ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems

Thomas Baudel, Manon Verbockhaven, Victoire Cousergue, Guillaume Roy, Rida Laarach

▶ **To cite this version:**

Thomas Baudel, Manon Verbockhaven, Victoire Cousergue, Guillaume Roy, Rida Laarach. ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems. 18th IFIP Conference on Human-Computer Interaction (INTERACT), Aug 2021, Bari, Italy. pp.300-320, 10.1007/978-3-030-85613-7_22 . hal-04292393

HAL Id: hal-04292393

https://inria.hal.science/hal-04292393v1

Submitted on 17 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems

Thomas Baudel[1][0000-0002-7505-2505] , Manon Verbockhaven[1,2], Victoire Cousergue[1,3], Guillaume Roy[1,4] and Rida Laarach[1,5]

[1] France Lab, IBM, Orsay, France
[2] ENSAE, France
[3] Université Paris-Dauphine & Mines ParisTech, France
[4] ENSAI Rennes, France
[5] Telecom ParisTech, France
baudelth@fr.ibm.com

Abstract. Business process management organizes flows of information and decisions in large organizations. These systems now integrate algorithmic decision aids leveraging machine learning: each time a stakeholder needs to make a decision, such as a purchase, a quote, or hiring someone, the software leverages the inputs and outcomes of similar past decisions to provide guidance, as a recommendation. If the confidence is high, the process may be automated. Otherwise, it may still help provide consistency in the decisions. Yet, we may question how these aids affect task performance. Can we measure an improvement? Can hidden biases influence decision makers negatively? What is the impact of various presentation options? To address those issues, we propose metrics of performance, automation bias and resistance. We validated those measures with an online study. Our aim is to instrument those systems to secure their benefits. In a first experiment, we study effective collaboration. Faced with a decision, subjects alone have a success rate of 72%; Aided by a recommender that has a 75% success rate, their success rate reaches 76%. The human-system collaboration had thus a greater success rate than each taken alone. However, we noted a complacency/authority bias that degraded the quality of decisions by 5% when the recommender was wrong. This suggests that any lingering algorithmic bias may be amplified by decision aids. In a second experiment, we evaluated the effectiveness of 5 presentation variants in reducing complacency bias. We found that optional presentation increases subjects' resistance to wrong recommendations. We intend to leverage these findings to guide the design of human-algorithm collaboration in financial compliance alert filtering.

Keywords: Business decision systems, Decision theory, Cognitive biases.

1 Introduction

For the past 20 years, Business Process Management (BPM) [29] has streamlined processes and operational decision-making in large enterprises, transforming work

organization. BPM organizes information flows, from input events such as a purchase order or a hiring request, through chains of stakeholders involved in various parts of the process, to reach some outcome. In a nutshell, a BPM system allows programming an enterprise like one would a robot.

Because all inputs and operational decisions are stored, recent improvements involve applying machine learning techniques on past inputs and outcomes, to automate decision processes or to assist decision makers by providing suggestions on the likely outcome of each decision, assuming similar causes produce similar effects [39,53]. This technology ought to be largely beneficial, reinforcing the consistency of decisions and improving productivity, enabling "extended cognition", or "active externalism" as described by Chalmers [14], or as envisioned as "Human-Centered AI" by Schneiderman [52]. Now, augmenting decision processes is not without risks. There is a large institutional community focusing on the area of AI Ethics, that stresses the requirements of fairness, transparency, explainability, accountability [28]… In particular, the prevention of algorithmic biases is a major concern [42], which needs to be addressed by technology [6], best practices [3] and regulations [37]. There is less institutional visibility on the changes to work practices and human decision-making these tools introduce. Cognitive biases induced by decision support systems are largely studied, identifying patterns such as automation and complacency biases [5] or, on the contrary, algorithm aversion [9] and decision fatigue [27, 4] when higher productivity is expected. There is less work directly applicable to the context of business decision-making, in our present situation, where decision aids can be provided as a generic, system-level feature, regardless of their relevance for the task being performed by the human agent.

To ensure this type of assistance can be safely and profitably incorporated in business decision support systems, we first narrow down the type of tasks we are interested in augmenting, and the type of aids we wish to evaluate. Then, we review the literature to guide our designs. We present a performance and cognitive biases evaluation model, as a set of metrics that can be evaluated empirically for various kinds of decision aids and tasks. We conduct a study to evaluate the ability of our model to measure performance and biases. Finally, we propose a methodology to incorporate bias measurement and compensation in business decision support systems, to ensure they provide their expected benefits with minimal inconvenience.

