Journal of Behavioral
Public Administration
Vol 6, pp. 1-19
DOI: 10.30636/jbpa.61.261

Research Article

Public Administration Meets Artificial Intelligence: Towards a Meaningful Behavioral Research Agenda on Algorithmic Decision-Making in Government

Saar Alon-Barkat[†], Madalina Busuioc[‡]

Abstract: We propose a novel research agenda for public administration regarding one of the most important developments in the public sector nowadays: the incorporation of artificial intelligence into public sector decision-making. These developments and their implications are gaining increasing attention from public administration scholars. We argue that public administration research in the behavioral strand is well-positioned to contribute to important aspects of this debate, by shedding light on the psychological processes underlying human-AI interactions and the distinct set of biases that arise for decision-makers and citizens alike in these algorithmic encounters. These aspects thus far remain considerably under-investigated in bureaucratic contexts. We develop theoretical propositions on two core behavioral aspects pertaining to algorithmic interactions in bureaucratic contexts. Namely, the first theme refers to the interaction between *public sector decision-makers and algorithms* and related cognitive biases arising in decision-making. The second theme pertains to the interaction between *citizens and algorithms* and how this affects citizen experience of, and responses to, government decision-making in its algorithmic iteration. A meaningful behavioral research agenda on the incorporation of AI in government requires taking the study of algorithm-related cognitive biases seriously in this context and the organizational and institutional challenges that can arise as a result.

Keywords: artificial intelligence; automation bias; algorithmic burdens; citizen trust; algorithmic transparency; public accountability

Introduction

Artificial intelligence (AI) algorithms are becoming a pervasive feature of administrative governance, in the process transforming public bureaucracies in non-negligeable ways. These developments are constitutive not only of administrative structures and routines (Meijer et al. 2021) but also shape the very nature of bureaucratic discretion (Young et al., 2019; de Boer & Raaphorst, 2023), public accountability (Busuioc, 2021) and fundamentally, citizen-state interactions (Peeters, 2023; Whiteford, 2021; Alon-Barkat & Busuioc, 2023). The aim of this contribution is to bring a public administration behavioral research agenda to the center of the study of AI in government and its implications, one that takes the study of bias seriously in this context.

Behavioral research in public administration has produced new insights for our discipline over the past decade. Applying theories and methods from psychology and neighboring disciplines in the context of government bureaucracies, it has contributed to our understanding of decision-making processes and citizens' attitudes to public policies and organizations (e.g., Jilke, Van de Walle & Kim, 2016; James, Jilke & Van Ryzin, 2017; Moynihan, 2018). A main theme of research in this strand is the exploration of cognitive biases that shape information processing and decision-making in these contexts.

The incorporation of artificial intelligence algorithms across public sectors is one of the most consequential developments in the public sector nowadays. Decision-making in public organizations is

† School of Political Sciences, University of Haifa, Haifa, Israel, ‡ Department of Political Science and Public Administration, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
Address correspondence to Madalina Busuioc at (e.m.busuioc@vu.n).
Copyright: © 2024. The authors license this article under the terms of the Creative Commons Attribution 4.0 International License.

undergoing a profound transformation, with information processing as well as countless decisional aspects increasingly delegated to automated tools, drawing on new generation algorithmic approaches like machine learning. These developments call for robust investigation as to their implications and effects. We argue that behavioral research in public administration is well-positioned to contribute to important aspects of this debate. As artificial intelligence permeates public administrations across jurisdictions, an agenda converging on the study of psychological processes and cognitive aspects pertaining to our interactions with the technology in bureaucratic contexts, and the unique set of biases and challenges that this can give rise to, gains especially high importance.

To that end, we bring together and build upon insights from different strands of relevant literatures namely, social psychology research on automation, data science and behavioral public administration. We lay concrete building blocks for a coherent research agenda around key issues as well as provide initial theoretical directions for exploration. We focus our attention on two core aspects pertaining to algorithmic interactions in bureaucratic contexts. The first theme regards *the interaction between AI algorithms and public sector decision-makers*, namely how decision-makers incorporate algorithmic advice into their decision process, and specific *cognitive biases* that arise at this interface. The second theme regards *the interaction between AI algorithms and citizens*, namely, how affected individuals and the public at large experience and assess algorithm-based government decisions and respond to them, especially so given their often-opaque nature.

These cognitive-related aspects brought on by the adoption of AI remain considerably underexplored in bureaucratic contexts. We argue that this is a significant oversight and that through its experimental outlook, natural synergies to psychology and traditional affinity to some of these topics (albeit so far largely outside an AI algorithmic context), a behavioral public administration approach can contribute in a meaningful way to highly relevant aspects of this debate, which we outline below. By accounting for key aspects pertaining to the complex nature of the public sector and broader connections to normative questions of governance and democracy in relation to AI, a public administration lens yields distinct insights as to the purported implications AI adoption stands to have, bringing a much-needed reflective contribution to the debate. In particular, we take issue with, and push back against, an influential yet highly problematic discourse (in behavior work outside our discipline) that would portray AI as having neutralizing effects as a supposed "fix" to human bias, despite mounting evidence to the contrary across bureaucratic contexts.

On the Rise: Algorithmic Decision-Making in the Public Sector

Algorithmic decision-making in the public sector is both ubiquitous and on the rise (Yeung & Lodge, 2019; Veale & Brass, 2019; Ferguson, 2017; Eubanks, 2018; Margetts & Dorobantu, 2019; Engstrom, Ho, Sharkey & Cuéllar, 2020). What sets artificial intelligence algorithms apart from their predecessors is that these new generation algorithms, such as machine learning (ML) algorithms, learn the rules that govern their behavior from data and can often be notoriously opaque in their functioning. By virtue of the sheer size of the parameter space and complexity of feature interactions, understanding the decisional pathways of how powerful ML models such as neural networks arrive at their predictions can be highly challenging (also for their developers). This associated opacity constitutes a distinctive and defining informational problem pertaining to AI systems that fundamentally differentiates them from earlier algorithmic precursors relied upon by governments in the past (Busuioc, 2021).

At the same time, AI algorithmic systems are increasingly adopted in a host of policy domains involving life-changing decisions. For example, they are relied upon in law enforcement (Ferguson, 2017; Richardson et al., 2019); criminal justice, as decisional aides in bailing and recidivism (Citron, 2016); or welfare, among others, for fraud detection (Gilman, 2020; Geiger, 2021) or to detect child neglect and cases requiring intervention (Eaton, 2019; Ho & Burke, 2022). Testimony to the ubiquity of AI tools in the public sector, a study in the US context found that "nearly half of all agencies use, or are investigating the use of, artificial intelligence" (Calo & Citron, 2021, 801; see Engstrom, Ho, Sharkey & Cuéllar, 2020 for an overview).

