts handed the task of verifying “whether or not a purported debt exists” from Centrelink to citizens (Senate Committee, 2017, p. 19). Requiring citizens to self-service represented another means of pursuing cost savings, by limiting DHS’* |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | *employees' interaction with citizens. Staff were instructed to redirect citizens to the online self-service portal in any debt-notification related matters even if they would have been able to help the citizen over the phone.*<br>*(P9.2d) In sum, managerial limits in process design removed the mindful human involvement that had previously characterised the complex process of identifying and raising debts, and transformed the role of the ADM artefact from a humans' decision*-support tool to being the main *decision-maker.*<br><br>(P9.3) The ADM programme produced decision outputs that were riddled with errors and had a biased effect on vulnerable cohorts. Limits imposed by incompatible data sources and the decision-making algorithm's inability to reconcile them was especially harmful for populations with a volatile income and numerous previous employers<br><br>*(P9.4) The data that was provided to OCI's algorithm as a basis for its decision*-making was *inconsistent mainly for two reasons. First, the two data sources, Centrelink and ATO, recorded data in different and incompatible formats (fortnightly vs. yearly). Moreover, the name of the same employer was sometimes recorded in a different way between the two databases, e.g., if a customer had made a spelling mistake when declaring their income to either party. These inconsistencies critically limited the extent to which the raw data could serve as a basis for conclusions about a given customer's welfare debt as "the information to enable an accurate debt assessment to be made" was insufficient (Senate Inquiry, 2017, p. 42*<br><br>(P9.5) The algorithm that estimated potential overpayments used simple averaging to match the two data sources and could not account for citizens' unique work-history circumstances. Citizens on welfare support payments often have fluctuating work hours and salaries due to casualised work, which the algorithm was not able to consider and would thus return inaccurate estimates. As such, "the discrepancy letters issued and the subsequent debts raised [were] a form of speculation" (Victoria Legal Aid, 2017, p. 7). Such a simplistic logic limited the ADM artefact's ability to represent the reality of ones' debt situation;<br><br>(P9.6a) Finally, OCI's citizen interface, materialised physically in the debt letters and digitally in the myGov portal, was limited in its ability to explain the basis of the debt calculations and the overall debt-recollection process. The Ombudsman's report (2017) stated, "[t]he letter did not include the 1800 telephone number for the compliance helpline. It did not explain that a person could ask for an extension of time or be assisted by a compliance officer if they had problems" (p. 9). Further, it noted that "the |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | OCI system does not clearly state it uses the averaging method or explain this may be inaccurate in some cases" (p. 12).<br>(P9.6b) In addition, the online portal was criticised for being complex and hard to use, further "black-boxing" the workings of the system<br>(P9.7) Finally, top management's limited vision resulted in the design of the ADM programme being plagued by a **lack of best practices**. Being predominantly informed by a welfare-critical ideology and an economic imperative, "the strategy was rolled out due to government pressure even when concerns with the process were being expressed" (CPSU, 2017, p. 4). Centrelink employees who raised red flags before OCI was implemented had their warnings dismissed by the department management: "Many members stated that concerns were raised during the design process but were simply ignored . . . " (CPSU, 2017, p. 13). Centrelink did not involve relevant stakeholders, such as DTA, ATO, legal experts, and domain specialists, in the development of OCI.<br>(P9.8) As such, little testing or piloting was conducted when implementing the system: "We asked DHS whether it had done modelling on how many debts were likely to be over-calculated as opposed to undercalculated. DHS advised no such modelling was done". (Ombudsman, 2017, p. 8).<br>(P9.9a) The distress was aggravated by the system's poor interface, which made understanding and contesting debts especially hard for people who have lower access to technologies like online portals.<br>(P9.9b) The difficulty of challenging debts and the short time given to respond meant that many citizens "who did not believe they owed a debt . . . paid it because some found it too difficult or too stressful to challenge the purported debt . . . " (Senate Committee, 2017, p. 4).<br>(P9.10a) Inability to help citizens made things worse. Some staff had received no information or training prior to OCI's go-live: "One Customer Service Officer stated that 'I had no idea about this initiative until I heard about it on TV and complaints started coming in from customers." (CPSU, 2017, p. 15).<br>(P9.11) This process was slow initially due to a lack of legal frameworks and structures to support citizens in challenging automated decisions (i.e., environmental limits) and given the novel and unprecedented nature of these issues. Governance activation refers to a constellation of governance and legal mechanisms that were mobilised in response to the negative effects of OCI, culminating in a formal court ruling that forced the system's decommissioning. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | Due to the **lack of legal support infrastructures** and oversight mechanisms in the broader environment, there were insufficient checks and balances to prevent the continued use of the destructive ADM programme. Many citizens who wanted to challenge or contest DHS decisions received little help to do so, with Centrelink making no legal support available to citizens and with community legal centres lacking capacity: the increased demand for legal services "placed extraordinary pressure on Community Legal Centres and Legal Aid Commissions" (ACOSS, 2017, p. 12).<br>(P9.12) **Limiting sociotechnical agency** is specific to the context of ADM in that it assumes that decision-making tasks can be delegated to an IT system with some degree of agency (Baird & Maruping, 2021). The mechanism suggests that a sociotechnical (human-machine) configuration (Grønsund & Aanestad, 2020) that minimises human agency and maximises algorithmic agency can create unintended negative outcomes at scale if the ADM system is unable to satisfy the principle of requisite variety in its context. This is comparable to Drummond's (2008) case where an IT system's inability to handle complexity caused destructive effects and raised questions about "how to combine technical and human capability" (p. 183) to avoid such outcomes. Robodebt's issues with requisite variety were captured in more granular terms by reflecting on the ADM artefact's limits—simplistic algorithm and inconsistent data— and their specific destructive implications in the broader system in which the sociotechnical sub-system operated. The mechanism identified in our case reveals how an ADM system produces discrimination at scale when humans are removed from the loop, creating unforeseen havoc that previous anecdotal literature has warned about (Boyd & Crawford, 2012; Favaretto et al., 2019; Newell & Marabelli, 2015).<br>(P9.13) This situation highlights the importance of ensuring effective governance and oversight functions are operating within the deploying organisation to compensate for this lack of external governance and legal oversight. Deploying organisations can proactively adopt best-practice frameworks for ensuring ethical use of ADM, including ensuring appropriate human oversight, ensuring the identification, mitigation and monitoring of risks to stakeholders before and during deployment (e.g., impact assessments, ethical review boards), checking the accuracy and robustness of algorithmic outcomes prior to deployment, and ensuring effective independent investigation processes in response to complaints (Gillespie et al., 2020).<br>(P9.14) Our study warns managers against considering ADM as a silver bullet that will yield automatic benefits when implemented in a sociotechnical system. The need for robust governance frameworks around ADM indicates that to do this successfully, governments need to develop sufficient capacity and competency to run ADM programmes in a responsible manner. These considerations suggest that managers should have sufficient domain sensitivity and willingness to invest not only in technical |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | capacity but also in human resources |
| P10 | Governing Ethical AI Transformation: A Case Study of AuroraAI (Leikas et al. 2022) | (P10.1) The sustainable development and deployment of artificial intelligence (AI) in the public sector requires dialogue and deliberation between developers, decision makers, deployers, end users, and the public.<br>(P10.2) According to this list of challenges it seems that the most demanding goal is to understand the system and to gain the knowledge to control its performance.<br>(P10.3) This problem is further complicated by the increasingly abstract nature of the digital world where there are even more layers of complexity that can be used to hide what is going on. In addition, AI introduces another layer of complexity as even many experts lack a clear understanding of the workings of many systems that are in use;<br>(P10.4) In addition, as AI evolves, we need a new kind of ethical reflection. This means constant ethical self-examination and vigilance alongside AI development. Ethical experts, scientists, technology developers, and other relevant stakeholders need to be brought together to deliberate the ethics of AI in a multidisciplinary way<br>(P10.5) In other words, the penalty for causing harm can be very high in the public sector. Consequently, the use of AI in the public sector needs to be transparent to the extent possible, in effect to gain citizens' trust (Bryson and Winfield, 2017), and to comply with the need for regular scrutiny and oversight (Desouza et al., 2020, p. 206).<br>(P10.6) The need for increased public engagement in the deployment and even in the design and development of AI services has been well-recognized by a range of organizations. Engagement with the public and raising public awareness serves two major functions. First, a transparent debate builds trust by involving the public, and second, better outcomes for design and implementation can be reached through public engagement. A sustained debate is important to influence policy decision-making and to ensure a more democratic and trustworthy process from the public's perspective.<br>(P10.7) Data openness and issues of privacy are important but not always easy to implement given the expertise that is needed to put them into practice.  In terms of data and privacy issues, the challenges faced by the public are well-documented and range from the unethical use of data (Gupta, 2019), lack of data privacy (Valle-Cruz et al., 2019), to challenges with data security (Toll et al., 2019). Public users are worried about novel challenges to the privacy of data in AI systems for governments (Fatima et al., |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | 2020) and how to curtail privacy violations (Kuziemski and Misuraca, 2020).<br>(P10.8) The challenge to creating AI-driven services for the public come largely from what Zuiderwijk et al. (2021) refer to as the skills challenge. In their review, they document different aspects of this problem including limited knowledge about machine learning and AI among the staff (Ojo et al., 2019). Differential skills levels of people in the organization based on their function and background, inhibits cross-sectoral collaboration around AI (Mikhaylov et al., 2018). Researchers have also documented a lack of in-house AI talent (Gupta, 2019; Sun and Medaglia, 2019) coupled with gaps in education for highly technical skills (Montoya and Rivas, 2019). Overall, a lack of expertise (Al-Mushayt, 2019) coupled with increased demand for a limited number of AI experts<br>(P10.9) The Ethics Board decided to follow a **forward-looking, proactive ethical deliberation process**, which includes participatory ethical design and aims not only at identifying problems but also at finding ethically sustainable solutions for implementation (Sengers et al., 2005; Stahl et al., 2010). The key elements of the process are: (i) Anticipation: Proactive ethical thinking in the development of design solutions; looking carefully at both the objectives and the potential unintended consequences of deploying an application or a service; (ii) Involvement: Involvement of technology users or user representatives and developers in identifying and discussing ethical challenges in a specified context; and (iii) Expertise: Involvement of experts of ethics, technology, social and behavioral sciences, and law in the discussions.<br>(P10.10) One of the main strategic questions was whether the focus should be on macroeconomic (a better national economy through improved welfare and individually empowered citizens) or highly personalized (AI helping users to achieve their own life goals). Therefore, in order to address the ethical issues effectively enough, the Ethics Board decided to take a systematic approach to the debate. The focus of the work was set on the non-technical dimensions of the AuroraAI program. This included: Background assumptions, value base, and social vision; Human and social impact; Power structures; The content of governance transformation; The meaning and interpretation of human-centricity; Interpretations of foresight and a foresighted society; and General application of AI to socio-economic and political issues<br>(P10.11) As the program includes all short, medium and long-term objectives, it has been difficult to understand the overall picture. One of the key strategic questions has been whether the AuroraAI program's emphasis is macroeconomic (a better national economy through improved wellbeing and individually empowered citizens) or ultra-individualistic (AI helps you achieve your own life goals, |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | whatever they may be). At the time of writing this article, this was not yet very clear<br>(P10.12) The goal of better and smarter service findability is obviously valuable in an environment of severe information overload. The use of AI to help a citizen find the relevant services or information and service providers in their situation is a worthy target. This includes finding responsibly and ethically sound ways of having personal information enrich the inquiries so that citizens can be given as accurate and relevant selection of services as possible without compromising their privacy and autonomy.<br>(P10.13) The attempt through the lens of data to better understand the state of wellbeing and needs of specific populations in order to design and produce services that come with genuine positive impact, was commended. The data policy and privacy issues that come with this were seen, however, as a matter to be looked into very closely.<br>(P10.14) It should be noted that service recommendations may not be based solely on the user's attributes; they would add the aggregate data of the user's reference cluster(s) the AI has identified to the mix. In this sense, the recommendation logic is quite similar to that of streaming media platforms: it combines your history and features with that of people it assumes are like you in certain critical respects. It remains unclear how the cluster data will be collected, what the data update frequency for that data would be and how a cluster's profile (aggregated attributes) would function for people whose profile places them, data-wise, at the outer edges of the cluster and makes them effectively anomalous in comparison to the people at the cluster's center.<br>(P10.15) If the trigger criteria are set too low, the recommendation engine might recommend these services to a young person going through normal "teenage pains" and who has not considered seeking professional help before. If a state-owned AI directs people like this to these services, an already difficult service situation may become notably worse.<br>(P10.16) The Board has noted that the service recommendation mechanism hinges on people's willingness and ability to either answer questionnaires on their situation, or wellbeing, to describe their situation or personal attributes in natural language, or to share their data from registers (this requires informed consent by the user). All these actions require different capabilities, ranging from linguistic ability to understanding what the sharing of personal information to a multi-actor network (Aurora Platform) means. The required ability to understand the overall algorithmic concept and working logic and critically consider the service recommendations may turn this into a service likely to benefit the more capable while leaving the digitally disenfranchised and some minorities by the wayside |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P10.17) Things become more complicated if we assume that the artificial intelligence in AuroraAI is largely another black box whose generated recommendations and estimations, let alone predictions, are not transparent and remain difficult to fully explain<br><br>(P10.18) The questions of **transparency** and **control** also become immediate in the context of the shared personal attributes. What tools will the user have to be able to track the use of their data? Will the users have access to the full list of service providers connected to the Aurora Platform? Can a user block a particular service provider from receiving their data? Can a user see which organizations and companies currently have that user's data on the platform?<br><br>(P10.19) Further on, however, the notion of user profiles and profile management came up and, even later, the concept of an AuroraAI account. The Board believes that with these steps, the issues of privacy and anonymity must be brought up and analyzed from scratch as the notion of automatically guaranteed anonymity, if there ever is such a thing, has become critically compromised. Ensuring the anonymity of users should be an essential part of the ethical use of AuroraAI. In the AuroraAI program, time periods covering even decades of a person's profile data have been mentioned in the vision of AuroraAI evolving into a personal "life guide" that learns from a person's attributes and events history over time and can thus generate a very accurate and timely guidance, service provision and predictive/proactive recommendations. It is therefore essential that particular attention is paid to this aspect from the anonymity and privacy perspectives.<br><br>(P10.20) The interests, values, and perspectives of citizens is essential (Levi and Stoker, 2000; Owen et al., 2013) in governmental actions related to AI. Governments should foster and facilitate societal discourse on the **desirability** of AI, and include active **participation** of various stakeholders and citizens. In reciprocal governance, AI experts should take the time to listen to and learn from users, especially their informal and emotional views on how the new service solution differs from existing (non-AI) arrangements, and what is expected of it. User involvement at every stage of the design process is therefore recognized as essential in public sector projects in Finland. However, in AuroraAI, this involvement has not been implemented.<br><br>(P10.21) Finally, **timing is critical** in ethics deliberation. The AuroraAI Ethics Board was set up quite late, considering the program's 3 year history at that point. More specifically, as the concept, vision, and planning for technical solutions have been mostly set already, it is questionable whether the Board will be able to have an impact on the program, especially its aims and goals and values, and there are openly |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | expressed suspicions of ethics washing. |