2 Augmented Business Decision Making

A business process is modeled in a flowchart. In this model, decision steps take as input some information, leverage external information, such as regulations or resource constraints, presumably only known to the decision maker, to advance the process through a predetermined set of possible outcomes. The type of decision tasks we in-

vestigate are constrained, leave little room for creativity and presumably rely on a combination of explicit rules and heuristics.

When the decision logic can be formally expressed, decisions can be automated deterministically, via deduction [10]. When sufficiently robust heuristics are available, scoring methods or more sophisticated algorithms may automate the decision via induction, with an escalation process to handle exceptions. Finally, many operational business decisions involve complex tradeoffs, which, we assume, involve so many factors that automation is for now out of question. In these cases, a decision aid fed with previous decisions and possibly external data sources, such as revenue targets for the team and staffing statistics, may provide information that can be useful, for instance to a manager who considers granting a hiring request.

More formally, we are interested in measuring, and possibly improve, the effects decision aids may have in the following circumstances:

- Some information regarding a case is available and is assumed to be reliable.
- A choice, among a predefined set of alternatives, needs to be made.
- The choice is partly constrained by explicit constraints (regulations, resource limitations…)
- Some contextual information, exact or probabilistic, explicit or intuitive, is available to the decision maker (e.g. guidelines & priorities regarding the business context). Those allow the decision maker to form an *internal decision model*, which is a non-explicited procedure a priori followed consistently by the decision maker.
- Other information is known only by the system, such as a history of cases.
- There is a best possible choice, but it may not be knowable in advance, if ever, for a given case.
- It is possible to provide, a posteriori, exact or approximate measures of which choices tend to perform better for which types of cases, for instance as an actuarial report. These measures can be used to create a *computable decision model*.

Numerous business situations implemented with BPM match this description, ranging from the mundane, such as a purchase decision, to more serious decisions such as hiring someone, granting a loan or selecting the winner of a bid, finally to morally heavy situations like deciding of a prison sentence [49]. Still, much decision-making activity, such as medical diagnostic or investment decisions, allow more creativity in choice-making and falls outside the scope of our work. What matters in our circumstances is that no algorithm may deliver a choice with 100% accuracy. An algorithm may be better than humans in general, but there is no substitute for human judgment and liability, when it is not possible, even a posteriori, to know if a particular decision was the right one.

A variety of decision aids can be automatically provided to users in this context, which we categorize, inspired from Miller's description of explanations [41]:

- Attributes deemed most important in the decision (inputs)
- Scoring (computed by deterministic rules, deduction)
- Comparable cases and their outcomes (nearest neighbors, induction)
- Decision tree branch and weights (probabilistic deduction).
- Counterfactuals (abduction).

While there may be more types of decision aids, these cover the use cases we have reviewed in the literature that can be implemented generically, without specific tuning for the decision task, for instance creating a custom visualization. Finally, the presentation of these decision aids may influence their usage. They may be provided as plain recommendations, inciting the user to follow them, or available only on request, after a delay, or even after the decision has been made, as a verification step.

Our goal is to assess if decision aids improve the decision-making process, moving it towards a definition of rational decision-making suitable to our context. For now, we focus on performance metrics:

- Can we provide a methodology to measure decision-making performance in our contexts of use?
- Can the combination of human and algorithmic aids outperform both the human and the algorithm taken alone? How can we reach this stage of human-machine "collaboration"?
- Machine learning biases are a major concern. Even if this human-machine collaboration is an improvement, it seems inevitable that underlying algorithmic biases may taint decisions. Various cognitive biases may interfere. Can we identify and separate those to design correction strategies targeted at each of them?

To address those questions, we propose a model, as a set of metrics, to define performance, various biases and resistance, and carry an experimental study on a simple decision task meeting our requirements to assess the capacity of the model to discriminate various presentations modes. The aim is to generalize this study in our contexts and provide continuous monitoring of decision tasks and decision aids. But before, we must acknowledge that there is a large body of literature addressing similar issues, which has guided our research.

3 Related Work

3.1 Decision Theory

Understanding decision-making is a full research area in psychology. For a start, there are several positions regarding the notion of "correct" decision. For Rational Decision Theory, a rational choice maximizes an expected utility function [26, p.237]. Non rational theories [23] consider effects such as risk aversion or naturalistic viewpoints. Ultimately, these approaches can be reconciled when assumptions are clearly stated [30]. We take mostly a Rational Decision Theory standpoint: simple hypotheses are acceptable in our context. We also assume good will: the decision maker and the decision stakeholders (the company) share the same utility. Under stress, pressure, or poor motivation, we may find a divergence, which leads to a complacency bias or decision fatigue [4]. Within decision theory, our work falls into the area of judge-advisor sys-

tems [8]. Although much of the literature in this area is focused on human advisors, we retain the importance of advice presentation on the decision outcomes [38].