These developments are driven by the promise of policy solutions that are clamored to be more effective, efficient and low-cost, especially appealing in a context of resource pressures on public services. In

addition, and importantly, AI algorithms are said to come with the promise of "neutrality" i.e., as means to neutralize or bypass human cognitive and social biases (e.g., Sunstein 2019, 2021). They are claimed to bring about more accurate and reliable predictions, allowing to bypass human limitations, with leading behavioral economics scholars explicitly arguing for their adoption to overcome human biases and the inconsistencies of human judgment (or "noise") (Kahneman, Sibony & Sunstein, 2021; Sunstein, 2021).

This argumentation is, to a large degree, a continuation of the behavioral economics and "nudge" agenda on the limitations of the human mind and emphasis on a rationality approach to dealing with human cognitive biases and shortcomings (Kahneman, 2011; Thaler & Sunstein, 2008). In this context, algorithms represent the latest solution purported by literature in this strand to do away with the weaknesses of human decision-making, this time with AI as a computational fix that is gaining high traction in practice. For instance, the appeal of algorithms in policing has, to an important degree, been ascribed to perceptions that their computational nature could help bypass human biases and discrimination (Ferguson, 2017), albeit algorithms themselves are necessarily produced by humans and trained on human data, with all its bias and noise.

At the same time, a different strand of research, psychology research on automation has been concerned with automation itself being a potential source and trigger of human biases (e.g., Skitka, Mosier & Burdick, 1999; Logg, Minson & Moore, 2019). This work, as we see below, rather than envisaging automation as a solution to human bias, has highlighted precisely its potential to give rise to distinct biases, errors and information processing challenges in decision-making, brought on (rather than alleviated) by automation (Skitka, Mosier & Burdick, 2000, 701).

The current realities of AI algorithmic deployment certainly seem to reflect the latter scenario with biases in algorithmic decision-making cropping up not only in human processing of algorithmic inputs but also embedded within AI systems themselves. A legion of far-ranging failures and harmful outcomes linked to AI deployments in public sectors expose the fallacy at the heart of default assumptions of AI functionality (Raji et al., 2022). Rather than "neutral", AI algorithms can come to reproduce human biases, to no small degree due to the way they learn. AI algorithms are a product of human data: they draw on historical data as their training data to learn the rules that govern their behavior. As a result, they can come to encode into their technical DNA unfair and discriminatory practices present in the data (in addition to the biases and heuristics of their designers), with resulting algorithmic models standing to reproduce and automate human biases and discrimination (O'Neil, 2017; Barocas & Selbst, 2016; Eubanks, 2018). Accordingly, while heralded by some authors (and providers themselves) as potential debiasing or neutralizing tools, paradoxically, AI algorithms have been found to replicate and reinforce on a large scale some of the same biases they were meant to remedy. This time however, under the guise of objectivity and with blurred responsibility and accountability for resulting outcomes. Growing evidence points to the need to take the study of automation and the incorporation of AI in government seriously, including accounting for its behavioral aspects and effects.

AI Algorithms and Public Sector Decision-makers: Human Processing of Algorithmic Advice

The appeal of algorithmic decision-making in the public sector has stemmed as noted above, from the perception that algorithmic systems, as computational devices, can overcome human cognitive biases and weaknesses. Yet, inescapably, algorithms are made by humans (from the choice of model, training data, fine tuning, hyperparameters, and the list goes on) and are trained on human data (which reflects our social biases and behaviors). In other words, algorithms learn from us, warts and all. What is more, for a large part, algorithms in the public sector do not make decisions on their own without a human intervener but serve as input to human decision-making.

For instance, ML-based recidivism models in criminal justice are predictive decision support systems, whose algorithmic outputs – i.e., scores quantifying individual risk of recidivism – judges are presented with to inform their bailing and sentencing decisions. Similarly, in policing, predictive systems serve as inputs to human decision-making on police activities and resource allocation. They are hybrid decision systems. In fact, in some jurisdictions, legal constraints (see the EU General Data Protection Regulation) outright prohibit the

reliance on solely automated decision-making, mandating human mediation in algorithmic decision-making. In other words, algorithmic decision-making in the public sector arises, to a considerable degree, in the interaction between humans and algorithmic systems.

This points to the pressing need to thoroughly understand specifically how *administrative decision-makers (civil servants) process* and *respond to AI outputs* and recommendations and their effects in this respect. Extant psychology literature on automation and behavioral public administration research on information processing provide us with valuable starting points to systematically unpack the proposed theme. Importantly, contrary to the promise of neutrality that has propelled the use of AI across public sectors, both literatures would clue us in to the ongoing relevance of cognitive biases, albeit in different – and somewhat diverging – forms.

Automation Bias: Automation as a Source of Cognitive Bias in Decision-Making

Evidence from social psychology literature suggests that automation might induce distinct types of bias, arising from human processing of automated outputs. Notably, research in psychology on automated support systems (which precede AI) has shown that individuals are susceptible to "automation bias" or default deference to automated systems (Parasuraman & Riley, 1997; Skitka, Mosier & Burdick, 1999; Skitka, Mosier & Burdick, 2000; Skitka, Mosier, Burdick & Rosenblatt, 2000; Mosier et al., 2001; Cummings, 2006; for a systematic review, see: Lyell & Coiera, 2017). The potential sources of automation bias are said to stem on the one hand, from the belief in the perceived inherent authority or superiority of automated systems and on the other, from "cognitive laziness", a reluctance to engage in cognitively complex mental processes and thorough information search and processing (Skitka, Mosier & Burdick, 2000).

The phenomenon has been documented in experiment-based psychology studies in areas where automated tools have been relied upon for a long time such as aviation and healthcare. A systematic review of studies within this literature concluded that the tendency for automation bias "appears to be a fairly robust and generic effect across research fields" (Goddard, Roudsari & Wyatt, 2012, 123). There is also anecdotal evidence thereof ranging from pilot error arising from blind deference to automated systems (Skitka, Mosier & Burdick, 2000, 703) to drivers slavishly following their GPS nearly off a cliff or into the ocean (Milner, 2016). Similar patterns of over-reliance on automated systems are also anecdotally reported in the context of self-driving cars (National Transportation Safety Board, 2017) or in medical contexts (Medium – Open Letter Concerned Researchers, 2019).

More recently, and in line with automation bias literature, studies investigating specifically AI-supported decision-making have similarly documented a tendency to assign greater weight to algorithmic advice (compared to human advice) in various contexts (Logg, Minson & Moore, 2019; Gunaratne, Zalmanson & Nov, 2018; Liel & Zalmanson, 2020; Hou & Jung, 2021; You, Yang & Li, 2022). Greater reliance on algorithmic advice is reported across various tasks and domains, again stemming from algorithms' perceived superior predictive performance.

