12-7-2022

Aversion vs. Abstinence: Conceptual Distinctions for the Receptivity Toward Algorithmic Decision-Making Systems Within Value-laden Contexts

Oliver Hannon
University of Sydney, oliver.hannon@sydney.edu.au

Uri Gal
University of Sydney, uri.gal@sydney.edu.au

Ilan Dar-Nimrod
University of Sydney, ilan.dar-nimrod@sydney.edu.au

Aversion vs. Abstinence: Conceptual Distinctions for the Receptivity Toward Algorithmic Decision-Making Systems Within Value-laden Contexts

Full research paper

Oliver Hannon
Discipline of Business Information Systems
The University of Sydney Business School
Sydney, Australia
Email: oliver.hannon@sydney.edu.au

Uri Gal
Discipline of Business Information Systems
The University of Sydney Business School
Sydney, Australia
Email: uri.gal@sydney.edu.au

Ilan Dar-Nimrod
School of Psychology
The University of Sydney
Sydney, Australia
Email: ilan.dar-nimrod@sydney.edu.au

Abstract

Whilst algorithmic decision-making systems (ADMS) become increasingly pertinent across several contexts, many remain reluctant to adopt such systems, preferring human alternatives – often explored as "Algorithm Aversion". However, the associated literature primarily frames this tendency in a utility-focused fashion, based on users' perceptions of efficacy or accuracy. This framing offers a narrow scope of "aversion" that neglects emotional and experiential elements that may be at play, as well as especially prominent in "value-laden contexts" (e.g., medicine). This study uses an inductive approach to identifying various concepts and themes emerging from open-ended responses to the potential use of a future ADMS in such a context. Different reactions (both reluctant and receptive) of ADMS are then discussed and offered conceptual distinctions that may inform future examinations of the resulting biases. In doing so, we start to respond to the call for qualitative research examining the underlying motives related to Algorithm Aversion.

Keywords: Algorithmic decision-making, algorithm aversion, decision aids, abstinence, bias.

1 Introduction

Algorithmic decision-making systems (ADMS) are used to inform and even automate decision-making for a variety of organisations around the world (Newell and Marabelli 2015). As such, both industry usage and scholarly discussion of ADMS have gained traction in business disciplines, such as marketing and management (Erevelles et al. 2016; Meijerink and Bondarouk 2021). With a focus on "strategic decisions" and measures of performance (e.g., accuracy and efficiency) as the primary rationale for adoption of these decision aids at the organisational level, a push toward their use in service of purportedly superior decision-making comes as no surprise.

Having said that, across a range of tasks and contexts, many remain reluctant to utilise ADMS. Specifically, much of the extant literature demonstrates that people may exhibit a bias against algorithmic systems in favour of human alternatives where available – a phenomenon often studied as "Algorithm Aversion" (Dietvorst et al. 2015). A recent systemic review found that contributions to this space are typically experimental works that seek to demonstrate when and why people choose to reject algorithms, and states that more qualitative work is needed to "get in-depth insights about the underlying motives for Algorithm Aversion", especially in order to foster the uptake of algorithms for superior utility in decision-making (Mahmud et al. 2022, p. 15). We agree that qualitative insights are needed and seek to assist in addressing this by exploring what some of the underlying motives for Algorithm Aversion are as part of our research problem. Put simply – how do people evaluate ADMS and what are some of the different rationales behind their receptivity (or lack thereof) toward them?

However, as a second component of our research problem, we also suggest that there are limitations within the existing Algorithm Aversion literature that require consideration for two primary reasons. The first reason is the changing nature of the decision contexts within which ADMS are used and examined whilst making reference to Algorithm Aversion. Over time, studies within the literature on Algorithm Aversion (and counterargument: "Algorithm Appreciation"; Logg et al. 2019) have moved far beyond their origins in forecasting (e.g., predicting student performance; Dietvorst et al. 2015). The phenomenon is now cited amidst much more high-stakes, value-laden decision contexts, such as judicial sentencing decisions (Bigman and Gray 2018) and political forecasting (Kennedy et al. 2022). It is also being explored in contexts where more autonomous (rather than advisory) systems are being examined, such as self-driving vehicles (Shariff et al. 2017) and automated job screenings (Noble et al. 2021). We argue that if the broader literature neglects to adequately account for this contextual shift, it risks oversimplifying the understanding of the motives for the rejection of ADMS in high-stakes, value-laden decision contexts, which may elicit more affective reactions and even moral considerations from users compared to that of more utility-focused tasks (e.g., predictive product forecasting).

Second, by appealing to factors like superior accuracy and enhanced utility as the rationale for adoption, the literature risks a managerial or modernisation bias toward the uptake of ADMS and against rejection. For example, Mahmud et al. (2022, p. 1) states that "it is logical to follow algorithms when it is evident that algorithms perform better than humans", and that "deviating from this rational behaviour may be viewed as a behavioural anomaly, which may reduce human subjects' expected utility". Within the extant literature, performance and utility serve as the main rationale for adoption of ADMS, and "aversion" is typically viewed as an obstacle to be overcome, with mitigation as the primary managerial implication cited (e.g., Castelo et al. 2019). This framing of aversion may only offer a partial account of a more complex phenomenon. An appeal to "logic" or "strategy" due to "better" performance or utility may presuppose that people do not, or should not, make decisions via means other than rationality (e.g., emotive, ethical, or experience-based reasoning). We argue that the current discussion of Algorithm Aversion in the literature may be oversimplified because of the focus on evaluations of utility as the primary reason to adopt, and the focus on "rational" bases for the rejection (or uptake) of algorithms.