| P11 | The importance of the assurance that “humans are still in the decision loop” for public trust in artificial intelligence: Evidence from an online experiment (Aoki 2021) | (P11.1) This article reports on the results from an online experiment testing these hypotheses in the context of Japan’s long-term nursing care sector, based on the responses of care users and their families (*N* = 1542). The study did not find strong evidence to support H2. However, it found some support for H1: the proportion of those who trusted a care plan prepared with AI assistance more than a care plan not involving AI was higher by 8.95 percentage points with the HDL assurance than without. This highlights the importance of the HDL assurance and reveals respondents’ reservations about a complete AI takeover in care planning.<br>*(P11.2) As far as the AI-aided delivery of public services is concerned, this study posits that, in addition to what past studies have proposed or found thus far, the information communicated to the public about the use of AI might matter to the public’s initial trust in the services, especially in light of societal uncertainties about AI’s potency and its societal impacts*<br>*(P11.3) In light of societal anxieties surrounding AI, if a public service provider announces its plan to use AI in the delivery of public services, how might the public react? They might vaguely guess that AI would replace humans, who, as a result, would no longer make decisions, and some people might not feel comfortable with this. For example, if AI-driven technology is going to be used to customize learning for individual students, parents might doubt whether machines can really understand the complex needs and situations of their children, and they might feel apprehensive about their children’s having to learn with a flawed curriculum and flawed instruction designed by AI. They might hope that human teachers will still at least be in the decision loop as a safeguard. When AI is about to be used to suggest medical treatments, patients and their families might be sceptical about its ability to understand complex situations, and they might feel uncomfortable with the idea of replacing human doctors with a machine that seems to lack compassion and a caring human heart. They might wish for human doctors to be available in case something goes wrong.*<br>(P11.4) In essence, the aforesaid public concern arises, because, amidst the speculation about AI takeover scenarios, the public wish to make sure that some functions are still reserved for humans, especially when they are not clearly informed about functional allocations, i.e. the way “functions are allocated to humans and machines” (Abbass, 2019, p. 161). Given all these considerations, this study proposed that communicating a piece of information, namely, that humans are still in the decision loop, makes a difference<br>(P11.5) Another factor that might make a difference in the public’s initial trust is communicating the benefits of introducing and using AI. This proposition is inspired by the purpose basis of trust conceived |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | of in the context of the operator-machine relationship mentioned earlier (Lee & Moray, 1992; Lee & See, 2004). More specifically, this basis of trust refers to "the degree to which the automation is being used within the realm of the designer's intent" and whether it "corresponds to faith and benevolence and reflects the perception that the trustee has a positive orientation toward the trustor," where "the trustee" is the designer of the machine (Lee & See, 2004, p. 59) – that is, in order for an automated system to inspire user trust, the designer's intention has to be benevolent and place the user's interests above their own<br>(P11.6) The current study applies the purpose basis of trust in operator-machine relationship to the relationship between the public and AI used in the delivery of public services. However, unlike operators using machines, public service users may not care as much about the intentions embedded in the design of the AI-driven machine as about the intentions of the government or other public service provider, which is the ultimate locus of responsibility in their eyes. Thus, the reason the service provider is using AI can be an important basis of public trust; to earn public trust, the provider's intention to introduce the AI decision aid has to be benevolent and place citizens' interests first, rather than seem to be intended to benefit those providing the services, such as saving time and reducing the burden on staff. This logic also applies to the public's initial trust in public services delivered with the help of AI relative to public services provided by humans without utilizing AI.<br>(P11.7) Nevertheless, in both cities, concerns were raised that the assessment of the user's condition and a doctor's diagnosis do not provide enough input for an AI system to generate appropriate and credible care service options. To arrive at appropriate care plans, the preferences and feelings of users and their families, their living environments and support systems, and their economic conditions also need to be taken into account.<br>(P11.8) Moreover, the AI proposals did not help care managers to explain why particular types of care service were proposed.<br>(P11.9) In Toyohashi, care managers reported that the service users and their families did not embrace AI and were resistant to the new technology, and that they could not think of AI proposals as real; they did not even bother to look at the tablet showing the service options. A care manager expressed the concern that service users and their families wish to be received by a human with a heart, not by a machine, and that they might think care managers are slacking by relying on automation.<br>(P11.10) This finding points to the fact that concerned individuals are not ready to see care planning |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | handled completely by AI, and it leads to a policy recommendation that government or other public service provider ought to engage in democratic communications about technology with the public and add the HDL assurance in communications announcing the introduction of AI. |
| P12 | Human Control and Discretion in AI-driven Decision-making in Government (Mitrou, Janssen, and Loukis 2021) | (P12.1) Hence, there is a need for human-control and the decision-makers should be given sufficient authority to control the system and deal with undesired outcomes.<br>(P12.2) Does the interpretation of the legal rules and the subsumption of facts to these rules presuppose the participation of humans or can be entrusted to AI to apply the law to the specific case and come to a decision that is compliant with the law? The answer depends on the degree of discretionary power conferred to the administrative agency and the use of algorithm-based decision process as a complement or substitute to human-made decisions<br>(P12.3) Automated decisions become more difficult if individual cases' details or still- unstructured information (that cannot be part of the historic data from which rules are extracted) has to be taken into account. Although AI is increasing in capacity, its use seems to be mostly rejectable/questionable when it comes to accomplishing discretionary tasks, or where there is a need for structuration of information or assessment (Etscheid, 2019). These systems are self-learning and automate the construction of criteria and rules (from historic data) to reach a decision. The determination of applicable norms is also disputable, as it necessitates a full understanding of the facts and complexity of the case at stake.<br>(P12.4) Additional complexity may occur where an automated system has to take not only a data-driven but a values-driven decision. Rule-based systems may be embedded into an AI decision-making application but this is hardly the case when the public administration has to strike a balance between competing rights and interests, not to mention the difficulty to ensure that a principles-driven system, "programmed with slave-morality" (Wirtz et. all, 2019) does not result in moral or legal rigidity with regard to individual circumstances.<br>(P12.5) Applying uniform rules and criteria across all decisions, without variation or particular consideration of the specific situation, may constitute a misuse of discretionary power that leads to an unlawful decision (Cobbe, 2019).<br>(P12.6) Therefore, in cases of complex government decisions, in which i) there are a lot of input data items to be taken into account, with some of them being unstructured, or even not known in advance; ii) in case of rules extracted from historic data, if these do not include all these numerous data items (structured and unstructured); iii) it is not clear which law has to be taken into account; iv) a balance |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | between competing values, rights and interests, then increasing human control and discretion is required (so AI should be used as a decision-support, which provides recommendations to human actors, or at least humans should have a role ‘in the loop’).<br>(P12.7) Also, some degree of human control, intervention and discretion might be necessary: (i) in order to avoid the use of data-sets that are non-representative, inaccurate, or reflect undesired past practices (or past practices that are not appropriate for the current context), or even reflect historic discriminations; (ii) if there might be problems of lack of consistency and predictability of the decisions provided by AI, especially in cases where new data are used for continuous learning of the AI algorithms; and (iii) for increasing transparency, explainability and auditability of administrative decisions, and therefore accountability for them<br>(P12.8) Humans remain accountable for the systems. Given the deficiency of AI systems, there is a huge need for meaningful human control of AI systems. In such situations, humans control the input data, information processing and output results and have the discretion to deviate from the suggested decisions by AI. This is different from classical discretion in public administration, and knowledge of algorithms, legislation and the situation at hand is needed. This makes this even more challenging. Meaningful human control is particularly important when there might be possible failures. Humans can play different roles and can deviate from decisions suggested by AI. This would require a sound argumentation and can be related to the input data, the algorithms used the understanding of part of the context not captured by data, the interpretation of regulations, or other aspects not captured by AI.<br>(P12.9) On the other side, “it is a mistake to assume [these big data analysis techniques] are objective simply because they are data-driven” (White House Report on Big Data, 2016). Data inherits the bias from the past (Janssen & Kuk, 2016). Serious concerns are expressed with regard to the “algorithmic neutrality” as the use of data as input to an algorithm, the very purpose (especially) of machine learning algorithms (categorise, classify, separate) and the inner working of the algorithm itself that may result into discriminatory decisions. We should not ignore that humans are not deprived from personal beliefs and biases that may affect their assessment (Vogl et al., 2019). However, people are most concerned about bias in the application of algorithms and direct or indirect discrimination is considered as one of the most crucial challenges in the use of AI-driven tools for decision-making areas. Various aspects of discrimination, including gender or race discrimination, can occur for several reasons. The selection and the quality of data fed into the system (lack of representativeness and accuracy), the data samples used to train and test algorithmic systems, choices of features, metrics and analytic structures that reproduce the designers’ perceptions and biases may predefine the outcome to be produced (Leslie, 2019). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| P13 | Managing Algorithms for Public Value (Selten and Meijer 2021) | (P13.1) To stimulate the creation of sigma-type values of algorithms and reduce the chance of adverse effects on theta-type and lambda-type values, it is therefore important that people who are deciding about the use of algorithms or working with them, are enabled to assess the quality of the information the algorithm provides. Policy makers need to be informed about how the algorithm is developed and its real world effects to be able to use them effectively and prevent automation biases (Saunders et al.,2016; O'Neill, 2016; Lyell, & Coiera, 2017; Veale et al., 2018; Zerilli et al., 2019a; Weller, 2019).<br><br>(P13.2) Accountability and transparency are traditional requirements to promote trust in public institutions (Harrison & Sayogo, 2014). However, providing transparency and accountability of algorithmic functioning is difficult, due to internal, external and process opacity (Burrell, 2016, p. 5; Goad & Gal, 2018; Lepri et al., 2018). For this reason, Kemper and Kolkman (2019) introduce the notion that a critical audience is needed to make algorithms transparent and accountable. See Wieringa (2020) a complete overview of these accountability relations in the context of algorithms, here a distinction is made between what are defined to be technical audiences and non-technical critical audiences. First, accountability towards the technical audiences, which are the administrative, professional and legal accountability audiences. Accountability to these audiences implies auditing of the algorithm by experts to test if it is functioning correctly and function in accordance with the law. Making technical accountability possible requires providing transparency about input and output data, have data management protocols, make the algorithmic code open source and work on techniques to make algorithmic functioning more explainable (Kemper & Kolkman, 2018; Burrell, 2016; Datta et al., 2016; Goodman & Flaxman, 2017; Wieringa, 2020). This is important because it can help point out potential flaws in the algorithm, such as coding errors or biases (Weller, 2019). While this type of accountability (partly) solves external and internal opacity problems, it does not tackle the process opacity issue. Second, non-technical audiences are the political and social accountability audiences. These non-technical audiences, political parties but also civil servants and citizens for example, are not necessarily able to obtain the technical knowledge to understand algorithmic functioning. For these audiences the use of algorithms creates a principal-agent relationship. Algorithms are developed by a team of programmers, but are used in practice by civil servants who have little understanding about how the algorithm functions and makes predictions (Meijer, 2018; Van der Voort et al., 2019). Similarly, citizens are affected by a decisions that are made using an algorithm but are not able to understand how these decisions were constructed (O'Neill, 2016, p. 8). Accountability thus also needs to be rendered to society as a whole. It is important that citizens broadly understand and become comfortable with the strengths and limitations of an algorithm to overcome fear of the unknown (Weller, 2019).<br><br>(P13.3) Building on Deley and Dubois (2020) who describe that trustworthiness of algorithms can only |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | be provided by demonstrating their reliability, it is therefore argued that public organisations must do precisely this – demonstrate the reliability of the algorithms they use. This is what Grimmelikhuijsen and Meijer (2014) refer to as transparency of policy outcomes or effects. Reliability is tied to all three value clusters. It entails that algorithms make public organisations function more effective (sigma-type values) and that decisions are not leading to adverse, unfair or discriminatory outcomes (theta-type values). Reliability also logically refers to accounting for lambda-type values. Clear back-up procedures for when the algorithm breaks are needed and a detailed strategy for maintaining the reliability of algorithms over time is needed to ensure the production of public value over time. Ultimately to enhance trust in algorithms also requires to demonstrating how they affect decision-making and policy outcomes<br>(P13.4) Providing reliability over outcomes, and thus make accountability of algorithms possible demands evaluation by all critical audiences that are described by Bovens (2007). Building on public administration research that aims to discern the causal mechanisms of policy interventions, implies this is only possible using empirical evaluation (James, et al., 2017).The use of field experiments is especially suitable to test the effects of algorithms (Hansen & Tummers, 2020). Fields experiments include a careful monitoring of the implementation of an algorithm and comparing this to similar situations where the algorithm is not used. This type of experiments enables policy makers and researchers to discern the advantages and adverse effects the implementation of an algorithms has in practice. This is what within experimental research is referred to as a as the ‘counterfactual’ (James et al., 2017). Simultaneously, testing the effects of an algorithm in a real world setting therefore requires careful consideration costs, ethics and validity (Hansen & Tummers, 2020).