A major lesson of Decision Theory is that performance varies between individuals: experts and novices approach problems differently, and personality traits can have a strong influence [48]. Classical cognitive biases such as the order effect can be significant [47]. Finally, the availability of more information does not necessarily lead to better decision-making [17]. For fast and frugal or other recent approaches [25, 23], human decision making does not rely so much on *risk* - when the decision rules and probabilities of outcomes are well modeled - than on *uncertainty* - where the indecisiveness is not just a consequence of unknowns quantities, but also of unknows on the suitability of the decision process itself -. Hence, making a decision heuristically, based on limited information, may yield better results than paying close attention to possibly irrelevant details. Burton [9] convincingly defends that, by design, algorithmic decision aids operate under models of *risk*, while humans need to consider the *uncertainties* of a situation. Consciously leveraging this difference may provide the means to make the best of human and algorithm complementarity.

3.2 Decision Support Systems

Algorithmic decision aids have existed for a long time, in slightly different contexts.

Semi-automation/process control: Our early focus was on measuring, and possibly reducing, automation bias, which we felt should occur in our context. A large portion of the literature on these biases focuses on tasks that involve less dedicated attention and analysis than business decisions. For instance, [44] finds attentional deficits at the onset of complacency biases. [5] identify this bias in process control tasks that involve verification rather than true decision making. Automation bias is clearly related to complacency bias. It can also result from attention deficits, which are even more prevalent in assisted driving tasks [21]. [2] finds a conformity bias: use by others increases trust in an algorithm. Finally, [24], [55] and [1] describe a major cause of biases: when the algorithm outperforms the human most of the time, motivation necessarily dwindles. Conversely, [43] provides some evidence that decision aids lose all usefulness when they provide less than 70% accuracy, which indicates that providing those in generic business processes requires some prior assessment of relevance.

Recommender Systems are algorithmic decision aids, but the tasks they support does not meet our focus: they don’t help making a choice among predetermined outcomes, and decision quality is elusive: we often assimilate decision performance with user satisfaction [32]. Several types of cognitive biases can interfere, such as exposure bias [31]. Interestingly, recent literature in decision theory for recommender systems focuses on algorithm aversion, the opposite of automation biases: subjects avoid following recommendations, even when the algorithms perform better than humans [9]. The literature stresses the need to provide explanations [54, 46], especially for expert users

[33]. We also notice that trust in the system degrades when bad recommendations occur [45], or when the task is highly subjective [13]. These findings should apply to our context of use.

Visual Analytics for decision making is an entire class of algorithmic decision aids. Visualization tools are geared towards exploratory analysis, which involves open decisions [20], thus is not in our focus. Still, work on identifying and reducing cognitive biases, such as the attraction effect [18] is relevant. More significant, the identification of the numerous biases [19] that may be found in visual decision aids provides a useful guiding framework.

3.3 Impact of algorithmic decision aids on work practices, and ethical considerations

Medical decision support systems include decision aids and are extensively studied. Once again, they do not enter our scope, as the type of decision aids they provide are highly customized for specific purposes, and a medical decision can hardly qualify as a choice among predetermined possible outcomes. They support critical decisions, which explains why decision aids may be met with suspicion and some level of algorithm aversion [11]. Even in successful propositions [12], suspicion may not arise from the tool itself, but from the way it transforms work practices, perhaps for the better, but in directions which open avenues for high uncertainty [15]. At this point, addressing both algorithmic and cognitive biases in decision aids reaches beyond the scientific undertaking, into the realm of professional ethics. Determining the role of algorithmic decision aids requires a rigorous assessment of their power and their limitations in benefiting society.

From scientific to ethical considerations. Health professions are not the only industry questioning the impact of decision aids on work practices. Morris [50] devotes a whole chapter on cognitive biases in decision-making for accounting, and how overcoming those biases is an ethical issue. Legal and administrative professions raise similar concerns [16], including how proper explanations improve the perception of legal decisions [7]. Decision automation also produces surprising adaptive behaviors to circumvent the loss of control associated with algorithm-driven activities. For instance, [40] shows that human players learn to control computer players in computer games. Kyung Lee [35] describes how Uber drivers use collaboration to understand and regain some level of control over the dispatch algorithms. One of the possible shortcomings of algorithm-driven decision making (and decision aids) is the potential to induce unwanted behaviors by reverse engineering the decision aids' logic: the decision maker becomes 'controlled' by the subjects impacted by those decisions.