Interestingly, at the same time, related literature in this strand in business management, focused on the private sector, points not only to disproportionate trust in algorithms, but also, under specific circumstances, to disproportionate aversion when decision-makers experience algorithms fail. Accordingly, these studies suggest people tend to be less tolerant when they observe an algorithm make mistakes and to more quickly lose confidence compared with similar human errors (e.g., Dietvorst, Simmons & Massey, 2015, 2018; Renier, Mast & Bekbergenova, 2021; Burton, Stein & Jensen, 2020). In other words, users "punish" algorithms more for observed failings compared to analogous errors by humans (Logg, Minson & Moore, 2019). This tendency is possibly at least in part due to being faced with evidence that contradicts unrealistic starting perceptions of algorithmic superiority and infallibility (Maasland & Weißmüller, 2022).

All in all, according to these strands of literature, algorithmic inputs potentially *create biases* of their own that arise, this time, in the *human-algorithm interaction*, effectively replacing one set of biases with another.

The few studies on the topic in the public sector context have yielded mixed evidence with regards to decision-makers' tendency to defer to AI. While several studies in the context of criminal justice do not provide clear evidence for default deference to algorithmic risk assessments (Green & Chen, 2019a,b; Stevenson, 2018; Grgić-Hlača, Engel & Gummadi, 2019), a recent study finds that people tend to have remarkably high trust in algorithmic risk assessments, compared with human expert assessments (Kennedy,

Waggoner & Ward, 2022). Respondents were highly trusting in algorithm performance even when explicitly told that algorithms do not perform better than lay humans or were ill-suited for forecasting in the task at hand. Recently, Alon-Barkat & Busuioc (2023) explicitly tested automation bias in an administrative context through a series of survey experiment studies examining participants' inclination to follow algorithmic versus equivalent human-expert advice, in light of contradictory external evidence. They did not find significant differences between the two conditions. At the same time, the authors emphasized it is too soon to rule out concerns with automation bias, which could become more prevalent as decision-makers become accustomed to (and trusting of) the use of AI in bureaucratic contexts over repeated interactions.

Concerns with automation bias or undue deference to AI algorithms have been also recurrently voiced by scholars in a context of growing reliance on AI tools in the public sector and high-stakes scenarios (e.g., Citron, 2008, 2016; Cobbe, 2019; Finck, 2020; Veale, Van Kleek & Binns, 2018; Yeung, 2019; Peeters. 2020). Such concerns are also of high practical relevance: "Automation bias" is explicitly mentioned as a key concern for human oversight of AI in the context of the AI Act set to regulate AI EU-wide. The Act explicitly requires providers of AI systems to develop such systems to ensure that individuals assigned with oversight "remain aware of the possible *tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)*"(Article 14(4)(b), draft AI Act, COM(2021) 206 final).

These cumulative insights raise important questions that need to be systematically prodded, tested (and reconciled) with respect to reliance on *AI algorithms* and implications thereof – with an *explicit public sector focus*. Concretely, they suggest algorithmic systems have the potential to introduce *distinct cognitive biases* for human decision-makers brought on by the interaction or exposure to automation.

From a public administration perspective, and given the high-stakes of AI public sector use, concerns with "automation bias" are especially relevant from a normative standpoint. While some authors proclaim positive effects associated with automation pertaining to less "noisy" or more uniform / standardized decision-making, public administration scholars have raised concerns precisely as to negative effects as automation supplants professional judgment and discretion in decision-making (Bovens & Zouridis, 2002; Zouridis, Van Eck & Bovens, 2019; Young, Bullock & Lecy, 2019; Peeters, 2020; Young, Himmelreich, Bullock & Kim, 2021). The latter also echoes early calls in our field warning about the potential of AI for "atrophying administrators' own judgment and sense of responsibility" (Barth & Arnold, 1999), the risk of "overly passive or deferential human beings" through the abdication of de facto decision-making to automation. As individuals are aggregated into predictive categories, "digital rigidity" and lack of affordances to special circumstances and individual considerations can become a source of injustice in its own right (Bovens & Zouridis, 2002). Research on street-level bureaucracies demonstrates that bureaucratic discretion can precisely play a critical function in navigating administrative rigidities, with frontline workers serving a bridging function between citizens and the administration, which automation risks doing away with.

Biased Information Processing by Bureaucratic Decision-Makers

In addition to automation bias, recent work points to a new, and somewhat diverging, concern: a human predisposition for so-called "selective adherence" to algorithmic advice (Alon Barkat & Busuioc, 2023). Namely, decision-makers display a propensity to follow algorithmic predictions selectively when these confirm pre-existing beliefs i.e., when predictions correspond to what they already believe to be true and reinforce stereotypical views. Accordingly, algorithms can end up being relied upon in ways that justify and reinforce existing biases.

These patterns are consistent with an established line in behavioral public administration literature which shows that even professional bureaucratic decision-makers are susceptible to various cognitive biases shaping the way they respond to new information and incorporate it into their judgements and decisions (Battaglio, Belardinelli, Bellé & Cantarelli, 2019; James et al., 2020). Motivated reasoning research shows decision-makers tend to search for, accept, and place greater weight on information that aligns with their prior beliefs and to contest and disregard evidence inconsistent with these beliefs (Baekgaard et al., 2019; Baekgaard & Serritzlew, 2016; Christensen et al., 2018; James & Van Ryzin, 2017; Jilke, 2017; Jilke & Baekgaard, 2020). These tendencies serve to protect peoples' positive self-conception and to rationalize their behavior and

predetermined positions while maintaining an "illusion of objectivity". Such patterns are established in the context of processing of non-ambiguous performance information.

Alon-Barkat and Busuioc (2023) theorize similar tendencies at play with respect to the processing of *algorithmic information*, leading to similar selective patterns of information adoption. They further theoretically link this mechanism to bureaucratic discrimination literature, which points to the tendency of civil servants to be unconsciously affected in their decisions by prevalent group stereotypes, which can result in discrimination of disadvantaged citizens (e.g., Jilke & Tummers, 2018; Andersen & Guul, 2019; Assouline, Gilad & Bloom, 2022). Altogether, these literatures raise concerns that decision-makers are likely to process information in a selective or biased way, depending on whether it is compatible with their prevailing beliefs and group stereotypes.