<br>(P13.5) Finally, it is important to explicitly discuss the public value dilemmas that have been presented in this article with citizens. In this discussion it is important to remember that not all value trade-offs are equal. There is a distinction between what Tetlock et al. (2000) define to be ‘secular values’ and ‘sacred values’. Trade-offs between secular and sacred values are what Tetlock et al (2000) refer to as being taboo trade-offs. No matter how big the monitory advantages an algorithm might be able to deliver, if it negatively impacts these sacred values citizens will not be willing to accept the use of algorithms (Tetlock et al, 2000). For the use of algorithms, many of the theta- and lambda-type values (e.g. discrimination, biases, risk of collapse) might fall in the sacred value category. This means that many of the public value trade-offs civil servants will be dealing with are not routine, but will be taboo trade-offs. When do the advantages algorithms bring outweigh their negative conquests - which values are secular and which are sacred? Most importantly, answers to these questions need to be formulated together with citizens, because public value is the collective expectation of citizens (Moore, 1995). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P13.6) Theta values comprise of values such as fairness, transparency and accountability (Hood, 1991). The biggest critiques on the use of algorithms by public organisations focusses on violations of these type of values. Algorithms are linked to increase unfair outcomes and to introduce biases in the decision making process (O'Neill, 2016; Mittelstadt et al., 2016; Crawford & Schultz, 2014). For example, algorithms used to predict the risk of reoffending have been proven to be almost twice as likely to mistakenly flag black than white defendants (Crawford & Calo, 2016). These issues are mostly caused by the fact that the underlying data the algorithms are trained on already contain these biases. Machine learning algorithms will learn and reinforce these existing structure – a problem that is especially problematic because of the above outlined abilities of algorithms to become a self- fulfilling prophecy and to be highly scalable (Willis et al., 2007; Ferguson, 2012; Diakopoulos 2014; O'Neil, 2016; Andrews, 2019).<br>(P13.7) Underlying all these elements is one main challenge; algorithms are in many instances not transparent. This lack of transparency makes it difficult to distinguish whether an algorithm is producing fair and accurate results or is actually discriminatory. In the current literature there are mainly three causes presented for this lack of transparency: internal, external and process opacity. External opacity means that many algorithms do not reveal their internal mechanisms. A key reason for external opacity is intellectual property protection but even when an algorithm is open-source, it is externally opaque for a majority of people since only a small part of the population can understand an algorithm's functioning by looking at the code (Lepri, et al., 2018). Yet, even when input data and the internal mechanisms are disclosed and reviewed by experts, algorithms can be non-transparent due to internal opacity (Oswald, 2018). Algorithms learn how to perform tasks by themselves, but what they learn and how they function is often not interpretable for humans (Guidotti, et al., 2018). Internal opacity refers to this fact that algorithms will always contain a "degree of unavoidable complexity when solving complex tasks" (Burrell, 2016, p. 5).<br>(P13.8) In addition to these two technology-related forms of opacity, process opacity means that is unclear how exactly algorithms are used by policy makers (Saunders et al., 2016). Even when an algorithm is working very well, if it is used incorrectly it can still produce unfair outcomes. Because of their inherent opaqueness, public decision makers are unable to control algorithmic outcomes (Peeters, 2020). This can result in the introduction of new types of biases in public decision-making processes (Deley & Dubois, 2020; Busuioc, 2020).<br>(P13.9) Even though it seems desirable to maximize scores on all values, the literature highlights that governing in the public sector will require trade-offs between values. This leads to dilemmas for public |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | organisations when deciding about the use of algorithms. Questions such as the ones posed by O'Neill (2016, p. 95): "[are] we as a society […] willing to sacrifice a bit efficiency in the interest of fairness [?]", and Young et al. (2019, p. 303): "If an AI system is half the cost of the current system, but 10% less accurate, is it better?", are therefore extremely relevant.. It is the task for public managers to find a balance between all public values. Do they prioritize lambda-, theta- or sigma-type values? Citizens expect efficiency, transparency and robustness, but it might not always be possible to realise all these values at the same time. Managing the adoption and implementation of algorithms thus entails balancing these expectations. This management of public expectations is what Bozeman (2012, p. 176) refers to as the management of publicness. |
| P14 | Improving public services using artificial intelligence: possibilities, pitfalls, governance (Henman 2020) | (P14.1) In the context of the use of AI, empirically-based concerns about bias around gender/sex and race/ethnicity are well documented. The use of the COMPAS system in American sentencing and parole decision making has also been well analysed for reproducing systemic biases against Afro-Americans (Allen, 2019; Benjamin, 2019). Facial recognition systems developed in the Western world have also been critiqued for having much higher error rates for non-White people (Bacchini & Lorusso, 2019). These problematic outcomes can result from the training data used to build the AI. If that data is incomplete, inaccurate or reflect historical structural inequalities, this gets learnt by the AI and informs its outputs.<br>(P14.2) Technical accuracy is also a challenge with the use of AI. This is particularly acute when used in a probabilistic decision-making or predictive manner, such as to predict child maltreatment or crime recidivism. In these cases, AI calculates future events based on probabilities, which may or may not occur. False negatives and false positives can have serious consequences, and the context of these decisions often requires carefully balancing between them (Henman, 2005). For example, a false negative that a child is unlikely to be maltreated, resulting in no child protection investigation, may mean the child dies. Alternatively, a false positive may result in a child unnecessarily being removed from the care of its parent/s. Accuracy is therefore especially pertinent in such high stakes decisions.<br>(P14.3) Firstly, using AI to make risk assessments (of individuals) is an exercise in calculating probabilities. These are not realities or certainties. Rather, such assessments are made by relating the current case (or individual) to cases (or individuals) with similar characteristics (or profiles) and deducing similar outcomes. The legal basis of acting based on a likelihood, rather than an actuality, can be problematic (Harcourt, 2008; Henman, 2005; Schauer, 2003). For example, criminal law is based on people being tried on an offence having been committed, not on what they may do in future. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | (P14.4) Secondly, in many areas of administration, human administrators are required to exercise professional judgement and discretion to best determine how the rules apply in complex situations. As is well documented, automating administrative decision making has reduced human discretion and gives rise to concerns that an individual's situation may not have been appropriately considered in making an administrative decision (Adler & Henman, 2009; Evans & Hupe, 2020; Garson, 1989). While AI has the capacity to be more nuanced than standard algorithms, it is also less clear how decisions on specific cases are made.<br><br>(P14.5) Thirdly, the foregoing observation points to the "black boxed" (Pasquale, 2015) nature of algorithms in general, and AI in particular. While complex algorithms have always made it hard to understand the basis for decision-making (Weizenbaum, 1984), they have always been coded by humans to directly implement rules and procedures. Machine learning algorithms, in contrast, generate their own internal rules to determine input to output. In doing so they add another layer of opaqueness to their decision making processes.<br><br>(P14.6) Sometimes new laws are required to ensure that an algorithmic decision is treated as equivalent to a human decision. However, given that an algorithm does not have human autonomy and agency, identifying responsibility when an AI decision is made is challenging. Who (or what) has responsibility or liability for the error: the machine, the creators of the machine, the coders, the managers who decided to deploy the machine? This situation is further exacerbated when AI tools are developed by external organisations or companies, and deployed by governments, a situation that is likely to increase as AI becomes more mainstream and used in an "off the shelf" manner.<br><br>(P14.7) Explainability is a considerable challenge for AI based decision-making. In short, how did an AI reach the decision it did with the input data it received? Computer and information scientists are working to develop algorithms that can independently provide an approximate explanation of how an AI generated a decision in a particular case (Samek et al., 2017). Such an approach involves giving the AI lots of separate sets of input data and assessing patterns in output data to identify the key input variables that appear to make a difference in AI decision-making. While such explanations are not precise, they are able to draw attention to key drivers emergent in the AI decision-making. Similarly "adversarial testing" is an approach whereby people try to "break" an AI or make it make very wrong decisions (Qiu et al., 2019).<br><br>(P14.8) While not new, software development processes are increasingly using a "privacy by design" approach (Cavoukian, 2012) whereby legal considerations of data protection and privacy are not left to |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | the end of the product development process, but built into the very architecture (and even software) of algorithmic decision making systems. Such approaches can be made more widespread and can also be extended to incorporate other ethical considerations in AI development (Morley et al., 2019). The International Standards Organisation (ISO) and its Australian counterpart – Standards Australia – are working on building technical standards for ethical AI (www.iso.org/committee/6794475.html).<br>(P14.9) In addition, there needs to be legal clarification around responsibility, appeal and redress in AI based decision-making.<br>(P14.10) Another approach to governing AI in government is to provide an independent quality assurance mechanism to test AIs' compliance with ethical/legal considerations. Australia's Chief Scientist, Alan Finkel, has referred to such an idea as a "Turing Certificate" (2018). Ensuring that AIs used in government are available for independent assessment is essential, and also allows the opportunity to subject them to the bias and adversarial algorithmic testing mentioned above.<br>(P14.11) A further approach is to develop practical tools and processes by which AIs can be assessed. To date, much of the discussion has been abstract and at the level of principles. Currently little work has been done on developing these practical tools, though these are emerging. Australia's Department of Industry, Innovation and Science (Dawson et al., 2019) provides a "toolkit for ethical AI" that includes impact assessments, risk assessments, review, best practice guidelines, industry standards, collaboration, monitoring, improvement and recourse mechanisms, and consultation. The UK government issued a guide to using artificial intelligence in the public sector (Government Data Service & Office of Artificial Intelligence, 2019), which focuses on assessing, planning and managing AI, and using AI ethically and safely. A more comprehensive Ethics and Algorithms Toolkit led by John Hopkins University's Centre for Government Excellence provides a "risk management framework" (www.ethicstoolkit.ai). Technology organisation VDE has also developed a framework to operationalise AI ethics (Hallensleben & Hustedt, 2020), while the Ada Lovelace Institute (2020) has provided an overview of tools for assessing algorithmic systems |
| P15 | Thinking problematically' as a resource for AI design in politicised contexts (Petersen et al. 2021) | (P15.1) AI solutions in politicised contexts often precede finding open problems, such that by the time ethnographic studies are conducted (if they are), formulations may have already been settled, including assumptions about caseworkers' roles in data-driven public services.<br>(P15.2) We wanted to explore whether AI could support an open-ended approach to public services in Denmark, avoiding the inflexibility that has been characteristic of this area. The lack of early openness to problem definition, characteristic of current AI projects in this area, also means that caseworkers - |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | who directly interact with both the systems and the citizens - are often left out of fundamental design decisions [47]. In taking an alternative approach, we (the authors of this paper) ask amongst each other: in what circumstances is AI seen as a solution and by whom, what problems does it solve, and who defines the problems that need solving, and on what grounds?<br>(P15.3) From an HCI design perspective, early engagement is motivated by the aim to posit tensions and misunderstandings as opportunities for 1) clarification [28], and 2) mutual learning across epistemological divides [10] that can be incorporated into problem formulations [7]. Previous HCI research aiming to 'bridge the gap' between ethnography and design suggests the need for better integration and cooperation through early and direct involvement. If working relationships are strong, the technological solution will be grounded in shared meanings, and stakeholders will have a voice in the process. With the right methods, we can discover the problems we want to solve, and these problems will originate solutions.<br>(P15.4) Using problematisation to understand the different epistemologies involved in AI design also helps us understand their effects and, through this understanding, counter them. By allowing the design space to stay open, we leave room open to thinking about alternatives. However, in our case, we needed to think problematically to understand what those alternatives might be. As such, we hope that our empirical moments and reflective practice can become an inspiration for future design projects. If we do not problematise the politics inherent in the technological solutions we build, we may prevent them from having their full effect in practice.<br>(P15.5) As made clear from our analysis, ethnographically informed design is not straight forward. However, it helps avoid leaving out essential aspects of work in initiatives to support it, as it resonates with actual circumstances and not some 'idealised' version of events [16]. Ames [4] noted that one way to fight charisma is to deflate or be 'anti-charismatic'. With this paper, we call on AI researchers and designers to enable problematisation as a strategy for concretising alternatives to the dominant ideals, such as those we found in our project. We learned from our analysis and ongoing work on this paper that we need to be ready to have these discussions across epistemologies. Thus, problematisation is not just a valuable resource for ethnographers to navigate the design process – it is also a meaningful process for everyone involved in the design of technologies. |
| P16 | Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA (Hartmann and Wenzelburger | (P16.1) The practitioner interviewed clearly believes that the evidence generated by using COMPAS along other available information will produce better decisions— and a clear conscience of the decision-maker. Interestingly, all actors still maintained that the COMPAS scores did not involve an "in and out"-decision and that the decisions by actors in the judicial system still "started and ended with facts". Some |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | 2021) | of them still emphasized the discretion of human actors and the non-binding role the risk scoring by the ADM system played.<br>(P16.2) COMPAS is praised by all interviewed practitioners to deliver new information and most of them seem to follow the categorization of low, medium, and high risk as put forward by the software. While it is true that most interviewed experts also emphasized that other sources of information were still consulted, and only the joint interpretation of all available information would lead to a decision, the emphasis of the research-based and validated character of COMPAS as well as the importance of the three-category risk classification which was mentioned in all interviews points to the central role played by the risk assessment tool in the entire decision-making process.<br>(P16.3) In contexts where human discretion is involved because "soft" factors play a role, such as the credibility of an offender to change its habits, algorithms may on the one hand provide important information that makes a human decision more evidence-based; on the other hand, an algorithmically generated score may also be an important anchor point from which a human decision maker will only rarely deviate.<br>(P16.4a and 4b) Second, as the changed process may have palpable downstream consequences on society, it is of prime importance to investigate not only whether the rules that are implemented *within* the algorithm are ethically defendable, but also to assess whether the *system as a whole* and the transformation of the decision-making process it involves is politically legitimate. |