To address these concerns, expert groups and regulatory bodies provide guidelines [28, 3] to design trustworthy decision support systems. *But in our context, we need more than design guidelines and regulations. We need metrics,* and possibly a partially automated method to ensure those metrics stay within safe boundaries in the variety

of contexts were algorithmic decision aids will be generically embedded in business decision support systems. As a first step towards this ambitious goal, we propose some metrics applicable to business contexts, and we have conducted an empirical study on a simple task to assess their discriminatory power, in the hope of generalizing them to real-world tasks.

4 A performance model for decision-making

In machine learning, performance is most often synthesized by the F_β score, where the β term translates the cost differences between false positives and false negatives, and F is the harmonic mean of the precision and recall of the classifier. Our target users, actuaries, financial analysts, budget planners and managers are more accustomed to a simple cost model, that includes cost of processing, and follows an equation similar to (for a binary decision):

$$P = (1\text{-}E_n)\ G_n + (1\text{-}E_p)\ G_p - C_pE_p - C_nE_n - C_t$$

Given a cost matrix:
C_p : $Cost\ of\ misclassification\ of\ a\ positive$
C_n : $Cost\ of\ misclassification\ of\ a\ negative$
G_p : $Positive\ identification\ gain$
G_n : Correct n$egative\ identification$ gain
C_t : $Average\ cost\ of\ treatment$ (typically human time and amortized time to develop the decision aid solution).
And some error terms:
E_n, E_p..: $estimated\ error\ rate\ in\ the\ proposed\ configuration$

In a given situation, the equation should be weighted by the expected occurrence ratio of positives and negatives, or other priorities, such as keeping decisions homogenous rather than maximizing value. In our HCI context, the only parameters we may control are the error rates, and thus, we will assimilate performance with a strict reduction in the error rates, while neglecting, for now, C_t.

4.1 Decision aid effectiveness & collaboration

Because we are interested in measuring the performance of decision aids, we define first a measure of human-algorithm collaboration. This is quantified as the ratio:

M_1 = performance with a decision aid /
max (performance without a decision aid, performance of the classifier).

This ratio depends on a lot of factors:

- If a classifier clearly outperforms humans, there is little chance for this ratio to be > 1, full decision automation is most likely the best solution.
- Conversely, a classifier will likely be useless when its accuracy is < 70% [43].

- If the classifier and the humans tend to make the same type of mistakes, typically because they leverage the same information, then there is no reason for M_1 to be above 1.
- Finally, when the *internal decision model* (of the user) and *the computable decision model* (algorithm) leverage different sources of information, or prioritize the components of expected utility differently, we may find a measure above 1, demonstrating a collaboration effect.

4.2 Automation bias & resistance

Collaboration is easy to assess, explain, and it provides a nice metric to guide augmented Business Decision system design. It tells however little about how decision aids influence decision-making. We need to consider how a wrong recommendation influences the subject to let them lose their rationality, i.e. Automation bias, which can be one of two sub-categories:

- Authority bias: the subject follows (wrongly) the algorithm instead of their own reasoning because they perceive themselves as less accurate
- Complacency bias: the subject follows (wrongly) the algorithm out of a lack of motivation.

While we cannot distinguish the reason for the presence of *an automation bias*, we may still define it as *the probability that a subject who sees a wrong recommendation will make a non-rational decision, considering he would have made a rational decision had he not seen this recommendation*. We can define a dual measure of *resistance* as *the probability that the subject makes a rational decision when given a bad recommendation, knowing that if the recommendation had been correct, he would have made a rational choice*.

More rigorously, we can define a linear panel model [22]. Subjects are indexed by *i*, α_i corresponds to individual effects, while decision number are indexed by *t*. $Y'_{i,t}$ = $\mathbf{1}_{\{response\ of(I,t)\ is\ rational\}}$. We compare the control group with the treatment group that received wrong recommendations: $X_{i,t} = (1, \mathbf{1}_{\{false\ recommendation\}})^T$. The linear panel model $Y_{i,t}$ is defined as:

$$Y_{i,t} = \mathbf{1}_{\{X_{i,t}^T\beta+\alpha_i+\epsilon_{i,t}>0\}}$$

$$E[Y_{i,t}|X_{i,t}, \alpha_i] = X_{i,t}^T\beta + \alpha_i$$

$$\forall i \in \{1, ..., n\}, t \in \{1, ..., m\}$$

With this panel model, automation bias is defined as:

$$B(\alpha_i) = P(\{Y_{i,t} = 0|X_{i,t} = (1,1)^T, \alpha_i\}|\{Y_{i,t} = 1|X_{i,t} = (1,0)^T, \alpha_i\})$$