This tendency has been empirically illustrated by several studies in computer science and in law, in the context of algorithm use in the US criminal justice system, building on survey experimental designs with online panelists (Green & Chen, 2019a,b) as well as on observational administrative data of judges' decisions (Stevenson, 2018). These studies found that decision-makers tend to be more influenced in their decisions by algorithms when they predict a high risk for black defendants or low risk for white defendants. Alon-Barkat and Busuioc (2023) further examined this tendency of selective adherence in a survey experimental study focused on an administrative context. The study found decision-makers are more likely to follow algorithmic predictions that match negative group stereotypes. Insights from a real-life case involving algorithm use in decision-making by public authorities further corroborated these findings, with decision-makers unlikely to override negative predictions involving minority applicants, resulting in state-sanctioned discrimination. Patterns of selective adherence to algorithmic advice were further corroborated by a subsequent study in the area of policing, which also found evidence of selective deference to algorithmic recommendations (Selten, Robeer & Grimmelikhuijsen, 2023). Similarly aligned with a motivated reasoning logic, participants (police officers) tended to follow algorithmic recommendations to the extent that these aligned with their intuitive judgements.

Altogether, these studies suggest that the use of AI algorithms can amplify bias in public sector in decision-making, not only due to well-established machine bias issues, but also due to biased processing thereof by human decision-makers, pointing at two distinct sources of algorithm-induced bias that can compound and reinforce each other.

Algorithm-Decision Maker Interactions: Future Directions

Taken jointly, these insights point at an important and coherent behavioral research agenda focused on studying the impact algorithms have on public sector decision-makers, the biases that arise in the interaction, and by extension, in public sector decision-making. Future behavioral studies should explore not only the prevalence of these cognitive biases among public decision-makers but also the conditions under which they are more likely to materialize in administrative contexts, their implications as well as potential solutions to ameliorate and mitigate their occurrence. We highlight below, in a theoretically-informed manner, specific conditions and possible "fixes" that could form the subject of empirical investigation.

As to the *conditions* prompting their occurrence, insights from automation studies suggest decision-makers are particularly prone to over-reliance on algorithmic decisional aids in *policy domains* characterized by *complexity, ambiguity, high-risk decisions* involving difficult moral judgments. In such domains, algorithms may serve as effective "moral buffers", leading to "psychological distancing" (Cummings, 2006). This would suggest that frontline workers including judges, police officers, welfare case-workers, whose decisions often involve complex moral judgements, and where algorithms are widely applied, could be especially susceptible to such biases. Importantly, these contexts are precisely the ones where bureaucratic discretion is said to be at its most relevant, involving, by virtue of their street-level nature, complex normative dilemmas and trade-offs. Hence, future research should examine the propensity towards such biases particularly in these contexts, for their enhanced theoretical and practical relevance.

Interestingly, early findings in the context of AI report, as noted above, *selective adherence* precisely in such high-stakes areas, while finding mixed evidence for automation bias in these areas. Future studies should strengthen the validity of these findings by examining these biases across high-stakes bureaucratic contexts as

well as ascertaining under which conditions they are more likely to become manifest. Moreover, studies should further explore the psychological mechanisms underlying these biases and their relation to decision-makers' stereotypes and prior beliefs, through designs that experimentally manipulate these variables: studies may prime subjects' beliefs and test whether this affects their tendency to selectively accept algorithmic advice. Future studies may supplement such experimental designs with observational administrative data, case studies and field experiments, which are less prone to social desirability bias constraints and have higher external validity.

When it comes to the *conditions* under which such biases are more likely to materialize, extant automation bias studies suggest that the propensity to defer to automation stems partially from the belief in the *perceived performative superiority* of automated systems i.e., enhanced user trust in their superior capacities over humans. Future studies should explicitly test to what extent automation bias is associated with enhanced user perception of superior capacities of algorithmic prediction tools.

In this relation, a potential explanation for the observed deviation in recent findings from automation bias studies might be related to the relative novelty of AI algorithm deployment in governmental contexts. This could be an important difference to earlier studies in fields traditionally well-accustomed to automation (aviation, medicine) – both through the repeated reliance on such tools as well as in the context of professional training programs – leading to high levels of trust in their performance. It could be expected that over time, *familiarity* and *repeated interactions* in this context with high-performing algorithmic tools could lead to enhanced user trust and thereby a *higher propensity for deference*. This can be explicitly tested through designs that provide for repeated, sustained interactions of decision-makers with automated tools. Such processes might prove more protracted however, in the public sector context, given *delayed effects* and higher *ambiguity of outcomes* for many tasks and where the counterfactual will often be lacking (e.g., individual risk of recidivism). These conditions further complexify the assessment of outcomes, rendering algorithmic (mis-)performance more ambiguous to users compared to automation in areas like healthcare and aviation.

Investigating the prominence and nature of these biases becomes pressing in light of potential *implications*. *Selective adherence* to algorithmic advice is a disconcerting prospect as studies link it to disparate outcomes that disadvantage minority groups. Similarly, *automation bias* too, is problematic, especially in light of well-documented deficiencies and failings of algorithmic systems currently in use (e.g., Calo & Citron, 2021; O'Neil, 2017; Eubanks, 2018; Buolamwini and Gebru, 2018; Raji et al., 2022) as well as considerable limitations to decision-makers' ability to identify these (Green & Chen, 2019a). A human proclivity for automatic deference under these circumstances would become especially problematic, raising fundamental questions as to the implications of automation for bureaucratic discretion and ensuing impact on citizens.

Public administration research on street-level bureaucracies suggests that bureaucratic discretion and accounting for individual circumstances can be salient to ensuring fair and equitable outcomes in administrative processes (Bovens & Zouridis, 2002). Discretion can play a critical role to enhancing the quality of services, decisions and outcomes and to promoting effective expertise acquisition in the bureaucracy (Gailmard & Patty, 2007) as well as to successful policy implementation (Bartels, 2013; Thomann, van Engen & Tummers, 2018). Importantly in this context, while recent behavioral economics work has specifically advocated for the reliance on algorithms to remove variability in public decision-making, as noted above, street-level bureaucracy literature would suggest that such variability can also play crucial functions in administrative processes. The exercise of discretion is crucial to making complex "normative and factual determinations" and trade-offs, especially so in sensitive task domains where that "factual and normative basis is incomplete" (Young et al. 2021, 247). It can help overcome excessive rigidity and formalization to safeguard the rights of vulnerable citizens (Elyounes, 2021).

In this regard, the behavioral economics quest to extricate "noise" from decision-making by means of algorithmic decision-making (removing human judgment from the process) could simultaneously serve to reinforce bureaucratic inflexibility and rule rigidness – a return to the Weberian "iron cage." (Behavioral) public administration scholarship is in a good position to nuance implications of the adoption of automation into public sectors, taking into account not only standardization and control considerations but also the broader institutional and organizational picture as to how administrative discretion, put to good use, can play an important and valuable function in bureaucratic processes.

Automation bias would also have important practical and normative *implications* not only for *discretion* but also relatedly, in terms of *accountability* and the allocation of responsibility for actions and decisions informed by algorithmic inputs. The risk is that what are in principle hybrid decision-making systems become *de facto* fully automated systems, with the human decision-makers effectively relinquishing or abdicating responsibility to the algorithm.