| P17 | Digital Discretion: Unpacking Human and Technological Agency in Automated Decision Making in Sweden's Social Services (Ranerup and Henriksen 2022) | (P17.1) *Citizens, caseworkers, and technologies.* The final decision on applications is made jointly by a caseworker and the technology.<br>(P17.2) A positive decision is much easier to programme. Negative decisions are much more complex. One might receive a partly negative decision when it should be a positive decision. There are many reasons for negative decisions. Our responsibility is to be very clear about our reasons. They must be justified. We also need to be clearer about how the robot makes decisions. (Manager No. 1, October 9, 2018)<br>(P17.3) We are not influenced by our emotions. [ . . . ]We are neutral. We make a decision based on the rules and regulations. That is what our work is based on. We have laws that we must follow. We can concentrate on that instead of focusing on applicants who may be disappointed, angry, or threatening. (Caseworker who works with social assistance, October 9, 2018) |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P17.4) According to the municipality, such decisions are legally binding. From the perspective of legal, appeal-proof decisions, the RPA is an objective decision-maker. The decisions are correct because they are based on the rules. (Trelleborg Municipality, 2017, p. 3) [The RPA] is legal. It is exactly in line with the Board's principles. (Manager No. 2, September 26, 2017). However, the interviewees said that people are still needed to work with technology development and to resolve complicated negative decisions.<br><br>(P17.5a) *Platform for applications.* The platform receives decisions from the separate routine. A text message and an email inform citizens about the decisions. Information about how to appeal negative decisions is also provided.<br>*Citizens.* Citizens can log onto their accounts in the platform and read their decision. If the decision is negative, they have the right to appeal.<br><br>(P17.5b) *Appeal against negative decisions Citizens.* A citizen can file an appeal against a negative decision.<br>*Civil servant at the help desk.* Assistance with the appeal is available at the help desk. A civil servant has a template to use with appeals. As one civil servant explains: We print the decision so that the citizens really understand the basis for their negative decision and the argument for their appeal. We explain how to make an appeal, but we don't write it. We cannot be held responsible if the appeal is unsuccessful. (Civil servant at the help desk, October 9, 2018) After the appeal has been prepared, it is submitted for review by caseworkers who work with social assistance.<br>*Platform for applications.* In July 2019, a function for making an appeal was added to the platform. Formerly, appeals had been submitted on paper.<br><br>(P17.6) *Ethical values.* Digital discretion facilitates avoiding unethical actions and corruption. At Trelleborg, contact between the citizens and caseworkers who manage the social assistance decisions is by mail, telephone, or the digital platform. These communication modes are used in a process governed by rules and procedures that are carefully built into the technology. This process safeguards fair and uniform decision making (Zouridis et al., 2020).<br><br>(P17.7) At Trelleborg, RPA alone makes about one third of the decisions. Many more decisions are made partially by automation (Trelleborg Municipality, 2017). There is a continuous development of the texts explaining the motivations for all decisions, positive and negative.<br><br>(P17.8) Our theoretical framework indicates that one desired effect of digital discretion is to increase |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | governments' political legitimacy. From a municipal perspective, this legitimacy increases when citizens find employment and become self-supporting as a result of governmental programs. At that point, government policy is perceived as valid (Trelleborg Municipality, 2015, 2017, 2019). Ultimately, support increases for governmental digitalization programs when decisions are standardized, rules and procedures are followed, and potential biases and preferences are eliminated.<br><br>(P17.9) Digital discretion, which facilitates adherence to rules and procedures, means decision-making accountability increases. At Trelleborg, although the caseworkers worked in close cooperation with automated decision making, they understood that they had been relieved of some decision responsibility. As Lipsky (2010) observed, street-level bureaucrats are greatly concerned with their own accountability. Over time that accountability has become embedded in the system. The automation process began in late 2016 and early 2017. By 2018, the process had reached the point at which the digital platform could explain negative decisions. In contrast to cases that involve children in need (Petersen et al., 2020), decisions about social assistance are more standardized.<br><br>(P17.10a) Preventing errors is also related to the new management model in terms of digital applications, algorithms for decisions, and the ongoing development of motivations communicated in both positive and negative decisions. Yet caseworkers are still needed for the review of complex negative decisions and for complicated applications.<br><br>(P17.10b) In contrast to Reddick's possible outcome, our case shows that a group of qualified caseworkers is still needed to review negative decisions and to manage citizens' appeals. A repeated theme in this article is that technology is not simply a technical issue. Even when automated processes replace manual ones, experts are still needed, not least to oversee the use and development of the automation. This echoes the findings of Petersen and colleagues (2020).<br><br>(P17.10c) As a result of these processes, despite some changes in their professional roles, caseworkers remain in close contact with citizens who apply for social assistance and who appeal denials of their applications. Caseworkers in social services in the area of decisions about social assistance also provide help during regular obligatory meetings with a focus on becoming self-supporting and with an array of job-related activities from interview coaching to employer contacts. However, it is clear that the primary role of caseworkers as decision makers has diminished (Zouridis et al., 2020). Although caseworkers still have the final responsibility for decisions, RPA has an increasing role in daily practice<br><br>(P17.11) Thus, a hybrid model involving humans and technology (Callon, 1986) can be used for core |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | decisions for social assistance. We observe that RPA, as a new standardization tool, has a place in the search for balance between human and technological discretionary practices and other concerns such as legality, security, and transparency (Brauneis & Goodman, 2017). |
| P18 | Artificial Intelligence and Administrative Evil Young et al. 2021) | (P18.1) The AI system allows an individual agent to uncover previously undetected negative externalities associated with, for example, means-testing eligibility thresholds that in aggregate are so large that they constitute an evil despite the cause being rooted in the desire for programmatic efficiency and effectiveness. Thus,<br>**Harm Discovery:** AI can decrease the chance of administrative evil by improving the information available to agents about potential harmful consequences associated with the decision, for example by revealing new correlations between decisions and outcomes<br>Our *Harm Discovery* proposition suggests that a future role for individual bureaucrats might be to audit and oversee the operation of AI systems – potentially utilizing further AI systems. This opportunity to evolve the role of bureaucrats however likely requires large, systemic changes to the ecosystem of statutory regulation, training opportunities, and individual incentives.<br><br>(P18.2) At the same time, the interplay between AI and an individual's capacity for collecting and weighing evidence may also *increase* the risk of administrative evil. Individuals are only boundedly rational, and often resort to the use of heuristics and other satisficing – that is "good enough" – choice processes, which are subject to various biases (Tversky and Kahneman 1974; Simon 1997). Specifically, when individuals interact with technology, they suffer from what is known in computer science and industrial psychology as *automation bias*: when individuals share control with an automated decision support tool, they become overly-reliant on the tool and do not critically review its recommendations or behaviors before taking action or making a decision (Manzey, Reichenbach, and Onnasch 2012; Mosier et al. 1996). For example, consider again our hypothetical social services organization but now with a different AI implementation. Suppose AI is used to help individual street-level bureaucrats determine whether applicants are eligible for benefits, and if so at what level. Over time, the individual can become lulled into a false sense of security that the AI will make the correct recommendation, particularly if it is highly accurate under most circumstances. But when an applicant arrives whose circumstances constitute an edge case that the AI is not calibrated to handle properly, the tool using bureaucrat may fall victim to automation bias and deny the applicant's claim when they are in fact eligible. In agency theoretic terms, in this case the bureaucrat is the principal, the AI-enabled decision tool is the agent, and the difference between the AI's true error rate and its presumed or reported error rate constitutes an information asymmetry between the parties. This suggests in particular the first point of **AI Exuberance:** AI can increase the chance of administrative evil because it (1) introduces the risk of |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | automation bias; (2) thrives within the cultural ideology of technological rationality; and therefore (3) may be enthusiastically deployed without appropriate testing or to address an issue for which AI is not the optimal available solution.<br><br>(P18.3) A third effect of AI on the micro level is that it typically reduces the discretion that is afforded to individuals. Technology in general, and ICTs and AI in particular, shift the locus of control away from front-line staff and street-level bureaucrats (agents) towards management (organizational principals) by curtailing the former's ability to exercise discretion (Garson 1989; Busch and Henriksen 2018). This is part of a more general tendency of automation. AI can perform increasingly complex tasks and it thereby threatens an increasing share of white-collar professional jobs (Frey and Osborne 2017; Lee 2016). This shift of control away from individual agents and towards principals will likely affect the risk of administrative evil in a nuanced way, in that this risk is likely jointly determined by the attributes of agents and principals. If agents have a greater disposition to commit evil than their principals, then this shift will likely decrease the risk of administrative evil, and *vice versa*.<br>Suppose first that individual agents are more disposed to contribute to evil. This individual disposition can be attributed to prejudice, malice, or other malfeasance on the individual's part. But as the theory of administrative evil suggests, this disposition can also arise from masking, that is, that the agent has incomplete information or is unable to correctly parse the information they have. Examples of this sort of evil-through-misfeasance include decision biases introduced from organizational loyalty, aversion to interpersonal conflict, and the bounded nature of human rationality in general. AI systems do not share these risk factors. Curtailing the discretion of individuals who are disposed to contribute to evil hence reduces the risk of administrative evil. But often the principal might instead be the source of the ethical problem. More precisely, agents may instead be less disposed to contribute to evil compared to principals. Human decision making is sensitive to values and individual discretion can be motivated ethically. In result, individual decisions need not comport with an organization's explicit tasks and goals. Such a conflict between individual values and organizational goals poses a dilemma for human agents who then can respond in several ways: they may accept the organization's goals (loyalty), raise objections (voice), resign in protest (exit), or sabotage the effort in different ways (neglect) (Hirschman 1970; O'Leary 2013; Rusbult et al. 1988). AI agents, in contrast, do not have these options. Hence, AI may increase the risk of administrative evil when it limits the discretion of agents disposed to prevent or resist evil. Thus, **Control Centralization:** AI consolidates discretion at higher levels of organizational hierarchy and thereby moderates the risk of administrative evil; the direction of its moderation is a function of whether organizational leadership is more or less predisposed to commit evil than street-level staff. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P18.4) Although we hypothesize that the adoption of AI centralizes control – because it centralizes decision-making power – the adoption may yet reduce control in the specific sense that AI are difficult to interpret and that their decisions can be hard to explain. One of AI’s principal advantages over human agents is its ability to analyze large, high-dimensional, and complex data. Unfortunately, this advantage comes with an inherent tradeoff: As its ability to process complex data increases, an AI system becomes harder to audit. The most powerful AI are often the least understandable with respect to their decision-making process *ex post* (Weld and Bansal 2019). This suggests that information-based agency problems are likely to materialize in a way that may lead to administrative evil. The organization can be seen as the principal, the AI system as the agent and the lack of interpretability as an information asymmetry. For illustration, consider the implementation of an AI system in our hypothetical social services organization where the system monitors the behavior of disability benefit recipients to detect fraud. But insofar as the AI system is not interpretable, and that automation bias can lead many in an organization to implicitly trust machine judgement as superior to human judgement (Hoff and Bashir 2015; Parasuraman and Manzey 2010; Dzindolet et al. 2003), classification errors become increasingly difficult to rectify (O’Neil 2016; Eubanks 2018). Type I Errors, that is false positives where an innocent benefits recipient is identified as engaging in fraud, are bound to occur but might be less likely to be recognized and rectified. Analyzed through the lens of agency theory, this is a problem of moral hazard: the agent has behaved in a way that is at odds with the principal’s stated objectives, but information asymmetries make it costly for the principal to notice or understand the discrepancy. Thus, **Technical Inscrutability:** AI can increase the chance of administrative evil because it masks decision making (due to increased complexity and decreased transparency): AI lacks explainability or requires technical expertise to understand decisions.<br><br>(P18.5) This form of moral hazard and its implications for administrative evil also contribute to the problem of goal displacement. Goal displacement occurs when organizations pursue intermediate and easily measurable goals instead of their originally intended but relatively unmeasurable goals. This often inadvertently leads to worse performance in terms of the originally intended goals (Lavertu 2016; Moynihan 2008). More relevant for our purposes, goal displacement may lead to administrative evil, even if only to minor forms of evil. Suppose an AI system is tasked with assigning benefit recipients with the appropriate benefit amounts. Suppose further that the system is designed to set the lowest benefit rate that meets statutory obligations. For some benefit recipients, the AI agent may identify a positive correlation between lower immediate benefit rates and greater long-term cost savings, and it may thus begin to systematically assign more recipients lower average benefit amounts when the system could have used administrative discretion to assign increased benefits. But suppose that the long-term cost savings are in part attributable to the fact that these recipients died earlier than they would have if |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | they had received increased benefits. In pursuit of the intermediate and more easily measurable goal of cost savings over time, the AI system minimized benefits levels as low as legally possible by indirectly maximizing recipient's mortality. In fact, AI agents demonstrate goal-displacing behavior to the extent that the field of AI research has its own associated terminology: *reward hacking* (Amodei et al. 2016). Thus, **Organizational Value Misalignment:** AI can increase the chance of administrative evil by increasing organizational goal displacement.