While resistance is defined as:

$$C(\alpha_i) = P(\{Y_{i,t} = 1|X_{i,t} = (1,1)^T\}|\{Y_{i,t} = 1|X_{i,t} = (1,0)^T\})$$

Estimating those probabilities can be performed with pooled ordinary least squares (pooledOLS) when performing a between-groups study (individuals are randomly assigned to the control and treatment groups and trials are randomly distributed). Other biases and effects, such as decision fatigue, order effects, timing effects, expertise effect, can be measured with the same econometric tool, provided our experiments record the appropriate information. To the reader unaccustomed to econometric tools: in a trivial panel, B is simply the extra error rate introduced by false recommendations, while C is the error rate when bad recommendations are given over the error rate when no recommendation are provided.

The goal of our research is to embed this evaluation model in various business decision systems, to automatically assess the proper decision aids to provide users, on a case by case basis.

5 Study

Before we may consider embedding those metrics in an augmented BPM system, we need to assess their effectiveness and discriminatory power. To this effect, we have conducted an online study based on a real use case but gamified so as to attract a large and varied population of subjects. Our study relies on a simple decision task: some information describing a case, a choice to be made among predefined possible outcomes, a right choice that depends on some rules and heuristics presumed known to the user (with an acceptable degree of variability). We test a single decision aid, presented as a recommendation from a generic classifier, under a variety of presentation modes. Sometimes the recommender will be misleading. We want to measure the success rate of the subjects, the impact of “wrong” recommendations as well as other measures such as decision fatigue or time taken to reach a decision. The deviation from rationality is attributed to an automation bias which denotes that the subject choses to follow the recommender over their own reasoning or intuition. Once we have obtained measures for a control condition (without decision aids), we provide decision aids with various presentations and observe the performance variations.

5.1 Choice of a decision task to evaluate

The task inspiring our use case is a fraud detection task: in large financial institutions, tens of thousands alerts on transactions are raised everyday as potential frauds. Scores of analysts review each of these alerts for further inspection, spending an average of 16 seconds on each alert. ML tools are currently being studied to help with this task, but actual deployment raises reliability and regulatory issues. Proper performance measures are needed to determine the optimal combination of algorithm and humans to handle this alert filtering. Resorting to a real-world task to define and assess our metrics is difficult: it requires extensive domain knowledge, and access to many expert users. Early pilots aimed at presenting a simplified version of the real task with a tutorial failed to provide the type of engagement we thought was needed: subjects

would fall for a type of "impostor's" syndrome, and blindly follow recommendations rather than give a chance to their own intuition and their own *internal decision model*. Others have followed this path [56] and failed to demonstrate nuanced results regarding the possibility of actual complementarity of human and algorithm decision processes.

Instead, we propose a simple decision task, for which a large population can create their own *internal decision model* easily, and will therefore have some autonomy in deciding whether to trust their judgment or the algorithmic recommendation. We leverage the well-known Titanic dataset[1], a database of passengers on the Titanic ship that sunk in 1912. This dataset is widely used to teach classification algorithms, because it exhibits some obvious patterns: most women in 1st and 2nd class survived, while most men in 3rd and 2nd class died, and other attributes such as number of relatives on board have a significant but lesser influence on the fate of each passenger. Simple machine learning classifiers, as well as humans after a short study of the data, exhibit success rates of 70-80% in guessing correctly if a passenger survived or not.

The decision task consists in, upon being presented with a passenger's information, choosing if it is more likely to have survived or died. The task is repeated 20 times. To create an incentive, the presentation is somewhat gamified: subjects are enticed to maximize their score of correct guesses. Unbeknown to the subjects, the passengers presented all follow the expected distribution of survivors: a logistics regression classifier has >70% chance of correct classification on those passengers. Hence, we are asking the subject to make a rational choice - maximizing their probability of scoring high -, not a chance guess. This means reaching a perfect score is possible and even likely. Obtaining less than 50% (less than random chance) is a sure sign the subject is not properly committed to the task.

5.2 Stimulus presentation

The experiment starts with a few demographic questions: age range, level of studies and type of studies (humanities, business, engineering/science or other). Then the subject is presented with the goal, as well as some interactive visualizations (treemap) that let them create their *internal decision model*. We do not present explicit decision rules so as not to taint subjects (Figure 1). Next, we introduce the task, indicating that the recommender (in the experimental condition) has about 76% success rate of guessing correctly. Then, we present the stimuli (Figure 2). In accordance to the stated success rate of the recommender, 5 times in the run of 20 trials, the recommendation is wrong: it says "survived" when the subject has died or vice-versa.