In order to better understand how people evaluate the use of ADMS in decision contexts that might elicit a wider range of possible reactions, we examined open-ended survey responses from a large cohort of Australian university students. The stated reasoning behind their reluctance or receptivity toward utilising ADMS in a high-stakes, value-laden decision context (decision-making in medicine) was analysed using an inductive approach and incorporating processes for data reduction and presentation as themes put forth by Gioia et al. (2013). A variety of concepts and themes emerged throughout the analysis, that were distilled further into aggregate dimensions to provide conceptual distinctions that may be useful for developing a theoretical understanding relating to the receptivity toward ADMS, and even assist with the development of constructs and measures, as is also called for within the literature (Mahmud et al. 2022, p. 16). This approach was use, with these aims in mind, because "concepts are precursors to constructs" and "it is first necessary to discover relevant concepts for the purpose of theory building that can guide the creation and validation of constructs" (Gioia et al. 2013, p. 16).

In summary, language emerged that shows some making evaluations of ADMS that are outside of, and even with complete disregard for any consideration of the system's efficacy, accuracy, or utility. Some reasoning cited instead revolves around ethics and affective elements that are not adequately explored within the work on Algorithm Aversion. Different to "Algorithm Aversion" (i.e., the *rejection* of algorithms based on "utility-focused" assessments, such as expected accuracy or efficacy), we assert that "Algorithmic Abstinence" is the *objection* to the use of algorithms based on "value-laden" assessments (e.g., ethical or affect-based reactions). Similarly, unlike the antithesis to Algorithm Aversion explored by Logg et al. (2019), "Algorithm Appreciation" (i.e., preference for an algorithm's advice based on utility), we assert that receptivity to an algorithm's advice within value-laden contexts may offer "Algorithmic Appeasements" – whereby users gain greater comfort in decisions supported by ADMS, based on value-laden assessments of its use (e.g., affect-based, ethical, etc).

This work makes three contributions. First, the conceptualisation of "Algorithmic Abstinence" and "Algorithmic Appeasement" provides useful distinctions for scholars that may deepen the understanding of people's decisions to use (or not to use) ADMS in situations that are not entirely utility-focused (i.e., "value-laden" decision contexts). Second, this work addresses the call from Mahmud et al. (2022, p. 15) for qualitative work that examines the "underlying motives" of Algorithm Aversion in ways that have not previously been examined, as well as that may assist with the ongoing development of a comprehensive framework for this literature. Finally, the study may assist with the call made for developing a measurement relating to the phenomenon (Mahmud et al. 2022, p. 16) by offering exploratory findings in service of creating constructs or complimentary dimensions to the base understanding of "Aversion", instead based on value-laden rejection (or uptake) of ADMS.

2 Conceptual Background

From "recommender systems" to "robo-advisors", decision aids that utilise algorithms take on many forms under many different names and definitions within the literature (Burton et al. 2020; Mahmud et al. 2022). However, "algorithms" can be largely thought of as sets of procedures, calculations, and rules for transforming input data into some form of output (Gillespie et al. 2014). Though originally thought to be suited to certain tasks and perhaps not others (e.g., suited to numerical tasks and not subjective tasks), ADMS are becoming increasingly capable. Such systems are now influencing complex organisational decision-making, including informing managers' workplace decisions with things like evaluating and predicting employee performance (Lee 2018). However, the implementation of ADMS presents a range of ethical issues and ongoing challenges that need to be considered as well.

2.1 Algorithms, Ethics & "Value-laden" Decision-Making in Medicine

As ADMS become increasingly prominent across social domains, the literature has started to address a variety of associated ethical ramifications. Matters such as discriminatory or unfair outcomes (Mittelstadt et al. 2016), the perpetuation of social inequalities and systemic biases (O'Neil 2016), and overall "datafication" of humans (Newell and Marabelli 2015) have all become prominent discussions within academia. It could be argued that considerations for ethical issues become particularly acute in realms such as medicine, "where lives are at stake and there are significant sensitivities that are not as pertinent in other domains" (Trocin et al. 2021, p. 1). Within this realm, ADMS are increasingly used for tasks surrounding tasks like diagnoses and are even expected to replace their human counterparts within certain medical disciplines in the very near future (Obermeyer and Emanuel 2016).

However, the use of ADMS as part of medical decision-making becomes increasingly complex, as not only do many medical decisions have high stakes, but what constitutes a "good" medical decision requires the consideration of a patient's "values and preferences, not just clinical information" (McDougall 2019, p. 157). To illustrate, the involvement of patients in decision-making is fundamental to what we understand as "shared decision-making" between the patient and the physician within a Western healthcare framework, and the approach upholds the autonomy of the patient and informed consent as pertinent principles (Brock 1991). With that said, where it may be possible to consult a medical ADMS, doing so in a way that is appropriate for any given patient or situation might be complicated by role of the algorithm if it does not "encourage doctors and patients to recognise treatment decision-making as value-laden" (McDougall 2019, p. 157). Within a treatment decision context, what constitutes a medically "correct" decision (i.e., that prioritises the preservation and life of the patient) may not always align with the "right" decision that is desired according to a patient's preferences or wishes. In situations where the values of a patient play a significant role, the decision not to utilise an ADMS may have less to do with one's assessment of its efficacy, and more to do with exercising a patient's autonomy and a reflection of informed consent. Despite this, recent bioethics literature has increasingly discussed the role that ADMS will play in medical decision-making, bringing

with it an "ontologically distinct situation from prior care models", and leading some to suggest such technology should only ever play advisory roles to human decision-makers (Arnold 2021, p. 121). Nevertheless, discussions are taking place surrounding the use of algorithms as decision-makers themselves, including cases where patients are unable to exercise their own decision-making capability.