<br>(P18.6) Yet even when an organization's decision to adopt AI is made with the best intentions – indeed, irrespective of the intention altogether – the risk of administrative evil can increase from AI adoption because AI can introduce or intensify decision-making biases that mask administrative evil and facilitate moral inversion. For example, a necessary precondition for the proposition of Harm Discovery to hold is that there are sufficient quantitative (and more specifically, machine readable) data available for AI to identify the appropriate patterns correctly and reliably. Despite the exponential growth of ICTs in the late 20th and early 21st centuries, which has made AI increasingly capable and useful, many forms of data and information remain unquantifiable and, therefore, unusable by AI. When these unquantifiable data are also the primary subject of interest, organizations set on using quantitative approaches like AI must use whatever quantifiable data they can find – often referred to as "proxy variables" – even if they are a less accurate representation of the phenomenon of interest. This tradeoff necessarily introduces the risk of both masking administrative evil and facilitating moral inversion. Proxy variables may mask administrative evil when they are related to primary outcomes in a biased way or when they make harmful algorithmic bias harder to detect (Johnson forthcoming). The extent to which using proxy variables is problematic is usually extremely difficult to measure; if it were easier there would be no need for proxy variables in the first place. Consider the example of health. Health cannot be directly measured but needs to be operationalized. The severity of chronic conditions of a patient – one important constituent of health – are usually measured based on the medial expenditures caused by this patient (not paid by them) over a given time. Recent research has shown that this proxy variable for health masks a racial bias (Obermeyer et al. 2019). A risk-scoring AI system is used by hospitals to determine which patients with chronic conditions should be included in a special treatment program. This AI system predicted health in terms of medical expenditures accurately and fairly. But the AI system was biased in that African American patients were much less likely to be suggested for inclusion in the special treatment program than white Americans with the same underlying chronic conditions. This bias in the prediction was due to the fact that how medical expenditures are accrued differs between racial groups. An African American patient generally accrues fewer medical expenditures compared to a white American with the same underlying chronic conditions. Unfortunately, underlying chronic conditions are difficult to measure and to aggregate (because of the highly sensitive nature of the data). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | Hence medical expenditure was used as a proxy variable. Thus, **Quantification Bias:** AI can increase the chance of administrative evil by reducing the amount or quality of data brought to bear for a decision: AI requires and reinforces a belief in the primacy of quantitative data that excludes other forms of information unless they can be readily and systematically quantified.<br><br>(P18.7) The underlying motivations that drive organizations to use second-best proxy variables also expose them to a more generalized risk of facilitating administrative evil. This occurs when they choose to use AI although it might not be the best available option, or when the use of AI introduces new harms or magnifies existing ones. Such an over- or mis-use of AI in pursuit of organizational objectives is more likely to take place when the organization identifies AI as a powerful and useful tool without deep understanding of how the technology works. The technology might then not fit a given task. Similarly, an organization may not properly understand the context of the task and still rely on AI despite this lack of contextual understanding. Both of these issues can be further exacerbated when the organization faces either normative or mimetic isomorphic pressure from its peer or neighbor organizations to adopt AI solutions (DiMaggio and Powell 1983; Jun and Weare 2011). Finally, organizations might adopt AI solutions without assessing potential risks. For example, facial recognition systems are widely used despite varying significantly in accuracy depending on the shades of one's skin (Buolamwini and Gebru 2018). Nevertheless, classifications of such facial recognition systems have already been used to make false arrests (Hill 2020). Both – the tendency to adopt AI despite problems and to yield to the judgments of AI systems – are key symptoms of *AI Exuberance*. **AI Exuberance:** AI can increase the chance of administrative evil because it (1) introduces the risk of automation bias; (2) thrives within the cultural ideology of technological rationality; and therefore (3) may be enthusiastically deployed without appropriate testing or to address an issue for which AI is not the optimal available solution.<br><br>(P18.8) AI systems are perceived to be a powerful solution and may be easier to market and procure than non-AI alternatives. Moreover, because AI systems are perceived to be a very general solution that could help with many different kinds of problems, such systems will be considered in many different domains. If AI is perceived to be more powerful and more general, then this could enable an insufficient attentiveness to the requirements of a given policy issue and thereby result in the deployment of technologies that are inappropriate or even defective given the task at hand. This is a further aspect of the *AI Exuberance* proposition. In addition to the problem of masking, in terms of the theory of administrative evil, the *AI Exuberance* proposition identifies a problem of moral inversion. AI is framed as a powerful and effective new tool that symbolizes progress and expertise, when in fact it may be ineffective or even do harm. In a discussion that is immediately applicable to AI, the theory of administrative evil postulates that the invitation to do evil might come as an invitation for an expert role. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | AI in this case is the expert invited to do something only on the basis of perceived expertise, and their judgments are uncritically assigned credence on the same basis.<br>Specifically, for the purposes of AI, we understand the “scientific-analytic mindset” as one that strives for and valorizes quantification and precise numerical measurement. The “scientific-analytic mindset” may unjustifiably increase the belief that quantitative data – and proxy variables in particular – are a simple and unproblematic representation of an underlying feature that is, instead, complex, messy and vague. Because AI deals in quantification and numerical representation and feature engineering, it may increase the chance of administrative evil through masking and moral inversion. We hypothesize that similar mechanisms of masking and moral inversion may be prominent in the case of AI, and that this is as much a cultural phenomenon as it is the product of any one decision to adopt and use AI. Whereas the overlooked lesson more generally is that it is wrong to think of something “dictated by science,” in the case of AI the analogue of this idea is that it is wrong to think of something being “dictated by data.” Thus, **Quantification Bias:** AI can increase the chance of administrative evil by reducing the amount or quality of data brought to bear a decision: AI requires and reinforces a belief in the primacy of quantitative data that excludes other forms of information unless they can be readily and systematically quantified. |
| P19 | Legal contestation of artificial intelligence-related decision-making in the United Kingdom: reflections for policy (Drake et al. 2022) | (P19.1) Above all, the groups commented on a third factor as a more distant but severe constraint on individual claims: the general lack of public information. Although these fundamental transparency issues were described in various terms, participants raised three broad aspects:<br>(P19.1a) above all, for ordinary people, achieving basic awareness (disclosure of the existence of a relevant system, notification, knowing what’s happening, overcoming secrecy);<br>(P19.1b) more relevant for technical steps in bringing actions, developing understanding about automated decision-making systems (building on basic awareness, assembling meaningful information, questioning decision-makers’ presentation of systems, revisiting previous authority based on incomplete understanding); and<br>(P19.1c) generally, realising that there is the potential to contest outcomes and standard setting (both in theory and in practice).<br>(P19.2) There was a general sense that the law will gradually illuminate more such harms as time passes, |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | with legal actions playing a leading role. Nonetheless, the groups discussed four main obstacles to this gradual progression towards a more balanced approach to system capability:<br><br>(P19.2a) Intellectual property (IP). In all of the groups except Finance, it was observed that public authorities looking to implement ADM systems are often not entitled to know much about the systems they are using because of suppliers' 'aggressive' commercial confidentiality standards and associated practices. This was repeatedly observed to be a major impediment to ADM-related transparency and therefore accountability. The Constitution group especially considered this to be a major reason why relevant harms are 'very well hidden'. This logic suggests public procurement practices, including relevant contractual standards, as a potential major focus for efforts to reduce ADM-related harms.<br><br>(P19.2b) Lack of clear responsibility. ADM harms were often discussed as occurring in ways that involve both: boundaries between institutional regimes (e.g. between service audit and medical devices in Health or between assessment standards and content delivery in Education); and incentives to devolve responsibilities to machines (e.g. trying to avoid the cost of large administrative teams in Health scheduling or the potential failures of human judgement in Education grading). This was described in the starkest terms in the Criminal Justice group as 'agency laundering', or delegation of discretionary authority to technology (Fussey, Davies, and Innes 2021). The Education group observed that there are significant silos within organisations as well as between them, indicating that ADM is causing profound confusion. The Health group discussed narratives about technological implementation being used deliberately to distract from other issues, suggesting that the confusion itself can serve other purposes.<br><br>(P19.2c) Relationships with clear power imbalances. This was a particular focus for discussion in the Education group, which observed that students have no real choice but to accept the terms imposed in relation to systems like those that are increasingly used for eproctoring (e.g. 17). However the issue also appeared in the Constitution (generally in terms of decisions by state authorities), Criminal Justice (mainly in terms of surveillance but also for example in pre-trial detention decisions) and Finance (where it was doubted that consumers read or understand the terms that are offered to them).<br><br>(P19.3) The highest priority is for organisations to promote transparency, which means at a minimum notifying people that relevant algorithms are being used. Lack of public information about significant ADM is now becoming a major driver of legal contestation (see Constitution group comments above). All groups (except Finance) stated that DPIAs should be published – this was the clearest and most frequent governance recommendation from the workshop. The Health group also discussed MHRA starting to publish details of its assessments and actions as a positive step towards transparency. In the |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | Finance group, one participant suggested that banks are voluntarily moving to publish digital ethics codes because of their assessment of reputational risks.<br>(P19.4) Broader risk assessment and mitigation strategies around ADM initiatives were considered desirable, including consideration of legal as well as ethical frameworks and paying special attention to stakeholder engagement. In the Constitution group, Canada's Algorithmic Impact Assessment Tool (Treasury Board of Canada Secretariat 2021) was contrasted favourably with the 'litany of frameworks and guidelines that the UK government's promulgated'. The Criminal Justice group considered lessons from the 'West Midlands' model in some detail (Oswald 2021) and an end-to-end approach was recommended (beyond the conventional before & after assessment); it was observed to be strange that intelligence agencies are subject to closer scrutiny than the police in the UK in terms of ADM-related surveillance.<br>(P19.5) It was striking that the groups spent more time discussing the contextual processes of design and deployment around AI systems than considering the specific problem of understanding their logic. This was evident in their clear preference for the concept of ADM over AI, which contrasts with some of the priorities in computer science. Although 'black box' AI opacity (Knight 2017) was considered relevant, the focus here was on 'process transparency' (processes surrounding design, development and deployment) more than 'system transparency' (operational logic) (Ostmann and Dorobantu 2021). System transparency is important, notably in helping explanations for people working with systems in organisations seeking to deploy ADM (supporting human intervention as mandated in the GDPR). But process transparency issues extend far beyond questions of machine learning, novel forms of data or even automation, for example involving initial decisions about where and how to try to involve AI techniques in existing processes or affecting external stakeholders who may never have any direct interaction with relevant machine systems. At least if we take social failures as seriously as mechanical failures as a design consideration (Millar 2020), this suggests a need to re-evaluate assumptions and practices. It challenges assumptions about reduced reliance on the 'domain knowledge' that proved so essential for expert systems (Confalonieri et al. 2021), for example, or the idea that technology can somehow liberate decision-makers from well-established obligations of transparency and public engagement. Much of the workshop was preoccupied with basic quality standards such as taking steps to inform external stakeholders that a system exists. The discussions implied that it is useful to start from relatively simple norms, like active communications describing an algorithm, its purpose and where to seek further information (BritainThinks 2021), or systematic record-keeping (Cobbe, Lee, and Singh 2021; Winfield and Jirotka 2017). AI needs to be encased in high-quality AI-related decision-making to achieve public legitimacy. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P19.6) Law and legal processes offer legitimate channels for public participation in efforts to implement ADM and the sense that they are being marginalised in the UK is provoking resistance. The workshop discussions suggested that UK policymakers could helpfully understand legal contestation as an integral part of these trust dynamics, a part that has grown rapidly and is expected to continue to grow. Rather than promoting AI, attempts to ignore relevant legal frameworks have simply led to a morass of regulatory uncertainty and unpredictable risks.<br>Developing means for ex ante / ongoing as well as ex post legal inputs to AI-related decision processes as defined above, in other words the surrounding non-technical aspects which are important to high quality technology implementation. These should emphasise transparency and compliance-related feedback to mitigate the risk of an increasingly adversarial relationship with relevant legal functions and, instead, seek to exploit their potential to contribute to improved quality standards in ADM implementation.<br><br>(P19.7) Regrettably for public policymakers in the UK, it is the state and not the private sector that is increasingly associated with examples of apparent ADM-related harm and legal contestation.<br>Although each of these fields have the potential generate policy reflections (for example, about effective market regulation or civil service standards), it is the public private intersection at which ADM systems are being procured that the policy reflections become sharper. Although it is unclear in many of the cases in Table 1 who the ultimate decision-makers were (or what they were told or understood about the relevant ADM or other relevant systems), it is, as the Committee for Standards in Public Life observed, a practical reality that governmental authorities purchase most of their computer systems from the private sector (Committee on Standards in Public Life 2020).The idea that the government may be not only failing to regulate ADM but even exploiting internal opacities to evade scrutiny is at the heart of one of the most recent and contentious examples in Table 1. In this example of the NHS Covid datastore (16), campaigners have successfully challenged the UK government's lack of disclosure in relation to a Covid-19 analysis contract with Palantir, but it remains unclear exactly what decisions are intended to result from the processing of routinely-collected NHS data (BBC 2021b; Mourby 2020). |