There are also important implications for the effectiveness of accountability checks to be put in place on algorithmic power. Keeping humans-in-the-loop is generally prescribed as a key safeguard on automated decision-making (see the EU GDPR, which provides for human intervention as a check on solely automated processing). The discussion above on automation bias raises important questions on the effectiveness of human intervention as a check in this context. Human oversight of AI remains understudied empirically in bureaucratic contexts, despite the crucial role human oversight is envisaged to play in such contexts. A much deeper investigation of our cognitive limits in overseeing AI – and how effective human oversight measures might look like – is required to be able to devise meaningful human oversight measures we can take comfort in as effective safeguards (see also Busuioc, 2022).

Future research should also research *potential "fixes"* or *solutions* to behavioral biases brought on by algorithms, and again here too, extant literature provides useful starting points. Both literatures on automation bias and biased information processing point at a common potential solution: workload reduction. Because automation bias is understood to be partly driven by attempts to reduce cognitive effort, studies have linked it with the presence of workloads, distractions and task complexity (Lyell & Coiera, 2017). Reducing workload was also found to reduce biases and discriminatory decisions in some instances (Andersen & Guul, 2019) and could thus potentially help alleviate biased adherence.

Moreover, existing research in psychology has suggested automation bias can also be attenuated by enhancing the *pre-decisional accountability* of human decision-makers (Skitka, Mosier & Burdick, 2000). The latter notion is aligned with considerable social psychology research on accountability and a growing line of public administration research, suggesting that introducing calibrated accountability mechanisms can serve to improve the quality of decisions and reduce cognitive biases (e.g., Tetlock, 1983; Schillemans, 2016; Aleksovska, Schillemans & Grimmelikhuijsen, 2019). Given its well-established effects of generating "pre-emptive self-criticism" (Tetlock, 1983) and integrative complex thinking, pre-decisional accountability (the expectation that one may be called to account) could potentially serve to mitigate a propensity for *biased adherence* to algorithmic advice. Through its demonstrated ability to lead to increased cognitive effort, it could similarly also be hypothesized to have an ameliorating effect on *automation bias*.

Finally, envisaged solutions to algorithmic bias could potentially also entail training interventions aimed at ameliorating cognitive bias, albeit studies in psychology report somewhat mixed findings on this (Bahner, Hüper & Manzey, 2008; Lyell & Coiera, 2017). These potential solutions should be explicitly investigated and experimentally tested with political and bureaucratic decision-making settings in mind to examine their effectiveness in such contexts.

Citizen-Algorithm Interactions: Of Algorithmic Burdens, Citizen Trust and Algorithmic Transparency

The previous section focused on issues related to the interactions of public sector decision-makers with algorithms. A second important avenue for behavioral research in this area pertains, we argue, to another type of algorithmic encounters in the public sector: citizens' interactions with algorithms. While incipient studies are starting to emerge in relation to algorithmic decision aids (e.g., Miller & Keiser, 2021; Grimmelikhuijsen, 2023; Schiff, Schiff & Pierson, 2021; Keppeler, 2024), the topic calls for greater attention by behavioral public administration scholars in light of the veritable transformation governments are undergoing.

How do citizens respond to and experience algorithmic decision-making? What drives citizen attitudes in this regard, especially as such systems are being increasingly deployed in high-stakes scenarios? What shapes *citizens' trust (and distrust)* in AI systems and in governmental services assisted by algorithmic systems? And importantly, how does *algorithmic transparency*, and other forms of information provision on algorithm

operation, impact *citizen trust* in this context? These questions become especially important given initial findings (Smith, 2018; Rainie et al., 2022) indicative of relatively negative attitudes among citizens towards AI technology and its deployment in public settings. This raises important questions about the implications that the broad use of AI in government would carry for trust in, and the legitimacy of, public institutions, more broadly.

We structure our discussion below along two main theoretical themes, which are particularly important for future studies to consider and can be linked to current theoretical discussions in the field. The first pertains to the implications of the use of AI algorithms for administrative burdens, what we term "algorithmic burdens" in this context. The second regards citizens' trust and how trust is shaped by aspects of algorithmic transparency and human involvement in algorithmic decision-making processes in administrative contexts.

Burdensome Encounters: Algorithmic Burdens in Citizen-State Interactions

Administrative burdens pertain to hurdles arising in citizens' interactions with government, giving rise to significant learning, compliance or psychological costs (Herd & Moynihan, 2018). The introduction of automation was hoped to minimize such burdens and lead to higher levels of inclusion and uptake in public services. In light of growing evidence of policy fiascos linked to algorithm use (see e.g. Eubanks, 2018; Whiteford, 2021; Geiger, 2021), initial optimistic assessments as to the potential of automated tools to improve citizen access to governmental services are increasingly giving way to concerns about negative impacts. Notably, scholars have flagged their potential to introduce (or further compound) administrative burdens (Peeters & Widlak, 2023; Madsen, Lindgren & Melin, 2022; Larsson, 2021) and to negatively impact disadvantaged citizens (Ranchordás 2022; Alon-Barkat & Busuioc, 2023). Importantly, such burdens – which we term here "algorithmic burdens" when resulting from algorithmic automation – stem not only from the digitalization of citizens' encounters with government (Peeters, 2023; Peeters & Widlak, 2018) but also specifically from the implications of the incorporation of AI algorithmic tools in the decision-making process.

Algorithmic technologies are reported to have disparate and exclusionary effects, with disadvantaged citizens wrongly accused of fraud (Geiger, 2021), deprived of critical welfare entitlements on unprecedented scale (Citron, 2008; Eubanks, 2018; Elyounes, 2021), disproportionately targeted by predictive policing tools (Ferguson, 2017; Richardson et al., 2021) or wrongly labeled as future criminals (Angwin et al., 2016; Citron, 2016). Such burdens are distributed unevenly with vulnerable and marginalized individuals and communities facing considerably higher levels of monitoring and scrutiny. Much in the manner that "facially neutral rules" can "launder racially disproportionate burdens" (Ray, Herd & Moynihan 2023, 141), seemingly "neutral" algorithmic systems can launder and conceal excessive burdens with heavily disparate effects.

Thus, additive *compliance costs* can arise for citizens flagged for closer scrutiny or automatically profiled by AI tools, disproportionately required to jump through cumbersome bureaucratic hoops to adduce evidence to disprove the algorithmic assessment in order to convince bureaucrats to overturn an erroneous algorithm-based decision. Poorly designed large-scale algorithmic systems coupled with shortages in data and algorithmic literacy skills (among both the bureaucracy and the citizenry) further amplify the likelihood of administrative error and associated enhanced compliance costs to reverse it. Typographical errors and/or self-reporting mistakes can be especially costly in this context, as even small errors or inconsistencies across different administrative data sources can be spotted by pattern recognition machine learning systems as discrepancies indicative of patterns of willful deception or misrepresentation. Burdens of proof shift to the citizen to disprove outcomes, often in the absence of insight as to the original source or cause thereof, given the often black-box nature of such systems.