[1] https://www.kaggle.com/c/titanic

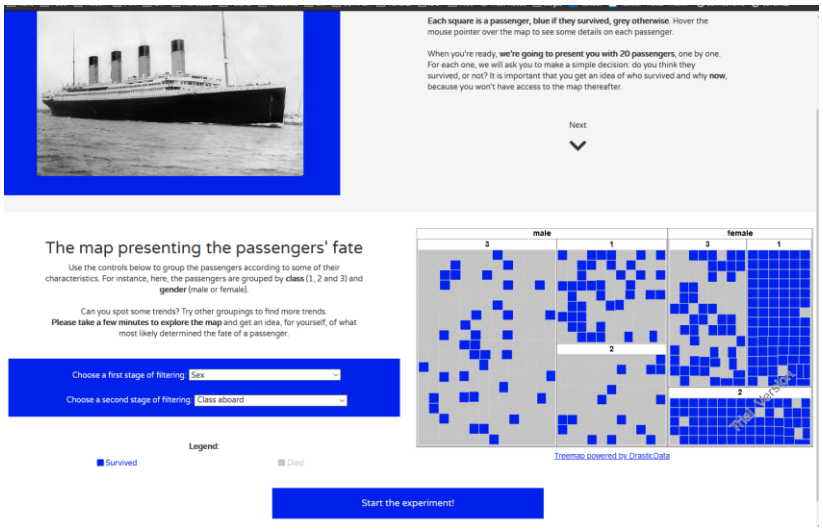

Fig. 1. interactive visualization (treemap), helping the user form an internal decision model of who predominantly survived on the Titanic: Females from 1st and 2nd class, and who tended to die: males of 2nd and 3rd class.

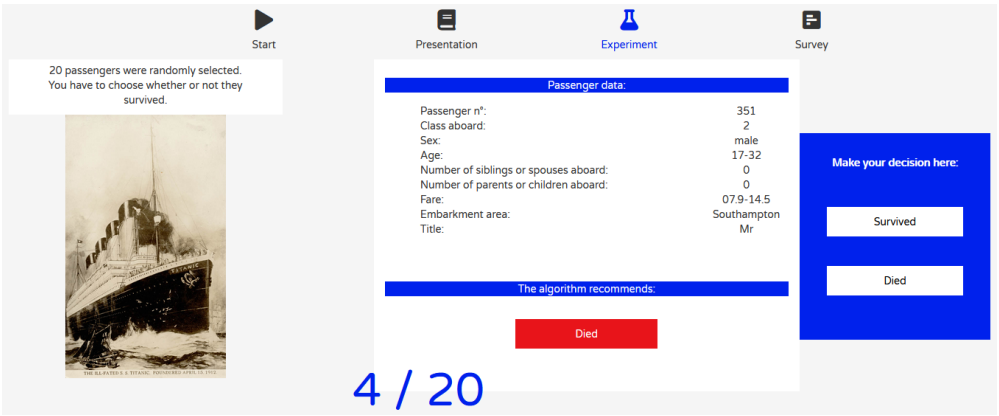

Fig. 2. Stimuli in the experimental condition: passenger data, followed by a recommendation (dies or survives), and, on the right, 2 buttons "survives" and "die". In the control condition, the recommendation panel is not shown.

Finally, after 20 trials, we ask a few experience questions: estimated success rate, did they choose intuitively or using self-made rules (or don't know), how many times they think the recommender provided the wrong answer, and a free comment box. Finally, we provide their score, and an invitation to an event that presented the early results. We set a cookie on their browser so they can't repeat the experiment for a few hours, so as not to taint our data collection with repeated trials by the same subject.

5.3 Second run: presentation variants

The first run was meant to compare a control condition (no decision aid) with an experimental condition. In a second run, conducted two weeks later on a different population, we tested 5 different presentations:

- Control: the same presentation as the first run
- Optional display: Instead of displaying the classifier result, the user has to click on a button "show recommendation" to display its result.
- Forced acknowledgement: the subject is presented with the passenger data, then must click a button to make the recommendation appear, and only then, after a small delay, they enter their choice.
- Reminder of 75%: instead of being only stated at the beginning, the subject is reminded that the decision aid has a 75% chances of success next to the recommendation.
- 80% success rate: instead of "guessing wrong" 75% of the time (5/20), the decision aid "fails" only 4 times out of 20. Of course, this is stated to the user, so it should not impact a rational behavior.

Each subject was assigned to the same condition for all their trials (between-groups design). The goal of this run was to assess how different presentation strategies and recommender reliability affect the measures of performance, authority and resistance (defined below) in a significant way. This would be a strong indication that our metrics can be generalized to other context to drive the design of augmented business decision support systems.