One potential future use case, coined the "Autonomy Algorithm", is purported to be a complex algorithm that will mine an incapacitated patient's electronic health records, digital footprint (e.g., social media), and other related medical data, to assist in predicting their preferred medical course of action (Lamanna and Byrne 2018). Whilst its proponents still caution against the dangers of using algorithms and call for input from clinicians and family where appropriate, they also state that the algorithm could be "truly patient centered" and "reduce the significant burdens of patients' incapacity by lowering the emotional strain on proxies and reducing the economic costs of unwanted tests and treatments" (Lamanna and Byrne 2018, pp. 906–907). Hubbard and Greenblum (2020, p. 3217) go further, stating that it may be "better at avoiding bias and, therefore, choosing in a more patient-centered manner". They also argue that "in cases where the patient has not issued a medical power of attorney… against the standard practice of vesting familial surrogates with decision making authority, the Autonomy Algorithm should have sole decision-making authority" (Hubbard and Greenblum 2020, p. 3217). With the Autonomy Algorithm raised as a potential answer to the burdens associated with incapacity (as well as the many problems presented by traditional surrogacy), such an ADMS poses pertinent questions regarding the receptivity from the public and perceptions with regards to algorithmic decision-making in medicine.

2.2 Algorithm Aversion & Algorithm Appreciation

Even where algorithms are purported to be demonstrably superior to humans across a range of analytical tasks, when users are able choose between an algorithm and a human alternative to assist with decisions, many demonstrate a bias against algorithms (Burton et al. 2020; Mahmud et al. 2022). This tendency dates to early comparative works examining the accuracy of "clinical" (i.e., human or intuitive) versus "actuarial" (i.e., statistical or mathematical) approaches to forecasting (Dawes et al. 1989; Meehl 1954). In revisiting the bias in a more modern context, Dietvorst et al. (2015) conceptualised this tendency as "Algorithm Aversion", and raised an interesting question that is central to the phenomenon: why do people choose to use algorithms when given the option?

The initial studies by Dietvorst et al. (2015) suggested that the bias was born of poor experiences with the algorithm, such as having witnessed it underperform or fail. Specifically, the authors found that errors from an algorithm were evaluated more harshly than similar errors made by a human (Dietvorst et al. 2015). Several proposed causes for Algorithm Aversion have been examined since, however, it remains the case that "relatively little is known about the conditions that lead to the acceptance or rejection of algorithmically generated insights by individual users of decision aids" (Burton et al. 2020, p. 220). A recent systematic review suggests that a majority of the empirical literature focuses on either the attributes of the algorithm itself or psychological aspects of the user (Mahmud et al. 2022). For example, the "black box" nature of algorithms (i.e., lack of understanding how an algorithm works compared to the more readily interpretable human decision-maker) appears to drive aversion (Cadario et al. 2021). Alternatively, perceptions that algorithms are better suited for objective tasks rather than subjective tasks (Castelo et al. 2019), or a belief that algorithms are unable to account for individuals' unique characteristics (Longoni et al. 2019) have also led to people exhibiting aversion. Conversely, there are also contributions that have documented cases where algorithms are preferred to a human alternative ("Algorithm Appreciation"; Logg et al. 2019), as well as how aversion can be alleviated. For example, certain capabilities of ADMS can reduce Algorithm Aversion, such as giving users the ability to modify the algorithm and its outputs (Dietvorst et al. 2018) or demonstrating an algorithm's ability to learn (Berger et al. 2021).

The conceptualisation of Algorithm Aversion by Dietvorst et al. (2015) found its origins within the literature and context of forecasting tasks. Early works that are cited examine statistical algorithms that offer support and play an advisory role as decision aids to their human decision-maker counterparts (Dawes et al. 1989; Meehl 1954). However, the literature on the phenomenon has since expanded to far more complex and even autonomous algorithmic systems, which instead play performative roles for humans and organisations. These systems are now used, and Algorithm Aversion is now cited, across a wide range of contexts, including scenarios that may impact entire careers (Noble et al. 2021), or potentially harmful outcomes and much higher stakes, such as autonomous vehicles (Shariff et al. 2017) and medical advice or diagnoses (Cadario et al. 2021). Interestingly, differences in the values and stakes that are at play in ADMS use contexts (and what might be a large contextual leap for some uses) appears to be rarely acknowledged within this literature, and even seems to be "undersold" despite the radically contrasting ramifications these use cases present: "consequences of bad advice in medical healthcare

would probably be more severe for participants than bad advice towards a product demand forecast" (Daschner and Obermaier 2022, p. 14). Though there are some studies where moral or ethical considerations are explored (e.g., Bigman and Gray, 2018; Lee 2018), this appears to be underexamined (Mahmud et al. 2022). Having investigated as follows, we maintain that the gravity of high-stakes, value-laden decision contexts elicit very different cognitive processes for evaluating ADMS, as well as different justifications for using them, or preferring a human alternative, in a decision context. These rationalisations may focus less on the efficacy and accuracy of ADMS, and instead emphasise emotional, ethical, and experiential elements of the decision context, which also deserve to be explored.