Relatedly, citizens' experience of public services that involve AI algorithms may be characterized by considerable *psychological costs*, namely the feeling of loss of autonomy, disempowerment and stress. These costs become accentuated in the context of algorithmic decisions that impact citizens' livelihoods, well-being and basic rights in consequential ways, and in the absence of face-to-face frontline interactions serving as intermediaries to automated bureaucratic systems. What is more, while the use of AI technology for some functions may help reduce *learning costs* for example, by facilitating the targeting of policy information and tailoring communications, it can simultaneously result in increased burdens stemming from the difficulty of

obtaining the necessary information to understand an algorithmic system's decisions. Learning costs pertaining to acquiring such information and algorithmic literacy skills will play a role not only in averting administrative error in the context of increasingly standardized and mass-scale automated administrative systems but also in attending to and managing its effects, with vulnerable citizens facing disproportionately higher costs.

For instance, the infamous childcare benefits scandal in the Netherlands ("toeslagenaffaire"), an emblematic case of state-sanctioned discrimination in AI-informed decision-making, exhibited many of the elements above. Over 25,000 citizens were wrongly accused of fraud, flagged for negligible clerical mistakes such as a missing signature and required to retrospectively pay back as much as tens of thousands of euro in benefits, resulting in serious financial and psychological costs such as bankruptcies, mental health illnesses, suicides and broken families (Geiger, 2021). Burdens were disproportionately distributed: 70% of the affected victims had a migration background, half were single parents, and one fifth were social security recipients (NOS Nieuws, 2022). Those who attempted to challenge algorithmic decisions were told that officials could not relay the grounds for the decisions. Bureaucrats were not given information about the factors that led to the high-risk classification by the system, with public officials reportedly justifying decisions "because the algorithm said so" (Hadwick & Lan, 2021, 6), compounding powerlessness among victims.

This speaks to key concerns which have been raised more broadly, that with the growing automation of public services, core public values and protections are being displaced (Citron, 2008; Larsson, 2021; Schou & Hjelholt, 2018; UN Special Rapporteur on Extreme Poverty and Human Rights, 2019), affecting disproportionately already marginalized communities.

Thus, more sobering accounts of the adoption of automation in government raise the prospect of such developments, unless carefully and thoughtfully managed, extracting significant costs from the citizenry – *compliance, psychological* and *learning costs* that require careful and systematic investigation. Such burdens stand to negatively impact citizen experience of government, shaping their trust in public institutions.

Beyond Black-Boxes: Algorithmic Transparency, Human Involvement and Citizen Trust

Anticipating these concerns, algorithmic transparency has been at the center of technical debates as a crucial element to advancing user and citizen trust in AI algorithms as well as to foster system improvement and overall accountability (Busuioc, 2021; Grimmelikhuijsen & Meijer, 2022; Busuioc, Curtin & Almada, 2022). Many of the solutions discussed in the computer science community for algorithmic transparency attempt to make algorithms more explainable or more interpretable. Computer scientists have increasingly recognized that trust in the model is crucial for it being deployed in practice and that model understandability, the ability to understand the reasons behind predictions, is a key aspect thereof (Ribeiro, Singh & Guestrin, 2016).

Yet, transparency is particularly difficult and problematic for AI models to achieve (Pasquale, 2015) given their "black-box" nature. Many algorithms are subject to private intellectual property protections that preclude disclosure and thus are opaque in a *legal* sense. Their computational complexity additionally makes them opaque in a *technical* sense (Burrell, 2016; Rudin, 2019). The latter, technical opacity, is especially characteristic of dominant and widely popular AI approaches such as deep learning or artificial neural networks. The basis for a model's outputs, the decision-making rationale – which specific inputs (features) contributed to a specific output – remains hidden, raising difficult challenges of inscrutability for professionals when relied upon in administrative contexts.

Under the assumption that increased transparency of black boxes will lead to enhanced user and citizen trust, a variety of transparency techniques for black-box models are being developed (and debated). Existing options range from: *model transparency* (i.e. disclosing the source code, model, weights, training data); *explainable AI* (XAI) approaches (i.e. post-hoc explanation models that "approximate" the behavior of the underlying black box, see Phillips et al., 2021, 13-15 for an overview); or, foregoing black-box models in favor of *inherently interpretable ML models* (Rudin, 2019). The latter are instead transparent from the outset. Questions of how these various technical approaches impact user trust and understanding are central to algorithmic debates. Nevertheless, there is "not a common agreement of what is required for an explanation" and "no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans, although it is of fundamental importance" (Guidotti et al., 2018, 36).

This points at an important role to play for behavioral public administration scholars in contributing to these debates in a public sector context. Multiple works in the field, also building on psychology procedural fairness theory, link trust in authority with the transparency of decision-making processes (e.g., de Fine Licht, 2014; Porumbescu, 2017; Grimmelikhuijsen et al., 2021). Further behavioral research is needed to decipher which technical solutions result in greater user understanding of model functioning and serve to increase user trust in their predictions and under which conditions. The technical solutions proposed for algorithmic transparency are varied and largely untested both as to user comprehensibility as well as to how users and citizens respond to these various options.

At the same time, to further complicate matters, transparency studies (outside the context of AI), also indicate that the relationship between transparency and trust is not a straightforward one. The impact of transparency on trust is context-dependent (de Fine Licht, 2014; Grimmelikhuijsen et al., 2021), varies according to the type of transparency at stake (Cucciniello, Porumbescu & Grimmelikhuijsen, 2017), and in fact some studies report negative findings with respect to transparency effects on trust (e.g., de Fine Licht, 2011; Grimmelikhuijsen et al., 2013).

This points at the need to systematically investigate the *role of algorithmic transparency in shaping citizen judgements* of algorithm-based decisions as well as how *different types of information on model functioning* – namely: model transparency; interpretable models; explainable AI – fare in terms of meaningfully enhancing user and citizen understandability of algorithmic outputs. Incipient work on this, albeit not focused explicitly on AI algorithms (see Grimmelikhuijsen, 2023) or lacking a public sector focus (Lee, 2018; Binns et al., 2018), indicates that different techniques can have context-dependent effects. This line of research would be a natural continuation of behavioral public administration work in this area, which traditionally focused on investigating how *different types of transparency* (i.e., decision-making; policy; outcomes) impact citizen trust. This research can be extended to types of transparency relevant for algorithmic sources, as laid out above.