5.4 Subjects recruitment and filtering

After a pre-run to calibrate expectations with ~20 subjects, we recruited subjects online. We did not want to use a survey service such as Amazon Mechanical Turk as we felt the subject's engagement would be distorted by a financial reward. Instead, we presented the experiment as a "fun and useful challenge". We announced the challenge on a variety of venues, starting in the company and student's forums and slowly extending our call to a wider audience, such as focused reddit and facebook groups, over a 2 months span.

75% of the incoming participants completed their trials run, which we take as indicative of the motivation we had managed to induce from our anonymous subjects (Figure 3). A few participants (7) either failed to grasp the task or wanted to introduce noise and had less than 50% success (less than random chance), and we discarded them. In total, the first run had 231 participants and 155 usable trial runs, the second had 302 participants and 250 usable trial runs. The demographics reached is a mix of students and educated professionals, with more of an engineering/science background, roughly equally distributed in the 20-55 age range (Figure 3).

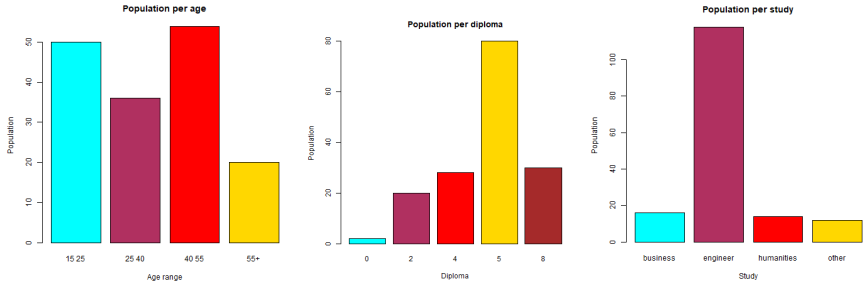

Fig. 3. Distribution of the subjects by age range, years of study after high school, and type of study (self-reported).

6 Results

6.1 Decision aid effectiveness

The first run shows a significant collaboration effect: subjects in the control condition obtain a score of 72.3%, while subject in the experimental condition (with decision aid) have a 76% success rate, giving $\mathbf{M_1 = 1.014}$.

	coefficient	95% confidence interval
Control condition (human alone)	0.7230	[0.6948, 0.7512]
With decision aid	0.7604	[0.7530, 0.7682]
"Algorithm alone"	0.75	--

Table 1: decision aid effectiveness (first run result)

This collaboration effect is most useful to compare decision aids presentation synthetically. Our second run shows some modest but consistent variations:

	coefficient	M_1
Control condition (human alone)	0.7230	1
With decision aid (new run)	0.7651	1.020
Optional display	0.7655	1.020
Forced acknowledgment	0.7660	1.021
Reminder of 75%	0.7619	1.016

Table 2: measures of collaboration for various presentation modes of the recommender

While this improvement may appear modest, it is statistically significant, and it should be reminded that our experiment is tuned to obtain average scores in the realistic range of 70%-80%. Finally, in the condition where the recommender has a 80% success rate, we have a coefficient of 0.7822 (p-value < 10^{-4}). Hence the collaboration

effect disappears (**M_1 = 0.977**), indicating that automatic classification would be better suited to the task. While we have not tested a recommenda1tion with 70% or less success rate, we can assume from [43] that we would find a similar negative impact on the collaboration.

These measures of M_1 allow deciding which type of presentation and decision aid to choose for a given task and a given effectiveness of the algorithm. However, the cost of wrong decisions, particularly if they involve an algorithmic bias, may not be symmetrical: false positives and false negatives may have different costs in different scenarios. Hence, the measure of biases is important to assess the full cost/benefit analysis of choosing appropriate decision aids.

6.2 Quantification of the automation bias

Coefficients β_1 and β_2 of the panel model are significantly nonzero at 5%. We can therefore reject the hypothesis that displaying a recommendation has no effect on the rationality of subjects. The model provides us with a metric that can be applied to various groups of the pool of subjects and compare their relative rationality ($B(\alpha_i)$, 0= little influence) and resistance ($C(\alpha_i)$, 1=maximal resistance). We can apply this model to study trends between different demographic classes recorded at the start of each run or presentation variants, between individuals, between trials (passengers) or any other available criteria. For instance, we display here the authority bias and resistance by type of studies (Table 4).