3 Methodology

Though qualitative approaches to research are multifaceted and encompass several traditions, Creswell and Creswell (2017) suggest that they can be especially suited to exploratory research efforts. In accordance with the call from Mahmud et al. (2022, p. 15) for an initial investigation into the "underlying motives" regarding Algorithm Aversion, an inductive approach for qualitative analysis was adopted in order to better understand such motives. In addition to the direct call for qualitative work in the realm of Algorithm Aversion, the systematic review by Mahmud et al. (2022, p. 1) also suggests that the nature of the task or context itself plays in the uptake of ADMS is an area where "very little attention has been given" in the literature, and that attributes of the user or the algorithm itself have been "investigated extensively" by comparison. As such, we sought to examine the rationales and justifications for peoples' receptivity (or lack thereof) toward the use of ADMS – with the explicit intent to incorporate a focus on high-stakes value-laden decision contexts that have not been adequately addressed in the literature. Therefore, medical decision-making was selected as the context with which to investigate this potential. As a subject matter for research it also appears to be increasingly examined within the Algorithm Aversion literature, whilst still appearing to have a focus on accuracy and functioning of the algorithm in exploring aversion, which may limit the understanding of its underlying motives (e.g., Cadario et al. 2021; Castelo et al. 2019; Longoni et al. 2019). Within this research, the "Autonomy Algorithm" was selected as an example ADMS with which to gather insights by presenting the hypothetical technology to participants and analysing their responses.

Prior to data collection, ethics approval was obtained from an Australian university, and a short notice was posted on a university website that disseminates information regarding student participation in research as part of their undergraduate course credit. All participants gave consent prior to taking part in accordance with the institution's requirements and could withdraw at any time. Participants were presented with a description of the hypothetical Autonomy Algorithm according to its proponents (Lamanna and Byrne 2018; Hubbard and Greenblum 2020). With the hypothetical technology in mind, they were asked to write openly responses to the questions: *"If you have a reason, what would be the primary reason for why you [DO/DO NOT] feel comfortable with the idea of the Autonomy Algorithm in a situation where a patient is unable to make a medical decision themselves?"* Due to the exploratory nature of the research, an attempt was made to keep the request for responses as broad as possible, allowing participants to describe their initial reasoning for being reluctant, receptive, or both. It was formulated in a way that is catered specifically to the situation at hand, and only where the Autonomy Algorithm is proposed to be utilised.

The inductive approach to analysis of the open-ended responses that followed was guided by the research problem – to examine how people evaluate ADMS and some of the different rationales behind their receptivity (or lack thereof) toward them. Thus, key themes in people's language were sought to be identified as part of potential theory building and even later construct creation. Using the Gioia Method (detailed below), analysis comprised of moving iteratively between a close reading of the original language that people used and the ongoing reduction and display of findings (i.e., categorisation of themes/aggregate dimensions, data structure, and write-up) until a point where new concepts were ceasing to emerge, and thus, no new data sought from further questionnaires, leading to 252 responses to the questionnaire being used (as Gioia et al., 2013 suggests, akin to theoretical saturation). Specifically, the key stages (pictured below) from the approach put forth by Gioia et al. (2013) were used, initially generating "1st-order Terms" (or "1st-order Concepts") that attempt to adhere to the original language used in responses, whilst deriving some sense of the potential categories for ideas that were put forth. Concepts were then examined for (dis)similarities in the language and ideas raised in order to provide some further potential for categorisation and identify emergent "2nd-order Themes" that reduce overlap to provide some structure using the researchers' inductive reasoning. Direct quotations that convey each of these concepts and themes were also noted, that would later demonstrate "data-to-theory connections" as an inclusion within the demonstration of research findings. Finally, the themes were distilled further into "aggregate dimensions" as a final form of categorisation where possible. At this

point, a "Data Structure" (Figure 1) can be constructed in order to provide a clear demonstration of the analysis process, as well as a visual aid for the discussion that follows.

4 Research Findings

In examining participants' responses to the Autonomy Algorithm as an example ADMS used in a high-stakes, value-laden decision context, several concepts and themes emerged from what participants highlighted as their reasoning for or against its potential use in medicine and surrogate decision-making. These themes were distilled into one of four aggregate dimensions: Functional "Aversions", Philosophical "Abstinence", Practical "Appreciations", and Personal "Appeasements". Whilst each aggregate dimension is discussed below, not all 1st-order Concepts and 2nd-order Themes are covered due to page constraints.