What is more, transparency in this context pertains not only to the substantive content of AI algorithmic decision-making, but also more fundamentally to disclosure of when such systems are deployed in the first place. For instance, administrative agencies are often not transparent regarding the extent to which they rely on algorithmic decisional aides (Raso, 2021), potentially misleading citizens in non-negligeable ways. Relatedly, relevant behavioral public administration work pertains to investigating how the disclosure and deployment of AI applications in public services impacts citizen trust and satisfaction with government services (see Keppeler, 2024).

In addition to the critical role of transparency, a related aspect that can also be theorized to be highly influential for citizens' trust in the context of algorithm-based government decisions regards the continued involvement of humans in the decision-making process, including investigations as to how to ensure this involvement is meaningful. This theoretical direction, which can also be linked to our above-mentioned discussion on psychological costs imposed by algorithms and resulting feelings of disempowerment / loss of autonomy, receives tentative support from recent vignette survey experiments (Aoki, 2021; Waldman & Martin, 2022). Thus, ensuring that there is a "human in the loop", and that there is an effective ability for human intervention and oversight is important not only as safeguard against potential errors and algorithmic biases, as discussed above, but also for securing citizens' trust and enhanced legitimacy (Busuioc, 2021; Kennedy, Waggoner & Ward, 2022; Keppeler, 2024).

At the same time, it is important to recognize that citizen trust in algorithm-based government services speaks to important normative debates and raises difficult and complex dilemmas pertaining to the nature of trust relations between citizens and government agencies. While citizen trust in governmental bodies is crucial to their authority and to them being able to deliver better outcomes effectively (and with less compliance costs), from a normative democratic point of view, high citizen trust is not always desirable. Citizens' judgments of the trustworthiness of government agencies should be rationally-informed, warranted in light of existing evidence of governmental performance as well as the existence of safeguards and mechanisms of supervision and accountability.

Accordingly, a skeptical "trust but verify" approach can lead to high or low trust in government agencies (Norris, 2022), contrary to "blind" or "credulous" trust on the one hand, and cynical mistrust on the other hand. From a democratic perspective, low trust can be a satisfactory outcome for instance, when the

performance of public bodies and actions falls short of expectations. In this context, low trust is normatively desirable and can be crucial to triggering crucial processes of democratic accountability. Correspondingly, in an AI context, it is desirable that citizens should apply a "healthy skepticism" approach towards agencies' incorporation of AI into their decision-making processes. That is to say, seek information cues as to such systems' quality as well as to the presence of effective institutional mechanisms of supervision and accountability, with the degree of algorithmic transparency itself representing such a cue.

For this reason, it is important that behavioral work in this context does not devolve into an exercise in artificially "manufacturing trust", attempting to "maximize" citizen trust and perceived legitimacy in AI decision-making, irrespective of actual system performance and ability to detect critical issues and malfunctions (cf. de Fine Licht & de Fine Licht, 2020). Recurrent scandals pertaining to algorithmic systems' adoption in government certainly suggest a healthy dose of skepticism among the citizenry is well-warranted when it comes to the deployment of such systems. Well-documented serious cases of poorly designed algorithmic systems in use in the public sector (e.g., Geiger, 2021; Citron, 2008, 2016; Eubanks, 2018; Algemene Rekenkamer, 2022) indicate that such systems often operate under the public radar and without proper vetting and supervision.

Conclusion

As AI becomes an integral part of more and more government functions and policy domains, it is important we study the nature of our interactions with the technology in this context, including specific cognitive biases that arise in this interaction.

We have proposed two crucial avenues for future behavioral research regarding the use of algorithms in the public sector. First, we proposed that scholars should investigate *algorithm-decision maker interactions* and the sources of bias arising at this interface in public sector contexts. Extant social psychology research on automation points to the susceptibility of decision-makers to unwarrantedly defer to algorithmic advice, or "automation bias". At the same time, extrapolating from current theories of information processing and decision-making, would point at selective (rather than automatic) adherence: a cognitive predisposition to rely on algorithmic advice selectively, in ways consistent with pre-existing beliefs. We suggest that future studies should extend current theories of information processing and decision-making and systematically explore empirically which of *these biases materialize* in a public administration context, *under which conditions,* study *their implications* as well as *possible "fixes*". We provide theoretically-informed, concrete elements for investigation on each of these aspects.

The *second* research avenue that we have discussed in this contribution regards the interaction between *citizens and algorithms* and citizen responses to algorithmic-based government decisions. We unpack this further by discussing the potential thereof to give rise to *algorithmic burdens* in citizen-state interactions, shaping citizen trust and experience of government in non-negligeable ways, as well as different technical *types of algorithmic transparency* that need to be tested as to *user understandability* of system functioning as well as to their *effects on citizen trust*. As AI is transforming public services, it becomes fundamentally important to understand how encounters with artificial intelligence in this context are shaping citizen attitudes. Citizen trust underpins public support for new technologies and is foundational to the legitimacy of government action and of public institutions, more broadly.

While these research avenues can be explored via various empirical approaches, experimental designs are especially well-suited for addressing these questions. If we aim to compare responses to algorithmic versus human expert advice in a robust manner, controlled experimental designs, which include random assignment to such conditions would be beneficial (e.g., Logg, Minson & Moore, 2019; Hou & Jung, 2021; Kennedy, Waggoner & Ward, 2022). Likewise, vignette survey experiments provide an effective tool for investigating how the perceptions of citizens are influenced by differences in algorithmic transparency and the involvement of human decision-makers in the process. Moreover, employing experimental designs can advance our knowledge about these causal relations by disentangling the psychological micro-mechanisms underlying them. At the same time, future studies applying experimental designs should dedicate considerable attention to issues of generalizability to real-world public sector settings.

A focus on human-AI interactions is already at the heart of algorithmic debates in the computer science community as well as among policy-makers. Policy initiatives on "human-centric AI" – placing "people at the centre of the development of AI" (European Commission, 2018), and similarly, technical debates on "hybrid intelligence" – focused on developing intelligent systems we can cooperate and interact with (Jonker, 2019) recognize the key importance of understanding human-algorithm interactions. For a social science discipline like ours, such a focus only becomes ever more relevant. As AI algorithmic tools are increasingly being assimilated in the public sector across different jurisdictions, there is an urgency to understand the ways in which we interact with algorithms in these domains and the broader implications thereof in shaping institutional and political environments. Beyond the techno-optimism hype, a meaningful research agenda on the incorporation of AI in government requires accounting for its behavioral aspects, including taking the study of algorithm-related cognitive biases seriously in this context. Rather than assuming automation will magically do away with human bias, we need to closely study and understand how it can replicate and encode bias in social institutions, our limitations (be it as citizens and/or decision-makers) in interacting with automation and the organizational and institutional challenges that can arise as a result. Throughout this article, we proposed several concrete starting points for investigation that can move us forward in the course of this research endeavor.