| P20 | Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems (Asatiani et al. 2021) | (P20.1a) To ensure that AI systems' abilities and limitations are controlled and therefore enveloped, the DBA decided to divide its AI development into a process of incremental stages by introducing multiple small-scale solutions, each dedicated to a certain set of relatively simple and well-defined actions. Thus, from a purely technical angle, the event-driven architecture and loosely coupled systems constitute a technique in which the various components of a larger architecture operate autonomously and malfunctions are limited to local impacts only. For instance, erroneous decisions are less likely to be |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | passed onward to other systems, and if this somehow does occur, the loose coupling allows the DBA to rapidly curb the failure's escalation. Each component is therefore operating in its own envelope, and larger envelopes are created to control AI components' operation as a network. However, as highlighted by the reference above to envelopes that meet various stakeholders' needs, boundary envelopes do not serve a technical purpose alone.<br>The following extract from the data shows how important the understanding of these boundaries is for those human stakeholders that are tasked with judging the correctness of the model's operation when, for example, the complexity of the environment exceeds the model's comprehension capability:<br>*We have around 160 rules. We have technical rules that look into whether the right taxonomy is being used, whether it is the XBRL format, and whether it is compliant. We also have business rules. For example, do assets and liabilities match? Some rules only look at technical issues in the instance report. Other rules are what we called full-stop rules … filers are not allowed to file the report until they have corrected the error. We also have more guidance[-type] rules, where we say, "It looks like you're about to make a mistake. Most people do it this way. Are you sure you want to continue filing the report?" And then [users] can choose whether to ignore the rule [or not].* (Mary)<br><br>(P20.1b) In addition to the technical issues connected with accounting for multiple kinds of failure, the comment attests to boundary envelopes' social dimension. The boundaries are clearly explained to internal users at the DBA, who can overrule the models if necessary. Moreover, customer-facing models operate within an environment that has clearly defined rules constraining their operation. Wherever non-expert employees interact directly with a model, these rules are explained to them, and the human always has the power to ignore the models' recommendations if they seem questionable.<br>Thus, importantly, for every customer-facing AI model at the DBA, the final boundary envelope is a human. A decision suggested by an AI model is always verified by a case worker. In simple terms, human rationality creates a boundary that envelops the model's operation. This serves a dual purpose: it denies any model the power to make unsupervised decisions while it also makes certain that every DBA decision is compliant with legal requirements.<br><br>According to Jason:<br>*(P20.1c) The agency can be taken into court when we dissolve a company, when we end a company [forceably] by means of the law. And we, in that situation, in court, will have to provide … full documentation of why that decision has been made. Now, legally speaking, as soon as there's a human involved, as there always is, we always keep a human in [the] loop, [so we are on the safe side].*<br><br>*(P20.1d) In that context, it's only legally necessary to present that human's decision. But we want to be* |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | *able to explain also decision support, so that's why we need explainability in our model and information chain. Explainability, on the micro scale, is beneficial to understanding [the] organization on a sort of macro scale.*<br>(P20.1e) In other instances, expert case workers are allowed to set thresholds for the model in question, to make certain it produces the most useful and precise recommendations. This has a knock-on effect in facilitating DBA workers' acceptance of the relevant model:<br>*For some [of our] models, there would be some guidance threshold set by us. And then case workers are free to move it up and down.* (Susan, ML Lab)<br>*The ability to "mute" a model or change the threshold has been a major cultural factor in [the] business adaptation of this technology.* (Jason)<br>(P20.2) The crucial importance of the data used in AI systems' training is widely acknowledged in the AI/ML community. If trained on different data sets, two models with otherwise identical structure produce vastly different outputs (Alpaydin, 2020; Robbins, 2020). Accordingly, close control of the training data and the training process form an important aspect of envelopment: if the spectrum of phenomena that the training data represent is considered with care, one can better understand what the model will—and will not—be able to interpret.<br>(P20.2a) Since the DBA wants to avoid any undesired outcomes from an uncontrolled model roaming freely on a sea of potentially biased training data, the organization has decided to maintain full control over the learning process; thus, it abstains from using online-learning models, which continue learning autonomously from incoming data. This aids the DBA in protecting its systems from the unintended over-fitting and bias that less tightly controlled training data could more easily introduce. The training may be implemented in a controlled, stepwise manner:<br>*We have taken a conscious decision not to use [online-learning] technologies, meaning that we train a model to a certain level and then we accept that it will not become smart until we retrain it.* (Jason, ML Lab)<br>(P20.2b) Avoidance of models that learn "on the fly" has a downside in that models' training at the DBA is a highly involved periodic process that requires human expertise. Successful training-data envelopment therefore entails numerous stakeholders at the agency cooperating periodically to assess the needs for retraining and to perform that retraining. Paying attention to training data stimulates internal discussion of the data's suitability and of possible improvements in detecting problematic cases that are flagged for manual processing. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P20.2c) To plan retraining appropriately, data scientists at the ML Lab communicate with case workers regularly with regard to analyzing the models' performance and new kinds of incoming data. Though time-consuming, this process supports employees' mutual understanding of how the models arrive at specific results. A case worker described the effect as follows:<br>*I'm not that technically [grounded a] person, but doing that—training the model and seeing what output actually came out from me training the model…—made my understanding of it a lot better.* (William, Company Registration)<br><br>(P20.2d) Through interaction during the retraining steps, the stakeholders gain greater appreciation of each other's needs:<br>*In the company team, we would very much like [a model that] tells us, "Look at these areas," areas we didn't even think about: "Look at these because we can see there is something rotten going on here," basically. Other control departments would rather say, "We have seen one case that look[s] like this; there were these eight things wrong. Dear machine, find me cases that are exactly the same." And we have tried many times to tell them that that's fine. We had a case years ago where there were a lot of bakeries that did a lot of fraud, but now it doesn't make sense to look for bakeries anymore, because now these bakeries … are selling flowers or making computers or something different.* (Daniel, Company Registration)<br><br>(P20.3a) Input and output determine, respectively, what data sources are used to create predictions and what types of decisions, classifications, or actions are created as the model's output. Any potential inputs and outputs that exhibit considerable noise, risk of bias, data omissions, or other problems are enveloped out of the AI's operation through these decisions. The selection of input sources is thus closely tied to conceptions of data quality.<br>In the concrete case of the ID-recognition model PassportEye, the benefits of input control in conditions of poor and variable end-user-generated content became clear to the lab's staff:<br>*I think our main problem was that, yeah, we had to go a little bit back and forth because the input data was [of] very varied quality. Mostly low quality. Out of the box, PassportEye actually returned very bad results, and that reflects the low quality of the input data, because people just take pictures in whatever lighting, [against] whatever background, and so on. So we actually figured out a way to rotate the images back and forth to get a more reliable result. Because, it turned out, PassportEye was quite sensitive to angle of an image. We didn't write it [the image analysis software], so this is maybe one of the risky parts when you just import a library instead of writing it yourself.* (Thomas, ML Lab) |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P20.3b) As for output envelopment, the interplay between social and technical is more prominent here. Instead of simply preventing production of outputs that may be untrustworthy, the DBA takes a more nuanced approach. Output of appropriate confidence ratings and intervals from the models is a subject of active deliberation at the DBA. When able to verify judgments on the basis of confidence ratings, the case worker can act in an accountable manner in the interactions with DBA clients (e.g., companies that have submitted documents) and respond convincingly to their inquiries. However, sometimes it is trickier to verify the model's output unequivocally, in which case the organization strives to understand the AI model's behavior by consulting domain experts who understand the social context of the model's output. As Steven put it, "When [it is] harder to determine if the model is right or wrong, we can push the cases to the case workers and say, 'Please look at this.'" These examples of input and output envelopment demonstrate clear interplay between the social and the technical. While an opaque model is able to process a large quantity of unstructured data efficiently and produce recommendations on whether to accept or reject particular documents, this process is closely guided by case workers who rely on organizational objectives and legislative limitations to be sure the AI-produced decisions are in line with their needs. Thus, final decisions are produced at the intersection of actions by humans and AI.<br>(P20.4) Second, in terms of the sociotechnical perspective, regardless of which envelopment method they were discussing, the interviewees never spoke of a purely technical solution for limiting AI agents' capabilities. Analysis revealed that, rather than in isolation, such actions were always carried out via iterative negotiations that took into account several stakeholder views, responsibility to society, and particular implications for the personnel's work processes<br>(P20.5) First, AI's mindless and, thereby, error-prone nature necessitates careful control of the AI's agency and autonomy in the implementation. Humans can serve as important counterweights in this equation (Butler & Gray, 2006; Pääkkönen et al., 2020; Salovaara et al., 2019). The division of labor and knowledge between humans and AI can be arranged in various ways whereby organizations can balance rigidity and predictability against flexibility and creative problem-solving (Asatiani et al., 2019; Lyytinen et al., in press).<br>(P20.6) Second, organizations' AI agents interact with many types of human stakeholders, each with a particular dependence on AI and distinct abilities to understand its operation (Gregor & Benbasat, 1999; Preece, 2018; Weller, 2019). Studies indicate that AI is rarely considered a "plug-and-play" technology and that an organization deploying it requires a clear implementation strategy that takes into account the wide spectrum of stakeholders (Keding, 2021). For instance, since the impact of AI's implementation varies greatly between stakeholders, decisions to decouple stakeholders from the process of designing, |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | implementing, and using it increase the likelihood of unethical conduct and breach of social contracts, often leading to the systems' ultimate failure (Wright & Schultz, 2018).<br>(P20.7) Our interviews showed that, given the drive to improve its operations by using AI models, the DBA must devote significant attention to making sure instrumental outcomes do not come bundled with ignoring humanistic ones. Two factors have shaped the organization's quest to find balance in terms of the explainability-accuracy tradeoff: its positions as a public agency and diverse stakeholder requirements. First, as a public agency, the DBA has significant responsibility for making sure that its decisions are as fair and bias-free as possible.<br>This comment from a chief consultant on the DBA annual statements team, Mary, addresses transparency's importance:<br>*I think in Denmark, generally, we have a lot of trust towards systems …. I'm very fond of transparency. I think it's the way to go that it's fully disclosed why a system reacts [the way] it does. Otherwise, you will feel unsafe about why the system makes the decisions it does … For me, it's very important that it's not a black box.*<br>(P20.8) Sometimes inscrutable models clearly outperform explainable ones, so the agency has a strong incentive to seek ways of expanding the range of AI models that are feasible for its operations, in pursuit of higher accuracy and better performance. However, it needs to do so without incurring excessive risks associated with inscrutable models:<br>*If the output of the algorithm is very bad when using the [explainable] models and we see a performance boost in more advanced or black-box algorithms, we will use [the more advanced ones]. Then, we will afterwards check like "okay, how to make this transparent, how to make this explainable…"* (Steven, ML Lab)<br>(P20.9) Secondly, the quest for explainable AI is made even more complex by the diversity of explanation-related requirements among various DBA stakeholders. The internal stakeholders comprise several distinct employee categories, including managers, data scientists, system developers, and case workers. Externally, the DBA interacts with citizens and the companies registered in Denmark, as well as with the IT consulting firms that maintain the agency's AI models deployed in the production environment.  Each of these stakeholders requires a specific kind of explanation of a given model's internal logic and outputs. While an expert may consider it helpful to have a particular sort of explanation for the logic behind the model's behavior, that explanation may be useless to someone who is not an expert user. For a non-expert user, a concise, directed, and even partially nontransparent |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | explanation may have more value than a precise technical account. David, a case worker with Early Warning Europe, offered an example: "When [a data scientist] explained this to us, of course it was like the teacher explaining … brain surgery to a group of five-year-olds." |
| P21 | AI systems as state actors (Crawford and Schultz 2019) | (P21.1) To assess constitutional liability for private parties under the state action doctrine, courts have generally applied three tests: (i) the public function test, which asks whether the private entity performed a function traditionally and exclusively performed by government; (ii) the compulsion test, which asks whether the state significantly encouraged or exercised coercive power over the private entity's actions; and iii) the joint participation test, which asks whether the role of private actors was "pervasively entwined" with public institutions and officials. In response, we argue that courts should adopt a version of the state action doctrine to apply to vendors who supply AI systems for government decision-making. Analyzing the state action doctrine's public function, compulsion, and joint participation tests, we argue that—much like other private actors who perform traditional core government functions at the behest of the state—developers of AI systems that directly influence government decisions should be found to be state actors for purposes of constitutional liability. This is a necessary step, we suggest, to bridge the current AI accountability gap.<br>The state action doctrine should be considered as a potential pathway to providing greater accountability for the government use of AI systems. As Professor Gillian Metzger argues, "State action doctrine remains the primary tool courts use to ensure that private actors do not wield government power outside of constitutional constraints."186 As discussions of AI regulation move forward, the state action doctrine should form part of the landscape of the reasonable and appropriate regime that is ultimately devised. In particular, as AI systems rely more on deep learning, potentially becoming more autonomous and inscrutable, the accountability gap for constitutional violations threatens to become broader and deeper.<br>(P21.2) For example, a new proposed rule from the U.S. Department of Housing and Urban Development creates a complete defense to a prima facie case of housing discrimination when the defendant uses an industry-standard algorithmic model to make its housing decisions. This rule, if adopted, would encourage many actors in the housing industry to use AI systems, knowing that they could avoid liability by blaming the AI itself, even if there was overwhelming evidence that they knew the use of the system would have a disparate discriminatory impact. No doubt there will be many attempts, such as the proposed HUD rule, to allow AI systems to be used as accountability-avoidance mechanisms when companies cause constitutional violations. This is why the state action doctrine must remain a powerful and flexible common law approach for courts to use to redress this gap as it widens. This will be particularly necessary if legislation or agency regulation is slow to materialize or inadequate |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