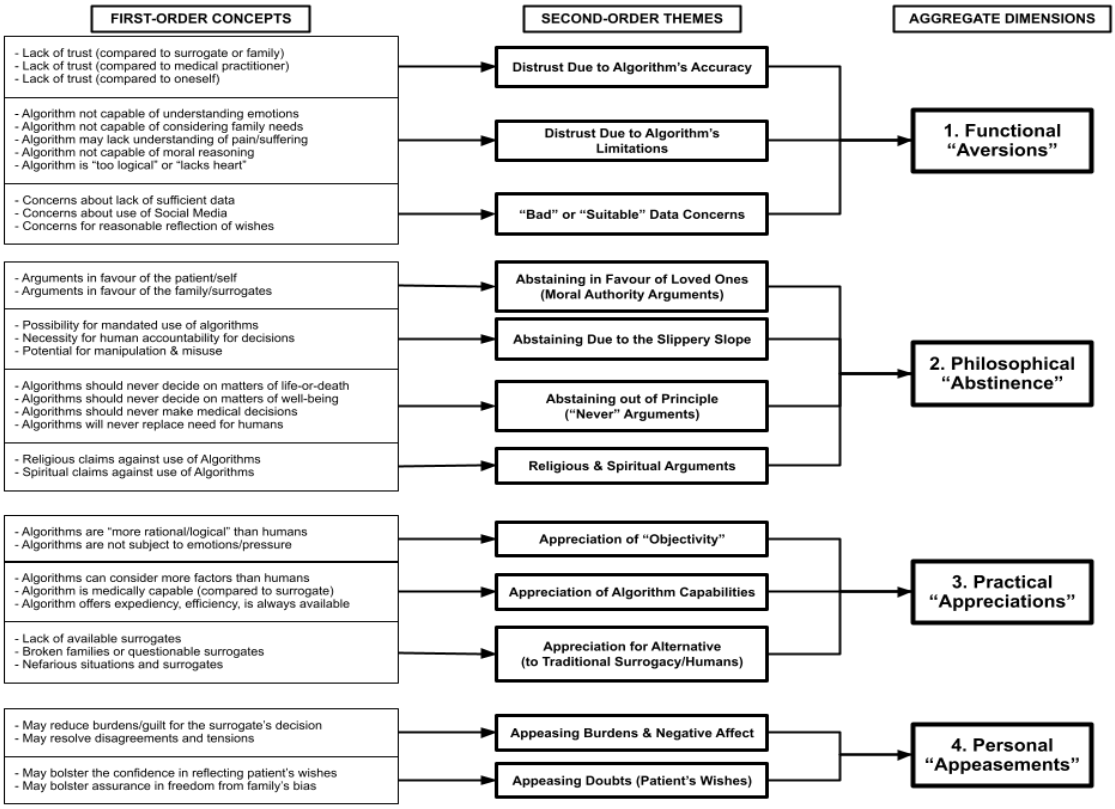

Figure 1: *Data Structure: Reactions to the Autonomy Algorithm*

4.1 Aggregate Dimension 1: Functional "Aversions"

The first aggregate dimension encompasses participants' concerns surrounding the algorithm and its function or feasibility. This aggregate dimension broadly aligns with how "Aversion" is currently examined in the Algorithm Aversion literature – that is to say that the focus for such themes centre around the efficacy of the technology, how it works, or its perceived utility for the task at hand.

Theme: Distrust Due to the Algorithm's Accuracy – A prominent theme emerging from the concepts raised by participants is simply a distinct lack of trust or a disbelief in the Autonomy Algorithm making accurate decisions. Participants cite concerns of inaccuracy: *"No matter how perfect the machine is, there will be times when it makes wrong decisions, or the machine has a certain error rate"*; even directly comparing it to loved ones acting as surrogates: *"Lack of trust in that an algorithm can predict what I would want compared to friends/family who have known me my whole life"*. Even where participants acknowledge that humans are also imperfect decision-makers, some voice disbelief in the efficacy of the algorithm's predictive capabilities: *"I am unsure whether the Autonomy Algorithm would be able to accurately predict my future decisions – after all, my values and opinions could change over time and the past is not necessarily the best indicator for future decisions"*. Such sentiments also apply when thinking toward the future and any possible feasibility, with some expressing that the Autonomy Algorithm will *"never be an accurate reflection of my choices"*.

Theme: Distrust Due to the Algorithm's Limitations – Some participants specifically raise the non-human nature of the algorithm, and how it lacks certain capacities compared to humans in the same context: *"I don't think I would feel comfortable using AI to make my medical decisions since they lack in the "heart" or the emotions, so [it] can only make logical decisions which I don't believe as a good way to live"*. Participants appear to speak to aspects of its function as limitations, rather than as merits, as it may not account for all relevant factors within these decisions: *"As the [Autonomy Algorithm] will predominantly rely on facts, this disregards the emotional aspects required when determining these situations"*. Some also make direct comparisons with a human alternative: *"I hate this idea. There are so many factors, from life experience to social cues and context, when considering someone's medical treatment options. We can never completely replicate these nuances of human consciousness in any robot or AI system…"* They appeal to the merits of human decision-makers' capabilities, rather than that of an algorithm: *"[The algorithm] is not human, it is a system. It does not have human's feeling and consideration…"*; suggesting it does not have the appropriate capabilities to make such decisions.

4.2 Aggregate Dimension 2: Philosophical "Abstinence"

The second aggregate dimension reflects moral and ethical reasoning for participants' apprehension toward the Autonomy Algorithm. Some of the emergent language and concepts relate to why some might even "Abstain" from its use, based on moral disapproval. Some appear to object to an algorithm in the situation, describing why it may not be permissible, or may even be "wrong", rather than simply not possible or useful. That is to say that some are reluctant because the use of such a system is *inappropriate* for the task, rather than *inadequate* (as is the case for Functional "Aversions").

Theme: Abstaining in Favour of Loved Ones (Moral Authority Arguments) – Some discuss the concept of traditional surrogacy, in favour of loved ones over an algorithm based on the moral authority of intimates. Some voiced their concerns that family or friends would be indispensable in this situation: *"…Nothing can replace the feelings and consideration a loved on [sic] would have for someone"*. Some even indirectly disagree with the notion put forth by the Autonomy Algorithm's proponents (i.e., patient's wishes would be respected above all and against family bias) in favour of the family's wishes: *"At the point of unconsciousness and not knowing if you would wake from a coma, the feelings of your family and friends would become more important than your own self-interested reasons and therefore they should make the decisions not [Autonomy Algorithm]"*.

Theme: Abstaining out of Principle ("Never" Arguments) – Several participants appeal to personally held principles regarding the adoption of a system like the Autonomy Algorithm. For example, some participants suggest that such use cases would "never" be appropriate, irrespective of the circumstances: *"Machines or computers can never replace the human mind"*; and some were even specific to the use of ADMS in medicine, or decisions with high stakes surrounding health: *"In my opinion, machine[s] should never make decisions relating to well-being of people"*.

Theme: Abstaining Due to the Slippery Slope – Some specifically address concerns about the "slippery slopes" of vesting too much trust in technology for important decisions. Some of the language used by participants is surprisingly foreboding: *"…One need only look towards dystopian literature to see the great fallibility of machines and the great potential for disaster in relying too heavily upon them"*; with participants alluding to the potential for malicious use: *"…it's scary to think it could be manipulated for profit, or other benefits"*; or a dystopian mandate: *"The decision made by [the autonomy algorithm] will be applied to the patient with mandatory requirement"*. Participants even raise ethical issues that are present within the literature surrounding ADMS, such as accountability: *"Need accountability for who made the action"*; and even bias: *"Based on the information given, I would not feel comfortable that the AI could accurately assume a person's decision making process, instead of relying on stereotypes of the sociological characteristics of a person"*.