Our contribution is also a response to the call to use behavioral approaches to answer "big questions" central to public administration (Moynihan, 2018). We highlight that understanding human-AI interactions in a public sector context is one of the biggest quests for public administration to tackle in the years to come, with significant societal impact. We believe that this research agenda responds to rightful critiques of behavioral public administration for its overly-narrow focus on experimental designs often lacking adequate theoretical basis, weakly integrated in broader institutional and political contexts (Bertelli & Riccucci, 2022). Unpacking the psychological mechanisms underpinning algorithmic decision-making in administrative contexts will be of crucial theoretical and empirical relevance to tackling core questions as to the impact of AI for bureaucratic discretion and expertise, citizens' experience of government and the nature of public authority in the age of automation. Behavioral public administration is well-positioned to contribute in a meaningful manner to critical building blocks of this important research agenda, should it choose to rise to the occasion.

References

Aleksovska, Marija, Schillemans, Thomas & Grimmelikhuijsen, Stephan. (2019). Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review. *Journal of Behavioral Public Administration* 2(2). doi:10.30636/jbpa.22.66.

Algemene Rekenkamer (Netherlands Court of Audit). (2022). *An Audit of Algorithms*. https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government.

Alon-Barkat, Saar & Busuioc, Madalina. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory* 33(1), 153-169.

Andersen, Simon Calmar & Guul, Thorbjørn Sejr. (2019). Reducing Minority Discrimination at the Front Line—Combined Survey and Field Experimental Evidence. *Journal of Public Administration Research and Theory* 29(3), 429-444. doi:10.1093/jopart/muy083.

Angwin, Julia, Larson, Jeff, Mattu, Surya & Kirchner, Lauren. (2016). Machine Bias. *ProPublica,* May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aoki, Naomi. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior* 114, 106572.

Assouline, Michaela, Gilad, Sharon & Ben-Nun Bloom, Pazit. (2022). Discrimination of Minority Welfare Claimants in the Real World: The Effect of Implicit Prejudice. *Journal of Public Administration Research and Theory* 32(1), 75-96.

Baekgaard, Martin, Christensen, Julian, Dahlmann, Casper Mondrup, Mathiasen, Asbjørn & Petersen, Niels Bjørn Grund. (2019). The Role of Evidence in Politics: Motivated Reasoning and Persuasion among Politicians. *British Journal of Political Science* 49(3), 1117–1140. doi:10.1017/S0007123417000084.

Baekgaard, Martin & Serritzlew, Søren. (2016). Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension. *Public Administration Review* 76(1), 73–82. doi:10.1111/puar.12406.

Bahner, J. Elin, Hüper, Anke-Dorothea & Manzey, Dietrich. (2008). Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience. *International Journal of Human-Computer Studies* 66(9), 688–699. doi:10.1016/j.ijhcs.2008.06.001.

Bartels, Koen P. (2013). Public encounters: The history and future of face-to-face contact between public professionals and citizens. *Public administration* 91(2), 469-483.

Barth, Tomas J. & Arnold, Eddy. (1999). Artificial Intelligence and Administrative Discretion: Implications for Public Administration. *The American Review of Public Administration* 29(4), 332-351. doi:10.1177/02750749922064463

Barocas, Solon & Selbst, Andrew D. (2016). Big data's disparate impact. *Calif. L. Rev.* 104, 671-732.

Battaglio Jr, R. Paul, Belardinelli, Paolo, Bellé, Nicola & Cantarelli, Paola. (2019). Behavioral public administration ad fontes: A synthesis of research on bounded rationality, cognitive biases, and nudging in public organizations. *Public Administration Review* 79(3), 304-320.

Bertelli, Anthony M. & Riccucci, Norma M. (2022). What is Behavioral Public Administration Good for? *Public Administration Review* 82(1), 179–183. doi.org/10.1111/puar.13283.

Binns, Reuben, Van Kleek, Max, Veale, Michael, Lyngs, Ulrik, Zhao, Jun & Shadbolt, Nigel. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *CHI' 18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

de Boer, Noortje & Raaphorst, Nadine. (2023). Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Management Review* 25(1), 42-62.

Bovens, Mark & Zouridis, Stavros. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review* 62(2), 174-184.

Bullock, Justin B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration* 49(7), 751-761. doi:10.1177/0275074019856123.

Buolamwini, Joy, & Gebru, Timnit. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81, 77-91.

Burrell, Jena. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1), 2053951715622512.

Burton, Jason W., Stein, Mari-Klara & Jensen, Tina Blegind. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2), 220-239.

Busuioc, Madalina. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review* 81(5), 825-826. doi.org/10.1111/puar.13293.

Busuioc, Madalina. (2022). "AI Algorithmic Oversight: New Frontiers in Regulation" In *Handbook of Regulatory Authorities* (eds) Maggetti, Martino, Fabrizio Di Mascio and Alessandro Natalini, pp. 469-486. Edward Elgar Publishing. https://doi.org/10.4337/9781839108990.00043

Busuioc, Madalina, Curtin, Deirdre & Almada, Marco. (2022). Reclaiming Transparency: Contesting the Logics of Secrecy within the AI Act. *European Law Open* 2(1), 79-105. doi:10.1017/elo.2022.47.

Calo, Ryan & Citron, Danielle K. (2021). The Automated Administrative State: A Crisis of Legitimacy. *Emory Law Journal*. 70(4), 797-846. Available at: https://scholarlycommons.law.emory.edu/elj/vol70/iss4/1.

Christensen, Julian, Dahlmann, Casper Mondrup, Mathiasen, Asbjørn Hovgaard, Moynihan, Donald P. & Petersen, Niels Bjørn Grund. (2018). How Do Elected Officials Evaluate Performance? Goal Preferences, Governance Preferences, and the Process of Goal Reprioritization. *Journal of Public Administration Research and Theory* 28(2), 197–211. doi:10.1093/jopart/muy001.

Citron, Danielle K. (2008). Technological Due Process. *Washington University Law Review* 85(6), 1249–1313.

Citron, Danielle K. (2016). (Un)Fairness Of Risk Scores In Criminal Sentencing. Forbes, July 13. https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/?sh=65ee9c734ad2.

Cobbe, Jennifer. (2019). Administrative law and the machines of government: judicial review of automated public-sector decision-making. *Legal Studies* 39(4), 636-655.

Cucciniello, Maria, Porumbescu, Gregory A. & Grimmelikhuijsen, Stephan. (2017). 25 Years of Transparency Research: Evidence and Future Directions. *Public Administration Review* 77(1), 32-44. doi:10.1111/puar.12685.

Cummings, Mary L. (2006). Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies* 32(1), 23-31. doi:10.21061/jots.v32i1.a.4.