|  |  | for the complex task that AI will present in the coming years.<br><br>(P21.3) While the plaintiffs were able to bring claims against the government to stop the ongoing deployment of the AI-driven program, the state agency lacked the capacity to address most of the specific causes of harm directly. The state had very little knowledge of how the AI software code had been written, where the mistakes were made, what data had been used to train and test it, or what means were required to mitigate the concerns raised in the case. The same is true for the *Houston Federation of Teachers* case, in which not a single employee of the school district could explain, let alone remedy, the methods or outputs of the proprietary AI at the heart of the constitutional liability concerns. Unless vendors are subject to the court's jurisdiction, the court cannot assert any real oversight or impose any specific injunctive relief on that party, even if it is in the best position to fix errors in how the AI performed.<br><br>(P21.4) For example, in the *Houston Federation of Teachers* case, none of the school district employees could provide any answers to the core substantive questions concerning constitutional liability in the case.183 Instead, all of those answers were within the technical and legal power of the vendor.184 In such cases, considering the vendor a state actor would allow courts access to the necessary information to decide cases while also directly addressing vendor trade secrecy concerns. As parties, technology companies litigate their technologies every day in courts. Allowing those who have been constitutionally harmed to sue the vendors directly would allow plaintiffs and courts to access all relevant information about the AI system, its function, and the role of the vendor in the alleged constitutional violation.<br><br>(P21.5) A range of "advocates, academics, and policymakers have raised serious concerns over the use of such systems, which are often deployed without adequate assessment, safeguards, [or] oversight."3<br><br>(P21.6) Thus, an algorithmic system itself, optimized to cut costs without consideration of legal or policy concerns, created the core constitutional problems that ultimately decided the lawsuits. These problems were also exacerbated as the result of a pattern that has emerged in which AI systems are adopted from state to state through a pattern of software contractor migration, by which AI vendors—like traveling sales representatives—usher the system from one state to another, training it on one state's historical data and then applying it to the new population.51 See id. ("Many states simply pick an assessment tool used by another state, trained on that other state's historical data, and then apply it to the new population . . . .").<br><br>(P21.7) Furthermore, "there are frequent flaws and errors in how these assessment systems are |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | implemented and how they calculate the need for care.” Id. Government agencies adopting these systems commonly enter into contracts with third-party vendors that handle everything. See supra note 26. The agency, particularly frontline staff that are most familiar with the Medicaid program and its challenges, has little to no involvement in how the tool analyzes data and produces calculations. See Litigating Algorithms, supra note 4, at 7– 8. Because these tools are often based on private systems licensed to government agencies, the design specifications and particularities of the technical system are considered trade-secrets of the vendor and are not publicly available. Id. at 8.<br>(P21.8) At the end of the day, the plaintiffs still had very little understanding of exactly how and why the AI system had reduced their benefits, and even less of an opportunity to hold accountable the private technology vendors who were primarily responsible for the harm. Constitutional accountability mechanisms in the courts inherently involve core judicial concepts such as access to the evidence of the harm57 and invocation of the court’s appropriate remedial and prophylactic powers.58 In the Arkansas and Idaho litigation, as well as their sister cases throughout the country, constitutional accountability for the creators of the AI systems responsible for the harms has been entirely absent.<br>(P21.9) The teachers sued the district through their union, arguing that the software was fundamentally inscrutable and that there was no way for teachers to know whether the software was accurately assessing their job performance. The court agreed, holding that the “teachers have no meaningful way to ensure correct calculation of their [evaluation] scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs.”61 The court based its holding on procedural due process, finding that the teachers could proceed to trial on this constitutional issue.62 The school district soon settled the case and stopped using the software.63 |
| P22 | Human rights and technology: New challenges for justice and accountability (Land and Aronson 2020) | (P22.1) As Piracés (2018) notes, crucial conversations about AI and inclusion are ongoing, but they “run the risk of being shaped by an ‘artificial intelligentsia’ that discusses inclusion without truly including the voices of the marginalized people likely to suffer most significantly in an AI ecosystem that didn’t consider their voice in its design.” Participatory design also means working with practitioners to ensure that technology meets the needs of practitioners and human rights defenders for privacy, security, and protection from harm.<br>(P22.2) Participatory processes also require transparency. People affected by automated decision-making technologies do not generally have the opportunity to review the data used for training or analysis purposes, the code underlying them, or the assumptions and value judgements embedded within them. Not only are these things difficult for the nonprogrammer to understand, companies often go to |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | great lengths to protect their code as trade secrets (Ram 2018, Wexler 2018).<br>(P22.3) These efforts make it impossible to test the accuracy of AI systems or better understand their limitations. As law professors Natalie Ram (2018) and Rebecca Wexler (2018) both argue, such oversight ability is crucial when automated systems are making life-and-death decisions.<br>(P22.4) Accountability gaps arise, for example, when governments circumvent checks on their power by outsourcing authority to private companies to censor content that they would not have been able to censor themselves (Land 2020), or when public actors are unable to provide remedies for harms caused by automated systems that they do not even understand (Crawford & Schultz 2019). In these contexts, a rigid distinction between public and private authority does not advance the underlying goals of accountability and remedy that are the foundations of human rights law. Rules regarding state action both on the international level (Land 2020) and in domestic law (Crawford & Schultz 2019), however, can be updated and deployed to ensure that states do not evade accountability for harms associated with the use of new technology.<br>(P22.5) Addressing the potential harms of new technology also requires states to take seriously the duty to protect individuals from harms by non-state actors (Karanicolas 2019–2020). For example, states might require private companies to institutionalize the practice of technology impact assessments, as well as instituting these processes themselves. Risk assessment must be integrated into systems of technological development and design, so that risks to human rights can be addressed before they are locked in. States must consider not just the obligation to promote technological innovation and access to technology but also the obligation to promote technological innovation in a way that supports rather than hinders the enjoyment of human rights.<br>(P22.6) Despite the popularity of projects that seek to deploy technology for social good, technology cannot and does not solve social problems. In a 2019 op-ed in the magazine *WIRED*, Data and Society's Mark Latonero (2019) argued that the deployment of AI for social good "smacks of tech solutionism, which can mask root causes and the risks of experimenting with AI on vulnerable people without appropriate safeguards."<br>(P22.7) Technologies must also be transparent and accessible to outside oversight and scrutiny, especially by those who bear the most risks associated with them. |
| P23 | The Moral Dimension of AI- | (P23.1) One approach is to make ethics an internal algorithm design criterion from the start. Doing so |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | Assisted Decision-Making: Some Practical Perspectives from the Front Lines (Carter 2022) | successfully may require substantial new conceptual invention in its own right, but this can be as exciting for the coding engineers as maximizing any other design feature, especially if value is attributed to it. The federal government, including the DOD, should fund basic research into ethics-by-algorithms, recognizing that companies will underinvest until some terrible wrong occurs. My experience in technology management suggests that the initial specialist refrain "it can't be done" is usually overcome by making the desired innovation a requirement-to-buy or a weighted factor in competitive source selection.<br><br>(P23.2) An additional approach is to focus on the *process* of algorithm design rather than the algorithm itself. The history of processes designed to prevent the misuse of nuclear weapons offers a valuable example. Bombs themselves are outfitted with elaborate coded locks to prevent abuse, which could have the gravest consequences. But any repair, movement, or contact–that is, any process in which bombs are handled, moved, repaired, or altered–requires two people rated in the same specialty (the "two-man rule"). Even I, as Secretary of Defense, was not authorized to be alone with a nuclear weapon. These many simultaneous approaches to security policy, some involving design and some involving process, recognized the ineffable variety of possible failure modes and the absolute necessity to prevent every one of them, all in an essentially unending custodianship of tens of thousands of bombs (the half-life of Plutonium-239 is twenty-two thousand years and Uranium-235 is 703 million years). Programs of established design principles backed up by dual or multiple checkers with equal training and qualifications and redundant safeguards are widely used in complex systems. Establishing such a design process control can not only reduce the likelihood of errors with advanced AI applications but mitigate, at least partially, the liability assigned to innovators if they do occur.<br><br>(P23.3) So the notion of requiring a process of qualified review for sensitive products is hardly new and should be the industry standard for AI. A dilemma arises from proprietary secrecy. A vendor will not want to disclose the inner workings of its algorithms and data sets; these are sensitive for competitive reasons. Given proprietary concerns, it is advantageous to establish industry-wide standards and a level of government involvement in the certification that these standards are being met.<br><br>(P23.4) Government routinely handles proprietary secrets of competing companies when it serves as a regulator or customer of advanced technology. Government security classification sometimes can be argued to slow the pace of innovation by preventing the free flow of ideas. But in the case of most AI, the preponderance of innovation is centered in companies, and intercompany secrecy is by far and away the bigger barrier to sharing information, the more so as the research frontier has moved out of universities that publish results openly and into industry. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | It is worth noting that AI itself can be a powerful tool in certification testing of AI systems whose workings are impossible for humans to fully grasp. The "checking AI" can perform an exhaustive search for oddities in large numbers of input-output runs and thereby identify design defects without unpacking the full mass of layered calculations. AI-assisted checking of algorithms can also speed up the process of ethical audit so it does not delay deployment.<br><br>(P23.5) The next thing to tackle in ethical AI is the data set the algorithm is trained on (if it is machine learning) or otherwise crunches to make recommendations to the human user. Data sets come from a wide variety of huge caches: enterprise business systems, social media, search engines, public data sets, the entire historical corpus of the written word, and Internet of Things (IoT) and sensor data of all kinds. The trickiest sets are "unstructured data": impressively large jumbles of data collected in an incidental manner. Generally, it really is true: "garbage in, garbage out." Some open-mindedness is needed, however, in the case of AI. Important hints or suggested solutions might come from running on bad data, but they should not be used for making determinations in sensitive applications.<br><br>(P23.6) An "ethical audit" of an AI database begins with its provenance. It seems well established that true anonymity cannot be promised: AI is so thoroughly penetrating that individual identity can almost always be unwound. It turns out that the risk of identification goes up in surprising ways when two databases, assembled "anonymously," are combined. There are technical approaches to enhancing privacy and true anonymity in databases used by AI that seem durable. One example is provided by the various forms of "differential privacy" in which fake data are mixed with true data in a quantified way, preserving some privacy but not entirely spoiling the data's use. The Census Bureau uses differential privacy in its data.<br><br>(P23.7) Assuming that the data sets used in AI are collected ethically to begin with, three features need to be carefully audited for inaccuracies and biases that could lead to morally fateful events when they are deployed. The audit should encompass the training set (in the case of machine learning), the application set, and potential issues in matching the two. As in the case of algorithms, an ethical case can be built on the characteristics of the data themselves and the process by which they are audited<br><br>(P23.8) To my knowledge, there is no substitute for a qualitative examination with a skeptical eye. Is the entire space of possible data points defined and is there a reasonable presence (or understood absence) of points in some corners (such as an edge subset representing a minority)? Is the set examined against a checklist of possible flaws: biased, outdated, or otherwise unrepresentative? How were the data originally tagged? Way back in the provenance of most data sets is a human tagger who originally |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | assigned a location to each point in a dataspace (“is this a dog or cat?”). And again, as in the case of algorithms, AI itself can help work through a proposed data set against a checklist of possible foibles before deployment. Finally, and again as with algorithms, the process of database audit itself can be given ethical standards: documentation, multiple qualified checkers, simulations, and sampling. |
| P24 | Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making (Grimmelikhuijsen 2022) | (P24.1) On June 25 and July 7, 2018, the City of Rotterdam used a system called SyRI (Systeem Risico Indicatie, or: “System Risk Indication”) to carry out a risk analysis of welfare fraud on 12,000 addresses in a deprived neighborhood. The risk analysis used an algorithm that was fed by 17 datasets containing personal data on someone’s fiscal, residential, educational, and labor situation. The city never published the algorithm’s parameters and decision rules, nor were investigated residents informed they were investigated for welfare fraud. Residents and activists protested and finally, in 2020, a Dutch Court prohibited governments to use SyRI. A core reason for this, according to the verdict, was a lack of transparency of the algorithm used by this system.<br><br>(P24.2) First, a new generation of algorithms uses techniques to detect patterns in data using only inputs (e.g. a training dataset provided by humans). How certain patterns and outputs based on these input data are generated has been referred to as an algorithmic ‘black box’. Such algorithms are not readily understandable to humans, making them unexplainable to citizens (Burrell 2016).<br><br>(P24.3) Second, algorithms are sometimes deliberately made inaccessible as they are often developed by commercial parties and subject and protected by intellectual property. Other algorithms are not accessible because governments fear that citizens subject to those algorithms will game the system once they have figured out how it works (Mittelstadt et al. 2016).<br><br>(P24.4) First, we consider the accessibility of algorithms.  Accessibility is an important issue because often the source code or model is not available to outsiders as it is considered the intellectual property of the company that developed it (Mittelstadt et al. 2016). Also, the algorithm might be kept inaccessible for privacy reasons or purportedly to prevent users from gaming the system (Burrell 2016; Kitchin 2017).1 This is risky since inaccessible algorithms are inscrutable and can more easily produce decisions or recommendations that are biased, discriminatory, or inaccurate.<br><br>(P24.5) Accessibility of algorithms goes beyond being accessible to the public since to most of the public it will be unclear what they are looking at when they see code (Annany and Crawford 2018). Even experts struggle to understand what software code will do in practice, as the source code only gives very limited information to predict how a computer system will behave in practice (Kroll et al. 2016). Hence, transparency as merely having a publicly accessible algorithm will not be sufficient to |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | improve an algorithm's trustworthiness. Therefore, on top of transparency-as-code-availability, scholars have argued for algorithms to be audited by an independent auditor (Tutt 2017) or to include technical tools that are incorporated in a system's design to ensure an algorithm complies with legal procedures and standards (Kroll et al. 2016). In this paper, accessibility means not just public availability, but accessibility means that external independent experts can access an algorithm for inspection and analysis to assess if it is compliant and does not violate any rules.<br>(P24.6) Finally, accessibility does not only concern the algorithm itself, but also the underlying data. Using single or combined high-volume structured and unstructured datasets ("big data") algorithms can detect new patterns. However, linking various fine-grained datasets has led to privacy concerns (Mergel, Rethemeyer and Isett 2016). Furthermore, a large body of research had pointed out that many datasets are biased, which leads to biased predictions of an algorithm (Eubanks 2018). To be able to assess an algorithm the quality of underlying data, these must be accessible to independent experts.<br>(P24.7) The second core element of algorithmic transparency is explainability. A lack of explainable models and outcomes is at the heart of the debate on the algorithmic black box explained earlier (Burrell 2016; Lepri et al. 2018; Miller 2019). Different techniques have been put forward to increase the explainability of AI algorithms.2 They could be developed to be inherently transparent in their decisions (Rudin 2019), or if the algorithm is a black-box, then various explanation techniques have been developed, such as XAI (Miller 2019). Algorithmic explainability can take different forms and the type of explanation that is needed depends on the user of the algorithm. In this article, we focus on simplified explanations that can be understood by a non-expert (citizen) audience. Such an explanation should, for instance, provide the reasons for what was the crucial variable that contributed to an algorithmic outcome (Friedrich and Zanker 2011; Kizilcec 2016). Citizens want to receive clear reasons for a (negative) decision to assess the fairness of a decision (De Fine Licht and De Fine Licht 2020; Tyler 2006) or to be able to contest an algorithmic decision (Mittelstadt et al. 2019)<br>(P24.8) For a person to trust decision-making when he or she is subject to a negative decision outcome—e.g. s/he has been denied benefits—it is crucial that the procedure of a decision was fair. In the case of welfare fraud detection, the consequences of the decision are much more pervasive than in the visa application case. This aligns with findings from the procedural justice literature, where scholars have found that perceived fairness is especially relevant for trust in authority when the decision outcome matters more to the recipient (Grootelaar and van den Bos 2018). In the event of such a negative decision outcome, people will want to seek information that helps them to interpret the situation (Fiske and Taylor 1991). |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P24.9) Algorithmic transparency taps into various elements of procedural fairness. Explaining an algorithmic decision will—ideally—show that a decision was taken in an unbiased and well-considered manner. Furthermore, the accessibility of an algorithm can be an important cue signaling openness. Therefore, we expect that transparent algorithms—algorithms that are accessible and explainable—improve their trustworthiness among the general public. |