4.3 Aggregate Dimension 3: Practical "Appreciations"

The third dimension emerged from participants demonstrating an "Appreciation" of the potential merits of ADMS, and where the Autonomy Algorithm may be helpful in service of medical decision-making based on evaluations of potential utility or its function and capability.

Theme: Appreciation of "Objectivity" – Some in favour of the ADMS emphasise the perceived merits of algorithms. Some speak about "logic" and "rationality" as positive aspects of algorithms: *"Machine is more rational than humans and could make a better or logic [sic] choice"*; or even a sense of neutrality: *"Make the most correct and neutral choice"*. Some appeal to an algorithm's "rationality" as ideal, stating that human emotions may obstruct decision-making and that the Autonomy Algorithm would be better *"Because it removes emotion from the equation. If the algorithm is making the decision*

it is likely to be based of off [sic] facts not emotions"; even in what are innately very emotionally straining, high-stakes decision contexts like medicine: *"…It makes sense to remove all irrationality or emotional thinking that can cloud decision-making in a dire medical situation"*. Some focus on the idea of making the most medically correct decision, and how the algorithm might service this, appealing to "objectivity" as a merit: *"I think the biggest benefit of Autonomy Algorithm is its objectivity… Autonomy Algorithm can make objective choices rather than arbitrary subjective choices…"*; and some suggest that this approach is potentially the foundation for "better" decision-making: *"It's more logical, providing suggestions with lots of statistics and can help us make better decisions"*.

Theme: Appreciation of the Algorithm's Capabilities – Contrasted to similar grounds for the Functional "Aversion" to the Autonomy Algorithm's capabilities, several participants highlight the capacities of ADMS as favourable. This includes the types of considerations that ADMS might be able to make related to medical data: *"Because it is medically trained while my family is not"*; or the sheer amount of data that ADMS can consider: *"…It would also be able to consider many more factors that the human brains of the family would not be able to do physically"*; as well as more specific examples of facts that a surrogate or family may not consider: *"Maybe the patient has hiddent [sic] values that they have not shared with any of their friends or family because of cultural pressures that the Autonomy Algorithm may be able to pick up and thus be able to more accurately make the medical decision"*. Considerations are also made for efficiency: *"If I feel comfortable with the idea of Autonomy Algorithm, the possible reason is that it can help and is fast to make a decision"*; including where it is crucial: *"In times of emergency, [the Autonomy Algorithm] may be more useful when doctors, practitioners, or other medical experts are unavailable…"*

Theme: Appreciation for an Alternative (to Traditional Surrogacy) – Several participants that favour the Autonomy Algorithm also highlight specific situations where it may be useful. Primarily, this centres around cases where traditional surrogate decision-making would not be possible, such as in the absence of loved ones: *"I feel that the Autonomy Algorithm could be useful in situations where patients are without family, or appropriate power of Attorney"*; even if participants are personally against using the algorithm: *"It would only be ok if the incapacitated patient had no family or friends to make a decision for them"*. Some participants that are sceptical still vocalise a justification for the algorithm, particularly as a last resort: *"It would be different if I had no close family members, in which case a poor algorithm would be better than nothing"*. Situations in which some see merit are more specific, such as due to problematic family dynamics: *"… [patients] may not be in a situation where leaving the medical decision up to their family is the best idea, i.e., abusive or broken families, fostering, etc."*

4.4 Aggregate Dimension 4: Personal "Appeasements"

The fourth aggregate dimension revolves around the potential merits of the Autonomy Algorithm in order to "Appease" and pacify doubts or negative affect in value-laden decision-making. This might be due to the further confidence it may instil in decision-makers (e.g., surrogates and their decisions), or in assuring that the ultimate outcome is, in fact, a reflection of the patient's wishes.

Theme: Appeasing Burdens & Negative Affect – Some speak of the pressures that are associated with surrogacy, much like proponents of the Autonomy Algorithm: *"I think it relieves some of the pressure from family members…"* They discuss the potential alleviation of guilt or blame that some of these pressures might induce in the family where they are unsure of the right decision: *"…It also takes away the feeling of guilt or pressure to make the 'right' decision and working out what they would want so this is a good idea"*. Participants were also conscious that the Autonomy Algorithm might instil confidence in decisions, or resolve disagreements: *"Stops arguments from the family and the difficult decision"*; or even assist with far more difficult situations where there may be bias from surrogate(s): *"When there have been reported tensions or disagreements between [the patient] and the person who wants to make a decision for them e.g., family member"*.

Theme: Appeasing Doubts About the Patient's Wishes – Some participants highlight that the Autonomy Algorithm might in fact be free from the influences of the family, and primarily in favour of the patient's wishes (similar to its proponents). They echo the Autonomy Algorithm's proposed patient-centred focus: *"Would only have the interests of the patient in mind"*; and that it may be able to shed light on matters that even loved ones are unsure of: *"If it is completely able to predict personal values and preferences, and people have complete faith in that, then why not use AI. It would allow families and others to feel like the right decision is being made based on the patient's own preferences"*; *"…for people who don't share any personal information or thought with no one, maybe this algorithm would end up knowing them better and hence taking [sic] better decision [than] their close people"*; with some even speaking out for the patient at all costs: *"Even [if] it's the patient's family or close friends they don't have the inalienable right to decide the destiny for him…"*

5 Discussion

In serving the aims of this paper – to investigate the ways in which people evaluate ADMS and the rationales behind their receptivity (or lack thereof) toward them – several distinct justifications for the reluctance and receptivity toward using the ADMS emerged. We examined this using the language from responses to a potential future use of ADMS in the high-stakes, value-laden context of surrogate medical decision-making. The summarised findings suggest that there are rationalisations and justifications that exist completely outside of the feasibility or potential utility provided by an ADMS – especially when one's values are at play. By reflecting on the participants' reasoning, and the differences in either utility-focused or value-laden assessments of the context and ADMS use case, the theoretical and practical understanding of what is currently conceptualised as "Algorithm Aversion" can be expanded using new dimensions, and a series of themes within each. These dimensions provide a greater resolution for the types of reactions to an understanding of the uptake of ADMS, dependent on the primary type of evaluation of the algorithm and associated decision context. They can be further summarised using the overarching framework below (Figure 2). This conceptualisation focuses on a participant's primary rationale or evaluation (characterised as either "Utility-based" or "Value-laden") for being either "Reluctant" or "Receptive" toward an ADMS in a given context. Using this view may determine which dimension is most applicable to understanding the basis of people's response and potential uptake of ADMS. Thus, greater understanding of the motives beneath the phenomenon, as well as construct creation for a measure as part of upcoming research, may benefit accordingly.

	Utility-based Evaluation	**Value-laden Evaluation**
Reluctant	Algorithm Aversion	Algorithmic Abstinence
Receptive	Algorithm Appreciation	Algorithmic Appeasement

Figure 2: *Conceptual Distinctions for Reactions to ADMS*

These conceptual distinctions also assist in addressing potential limitations and biases that may exist within the Algorithm Aversion literature currently. A broader systematic review of the extant literature on the phenomenon (Mahmud et al. 2022, p. 1) frames the adoption of algorithms as a "rational behaviour", and conversely, aversion to them as a "behavioural anomaly". The empirical literature also typically seeks to explore means by which aversion can be mitigated, often assuming it as an obstacle to overcome in service of "better" decisions, otherwise it "may reduce human subjects' expected utility" (Mahmud et al. 2022, p. 1). To some extent, this might be born of the fact that a large portion of Algorithm Aversion literature appears within research from business disciplines and forecasting, where managerial implications typically service strategic decision-making and organisational objectives. However, drawing upon Bauer's (1995, p. 2) notions on "resistance", this poses risks of "managerial and modernisation bias" toward the adoption of technology, where the focus remains on "progress" rather than any adverse consequences associated with such developments. In line with Bauer (1995, p. 413), it may be important to acknowledge where "resistance" to new technology may play a crucial, constructive role – where it is not "a deficit, but a resource for the development of technology… that indicates the need for alterations, and points to the kind of alterations that are required".

Further to the call for qualitative insights to understand the motives beneath Algorithm Aversion, the potential bias toward utility may also mean misattributing one's rationale for the rejection of ADMS to something that it is not. Attributing one's reluctance toward algorithms simply to a lack of trust in its expected performance or utility may be inappropriate in cases where evaluations are made based on more emotional or ethical grounds. Arguably, the Algorithm Aversion literature does not adequately account for the role of such subjectivity, stakes, and one's values, which might be intertwined with certain decision contexts, rendering the use of ADMS as unsuitable or inappropriate for some users (despite being capable). In such scenarios, and for such users, it may become a question of whether algorithms *"should"* make decisions, rather than whether they are *"good"* at making decisions. Thus, it may be necessary for the literature to further acknowledge the roles of such elements, and the deliberate decision to *"object"* to algorithms by abstaining, rather than *"reject"* them, based on value-laden assessments. The same applies to the "appeasement" that ADMS may offer to some users in difficult value-laden contexts, where what is "accurate" may not be of concern or even possible to determine (rather what is "right"), but where ADMS still offer some reassurance for decisions. In situations with such gravity that relate to ethics, values, affect, and high stakes, the focus of the current theorisation for

Algorithm Aversion (surrounding trust and "rational" assessments of utility due to an algorithm's efficacy) may not be sufficient alone.

As with any study, this research has some limitations that more extensive qualitative work on the evaluations of ADMS may improve upon. Like much of the research to date, this study was conducted on lay people, and furthermore, was limited to students. There is undoubtedly a need to conduct qualitative work on the opinions of experts in various contexts, as well as examinations of applied algorithmic systems and decision contexts in real-world settings, in order to validate the claims made by experimental contributions (Mahmud et al. 2022). Further research may also examine a longer list of practical uses and scenarios in which algorithmic agents are currently being used or expected for future use that might pose similar ethical objections raised presently. This conceptualisation and framework are also not suggesting that several decision contexts will not have multiple considerations existing behind users' evaluations of algorithms, and several elements co-contributing to reluctance or receptivity. That is, users' evaluation can be both utility-focused and value-laden at the same time. However, it is anticipated that these conceptual distinctions might assist in future explorations of the phenomenon, using a deeper understanding of its "underlying motives" (Mahmud et al. 2022, p. 15).

6 References

Arnold, M. H. 2021. "Teasing out Artificial Intelligence in Medicine: An Ethical Critique of Artificial Intelligence and Machine Learning in Medicine," *Journal of Bioethical Inquiry* (18:1), pp. 121–139. (doi: 10.1007/s11673-020-10080-1).

Bauer, M. W. 1995. *Resistance to New Technology: Nuclear Power, Information Technology and Biotechnology*, (M. W. Bauer, ed.), Cambridge, UK: Cambridge University Press.

Berger, B., Adam, M., Rühr, A., and Benlian, A. 2021. "Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn," *Business & Information Systems Engineering* (63:1), pp. 55–68. (doi: 10.1007/s12599-020-00678-5).