| P25 | Rethinking AI for Good Governance (Margetts 2022) | (P25.1) The use of prediction to deliver individual (as opposed to aggregate) risk scores is much more controversial. For local authorities that have used predictive techniques to identify the number of children that are likely to be at risk of abuse or neglect, the next step from forecasting (say) demand for childcare places is likely to be "which children?" Such a question would come naturally to social services departments terrified of being held responsible for the next ghastly case of abuse to hit the headlines, the next "Baby P." But should a technique that is essentially inductive be used in this way? A risk of 95 percent of being a victim of an abusive incident means that there is still a chance that the event will not happen, and if the figure is 65 percent, the meaning of the individual number is highly ambiguous. Social policy experts who advocate this kind of machine learning for decision support have built models to support childcare workers' decision-making in New Zealand, the United States, and Australia. But other studies have counseled a more cautious and thoughtful approach, and noted the importance of the data environment. The most feted version, in Pittsburgh, was built from a data-rich environment providing a 360-degree view of all children's and their families' interactions with state agencies throughout their lives, an environment that rarely exists in local authorities. And such systems are extremely vulnerable to bias, especially where data are derived from the criminal justice system.<br>(P25.2) Users of digital platforms know very little about the operation of search or newsfeed algorithms, yet will rightly have quite different expectations of their right to understand how decisions on their benefit entitlement or health care coverage have been made. The opaqueness of AI technology is accepted in the private sector, but it challenges government transparency.<br>(P25.3) If we do not design appropriate accountability frameworks, then politicians and policy-makers will take advantage of this blame-shifting possibility. This will range from cases like the UK prime minister blaming poor statistical processes to calculate public examination results after school closures in the 2020 pandemic prevented exams from taking place as a "mutant algorithm," to the more nuanced and unconscious shifting of responsibility to statistical processes involved in judicial decision-making with AI observed above. A public sector AI in which fairness, accountability, and transparency are prioritized would be viewed as more trustworthy, working against such perceptions. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P25.4) First, governments could prioritize the development of expertise and capacity in AI to foster innovation and overcome some of the recurring challenges. As noted above, the history of government computing has been characterized by large-scale contracting to global computer services providers, but AI does not lend itself to this kind of outsourcing, whereby governments lose control of key features. For example, the U.S. CBP was criticized in 2020 for being unable to explain failure rates of biometric scanning technology "due to the proprietary technology being used."33 Similar issues have dogged the adoption of facial recognition technologies by police agencies, with moratoria announced in several cities. There is evidence that government agencies realize the importance of developing capacity: the same U.S. study also found that "over half of applications were built in-house, suggesting there is substantial creative appetite within agencies."<br>(P25.5) Finally, perhaps the most ambitious use of AI would be to tackle issues of equality and fairness in governmental systems in a profound and transformative way, identifying and reforming long-standing biases in resource allocation, decision-making, the administering of justice, and the delivery of services.<br>(P25.6) Many of the causes of bias and unfairness in machine learning, for example, come from training data generated by the existing system.<br>(P25.7) Data and modeling have made these biases and inequalities explicit, sometimes for the first time. Some researchers have suggested that we might develop AI models that incorporate these different sources of data and combine insights from a range of models (so-called ensemble learning) aimed at the needs of different societal groups. Such models might be used to produce unbiased resource allocation methods and decision support systems for public professionals, helping to make government better, in every sense of the word, than ever before. |
| P26 | Alexa, Is This a Historical Record? (Venkata et al. 2022) | (P26.1) The current project **AI for Selection** deals with the selection issue exclusively. It aims at studying ML approaches and assessing the existing tools in the market to aid government departments in the selection process. It is perceived that any use of ML would aim to reduce the manual burden on RMs, still allowing them to retain the final decision on permanent preservation<br>(P26.2) While each algorithm's performance was evaluated using standard accuracy metrics it is important to understand that these are not necessarily representative of a product's performance in general. The dataset was highly curated and belonged to one relatively small government department. Good performance against this dataset does not imply that equal performance would be achieved against the documents of a larger, more complex department. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P26.3) Product A. Following training, unlabelled documents are imported. De-duplication and named entity extraction are automatically applied as documents are loaded, with the extracted entities being used as features for the ML. The newly loaded documents are given suggested labels by the ML classifier but they are not final until they have been approved or corrected by the RM. The model can be iteratively re-trained as the user works through this process.<br>(P26.4) This section details the development of the benchmark tool, *product T*, and discusses the general methodology, data pre-processing, and ML training and evaluation. It was created using the Python programming language and free open source ML libraries, with a command line interface (CLI) rather than a GUI. Although the solution requires technical skill to use it, the problem and resulting ML pipeline were modelled from a RMs perspective.<br>(P26.5) To test the effectiveness of an ML model it is trained with one set of data and then tested with a second. Since only one dataset was provided, two lists of record identifiers were sent to suppliers to standardise the training/test split. In a competitive scenario the labels for the test set would have been held back from the suppliers. Suppliers were also asked to use 10-fold cross validation, which generates 10 training and validation sets and averages the results. Cross validation shows whether algorithms generalize well when presented with new data. How to split the data with each iteration is important too. Since the classifications were highly imbalanced a weighted approach to splitting data was used to make sure every class was represented in the training data. The danger of not doing this is that if a class does not appear in the training data it will be unknown to the algorithm.<br>(P26.6) *Enduring Value.* A problem with moving to ML solutions is adapting processes for accurately assigning enduring value, which has often been a subjective and shifting classification. In terms of digital strategy practice at TNA, much would still be recognisable to Jenkinson [7, 20]. The government departments, as the creators of the records, select material based on two broad criteria: “the documentation of what government did, why and how” and “the value of the records for future historical research” [5]. Departments create and publish appraisal policies that outline how decisions are undertaken and what material will be selected [4].<br>(P26.7) *Data Quality.* It is clear from what we have seen that the quality of the training data is critical to good results. While feature engineering and model tuning are an important aspect of optimising results, *supplier B* reported that selection of training data had the biggest influence on accuracy. Even those products promising a user experience geared toward a RM still embedded data scientists in their project |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | teams suggesting we are still a long way from not requiring skilled data professionals in the process. This point is emphasised by the fact that every supplier began with a phase of data analysis and data visualisation, functionality that was not built into the tools themselves but is recognised as an essential part of the ML pipeline.<br>(P26.8) *Training models.* Traditional software products are designed from a set of requirements gained through user engagement. The expert systems in the 1990s were created by eliciting logical rules from experts. But ML is different. Training a supervised learning algorithm for selection means trying to imprint the context of an appraisal policy via the training data. Addressing concerns in data preparation is not a concern unique to archives, Jo et al. suggested that ML pipelines could be enhanced by improving data collection by using some of the methods undertaken by archives for appraisal, namely, recording the "process of data collection" and relying on multilayered, and multi-person systems rather than "a single ML engineer" when compiling a dataset [21].<br>(P26.9) Our view is that records management teams should be equipped with tools for data mining and analysis, and the requisite skills to use them. They must go through the process of curating training sets of documents that effectively represent the rules and processes they follow when selecting documents, before jumping into ML. To the uninitiated the ML may sound difficult and complex but it is not, as the data is the hard part rather than ML. At TNA, we have begun a series of internal workshops to explain and demystify the ML process for non-technical staff, and we would recommend similar training for government departments [8].<br>(P26.10) *Disorganised data.* In real-life applications data often comes disorganised, like the unlabelled data we received for the project (refer to Section 4). This unlabelled data suffered from two issues. First, document folders are nothing but the containers for a mixture of file types, and non-standardised file structures belonging to various departments dumped together. This scenario is relatable to common shared drive data in any organisation. It needs a thorough cleaning of data by data specialists and domain experts working together. Domain experts are needed to use their knowledge to provide some order and to identify high-level document features that can aid selection. Data specialists need to perform *feature engineering* to extract and generate features from the documents at scale to be fed into a ML model.<br>(P26.11) The second issue is with data representation. While performing classification with machines, it is necessary to have data points with good representation in each category. Data are said to suffer the *class imbalance problem* when the class distributions are highly imbalanced [24] in large data volumes. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | In our document corpus, the class imbalance is shown in the Figure 1.While there is a majority representation for categories: 02, 04, 05, 23, 24, and 33, there are hardly any files for classes: 06, 10, 24a, 25. Such huge class imbalance induces *bias* in the decision making. A serious problem in the given data is retention category 06, which should be categorised as "Selected." But with a negligible representation in the training data, the ML model might fail to recognise that category at all! As a result, there is a chance that all documents in that category will be misclassified as "Not selected." Theoretically, there are few ways to handle the imbalance problem [32, 35, 38].<br>(P26.12) What these methodologies have in common is that modelling requirements and understanding data are a large part of the process and are not ready to be automated. Additionally, the evaluation phase and on-going monitoring require data science skills, otherwise the automl is effectively marking its own homework. Highlighting the complexity of putting a ML pipeline into production, Breck et al. have proposed a 28 step ML readiness rubric [10]. Ongoing monitoring is a part of the rubric that includes tests for model "staleness," which can occur when the distribution of incoming data changes over time. One of the products we tested included the facility to weight training data according to its age.<br>(P26.13) *The context of decisions made.* Understanding the decisions made around selection of records is important not just to RMs who make selection decisions but also to the long-term understanding by future researchers of the collections. Dunley states, that to understand the archive fully you must understand the processes through which "the selection and preservation of 'valuable' information occurs" [14]. Understanding existing decisions relies on analysis of appraisal policies and procedures. The addition of ML approaches adds additional complexity to this understanding. Manoff has listed concerns around 'impenetrability of machine processes and algorithms' in her study assessing the potential for technology to create areas of "archival silence" [25]. While algorithms do not have to be impenetrable some are, especially proprietary ones using automated ML approaches. It could also be argued that human-based selection that involves interpretation of policies is also impenetrable; we are not always able to request the reason behind a record's selection.<br>(P26.14) *Explainability.* As ML makes more decisions affecting our lives, the issues of transparency and **explainable AI (XAI)** have come to the fore in the fields of Human Computer Interaction and AI. Following a workshop on Human-Centred Explainable AI, Bunn offers some reflections on XAI from a records management perspective [12]. Algorithmic decisions can be explained at the level of the model (e.g., how does it function? What assumptions does it make?) or at the individual record level (why was this record classified as A rather than B?). The first type of explanation can be achieved through transparency by using open source code and through documentation. |

| Paper ID | Title & Author | Raw Data (Extracted Text Fragments) |
|---|---|---|
| | | (P26.15) As well as technical skills there also needs to be an educational programme for non-technical staff to understand the concepts of ML and their role in creating the data to train the system.<br>(P26.16) Record selection is about enacting a policy and requires knowledge of the collections and events in the outside world. This requires experts leveraging technology to make their job possible, rather than relying on it to perform the selection task for them.<br>These tools are labelling at the document level, not at the folder level, and do not take into account the context of documents as a collection within a series or folder. Maintaining human control over the process can allay fears of machines making decisions.<br>(P26.17) Transparency is vital and more work is needed to explain why algorithms are making decisions and align them with policies. The processes and rationale behind training data curation should be published to help identify potential biases.<br>(P26.18) The good news is that all of the products we saw were in early stages of development, which means this is a great time to engage with suppliers and influence their future development so that they work for RMs and archivists. |