Bigman, Y. E., and Gray, K. 2018. "People Are Averse to Machines Making Moral Decisions," *Cognition* (181), pp. 21–34. (doi: 10.1016/j.cognition.2018.08.003).

Brock, D. W. 1991. "The Ideal of Shared Decision Making Between Physicians and Patients," *Kennedy Institute of Ethics Journal* (1:1), pp. 28–47. (doi: 10.1353/ken.0.0084).

Burton, J. W., Stein, M.-K., and Jensen, T. B. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making," *Journal of Behavioral Decision Making* (33:2), pp. 220–239. (doi: 10.1002/bdm.2155).

Cadario, R., Longoni, C., and Morewedge, C. K. 2021. "Understanding, Explaining, and Utilizing Medical Artificial Intelligence," *Nature Human Behaviour* (5:12), pp. 1636–1642. (doi: 10.1177/2053951718756684).

Castelo, N., Bos, M. W., and Lehmann, D. R. 2019. "Task-Dependent Algorithm Aversion," *Journal of Marketing Research* (56:5), pp. 809–825. (doi: 10.1177/0022243719851788).

Creswell, J. W., and Creswell, J. D. 2017. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, SAGE Publications.

Daschner, S., and Obermaier, R. 2022. "Algorithm Aversion? On the Influence of Advice Accuracy on Trust in Algorithmic Advice," *Journal of Decision Systems*, pp. 1–21. (doi: 10.1080/12460125.2022.2070951).

Dawes, R. M., Faust, D., and Meehl, P. E. 1989. "Clinical Versus Actuarial Judgment," *Science* (243:4899), pp. 1668–1674. (doi: 10.1126/science.2648573).

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General* (144:1), pp. 114–126. (doi: 10.1037/xge0000033).

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science* (64:3), pp. 1155–1170. (doi: 10.1287/mnsc.2016.2643).

Erevelles, S., Fukawa, N., and Swayne, L. 2016. "Big Data Consumer Analytics and the Transformation of Marketing," *Journal of Business Research* (69:2), pp. 897–904. (doi: 10.1016/j.jbusres.2015.07.001).

Gillespie, T., Boczkowski, P. J., and Foot, K. A. 2014. *Media Technologies: Essays on Communication, Materiality, and Society*, MA, MIT Press.

Gioia, D. A., Corley, K. G., and Hamilton, A. L. 2013. "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology," *Organizational Research Methods* (16:1), pp. 15–31. (doi: 10.1177/1094428112452151).

Hubbard, R., and Greenblum, J. 2020. "Surrogates and Artificial Intelligence: Why AI Trumps Family," *Science and Engineering Ethics* (26:6), pp. 3217–3227. (doi: 10.1007/s11948-020-00266-6).

Kennedy, R. P., Waggoner, P. D., and Ward, M. M. 2022. "Trust in Public Policy Algorithms," *The Journal of Politics* (84:2), pp. 1132–1148. (doi: 10.1086/716283).

Lamanna, C., and Byrne, L. 2018. "Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm," *AMA Journal of Ethics* (20:9), pp. 902–910. (doi: 10.1001/amajethics.2018.902).

Lee, M. K. 2018. "Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management," *Big Data & Society* (5:1), pp. 1–16. (doi: 10.1177/2053951718756684).

Logg, J. M., Minson, J. A., and Moore, D. A. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151), pp. 90–103. (doi: 10.1016/j.obhdp.2018.12.005).

Longoni, C., Bonezzi, A., and Morewedge, C. K. 2019. "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research* (46:4), pp. 629–650. (doi: 10.1093/jcr/ucz013).

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., and Smolander, K. 2022. "What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion," *Technological Forecasting and Social Change* (175), 121390. (doi: 10.1016/j.techfore.2021.121390).

McDougall, R. J. 2019. "Computer Knows Best? The Need for Value-Flexibility in Medical AI," *Journal of Medical Ethics* (45:3), pp. 156-160. (doi: 10.1136/medethics-2018-105118).

Meehl, P. E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.*, Minneapolis, US: University of Minnesota Press.

Meijerink, J., and Bondarouk, T. 2021. "The Duality of Algorithmic Management: Toward a Research Agenda on HRM Algorithms, Autonomy and Value Creation," *Human Resource Management Review*, 100876. (doi: 10.1016/j.hrmr.2021.100876).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. 2016. "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* (3:2), pp. 1–21. (doi: 10.1177/2053951716679679).

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'," *The Journal of Strategic Information Systems* (24:1), pp. 3–14. (doi: 10.1016/j.jsis.2015.02.001).

Noble, S. M., Foster, L. L., and Craig, S. B. 2021. "The Procedural and Interpersonal Justice of Automated Application and Resume Screening," *International Journal of Selection and Assessment* (29:2), pp. 139–153. (doi: 10.1111/ijsa.12320).

Obermeyer, Z., and Emanuel, E. J. 2016. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *The New England Journal of Medicine* (375:13), pp. 1216–1219. (doi: 10.1056%2FNEJMp1606181).

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, NY, Crown Books.

Shariff, A., Bonnefon, J.-F., and Rahwan, I. 2017. "Psychological Roadblocks to the Adoption of Self-Driving Vehicles," *Nature Human Behaviour* (1:10), pp. 694–696. (doi: 10.1038/s41562-017-0202-6).

Trocin, C., Mikalef, P., Papamitsiou, Z., and Conboy, K. 2021. "Responsible AI for Digital Health: A Synthesis and a Research Agenda," *Information Systems Frontiers*, pp. 1–19. (doi: 10.1007/s10796-021-10146-4).

Copyright