Study type	Authority bias $B(\alpha_i)$	Resistance $C(\alpha_i)$	95% conf. $B(\alpha_i)$,	95% conf. $C(\alpha_i)$
Engineering/science	0.0666	0.8581	[0.0626, 0.0705]	[0.8152, 0.9011]
Business	0.0708	0.8423	[0.0663, 0.0753]	[0.7898, 0.8947]
Humanities	0.0684	0.8471	[0.0641, 0.0726]	[0.7970, 0.8971]
other	0.0737	0.8423	[0.0687, 0.0787]	[0.7900, 0.8946]

Age range	$B(\alpha_i)$	$C(\alpha_i)$	Level of study	$B(\alpha_i)$	$C(\alpha_i)$
15-25	0.0661	0.8606	2-	0.0688	0.8553
25-40	0.0687	0.8473	4	0.0681	0.8552
40-55	0.0663	0.8522	5	0.0657	0.8537
55+	0.0713	0.8548	8	0.0694	0.8545

Table 4: authority bias and resistance for different demographic groups.

While the bias and resistance differences are small, those measures can be useful to apply to varying levels of expertise on a real task. Data and detailed results are available in the supplementary material.

6.3 Comparison between presentation variants

To choose the most effective presentation mode, depending if our goal is to maximize collaboration effectiveness, or to minimize authority bias while maintaining a high collaboration effectiveness, we compare the distributions of success under several conditions in Figure 4.

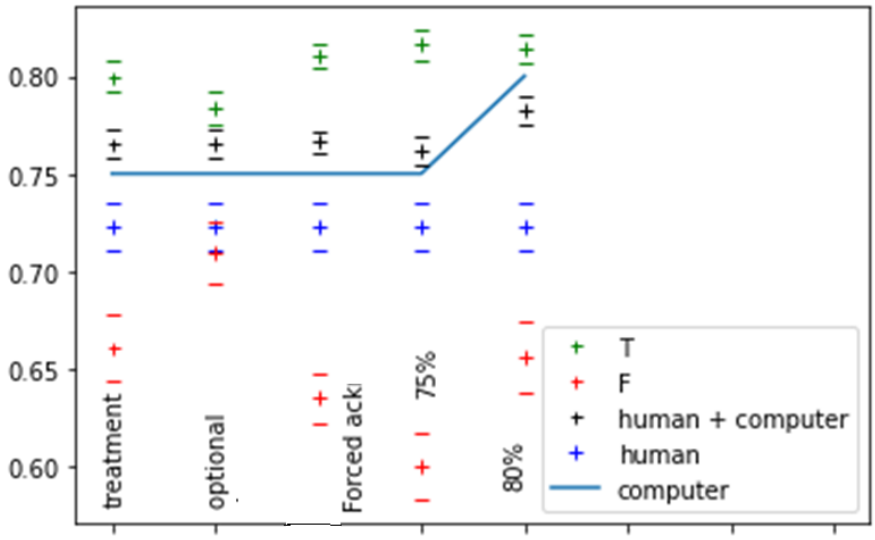

Figure 4: comparison of success rates with various stimuli presentations (black marks), success rates with wrong recommendation (red marks), success rate with good recommendations (green). Blue marks indicate the success without decision aid, and the line represent the "success rate" of the algorithm alone.

In Figure 4, we see success rate in the control condition as a horizontal line of blue marks, as a reference. Also for reference, the continuous horizontal line marks the "success rate" of the algorithm taken alone. The black marks indicate the success rate of the presentation mode. If the black marks are above both the blue marks (human alone) and the line ("computer alone"), then we can say the decision aid is effective in improving decision performance. This happens in all conditions but the last one, where the recommender has a much higher "success rate" of 80%. Based on this figure, the best performance is achieved with the "Forced acknowledgement" presentation mode (the subject must click to see the recommendation, then they can make their decision), although other presentations are quite close. This is the same information as tables 1 and 2.

But more importantly, the red marks show the distribution of success with a wrong recommendation. We can see that in the "optional display" presentation mode, the authority bias is weaker. This suggests using the presentation mode "optional display" over "Forced acknowledgement" if one is particularly concerned about avoiding underlying algorithmic biases.

6.4 Qualitative feedback

Time spent on this experiment is not relevant enough to justify elaborate statistics: we needed a task that could be completed fast to reach a large population and match our target use case of alarm filtering. Aggregated data shows that, by and large, our assumptions hold. The average time to answer a trial is 10.3s in the control condition vs. 9.5s in the 1st experiment condition, with medians at 6.3s and 5.4s respectively, suggesting a small performance improvement with the decision aid. In the experiment condition, there is a slightly longer average time (9.6s vs 9.1s) when the recommendation is wrong than when it is correct, suggesting that subjects perceive the need to reflect when presented with a counter-intuitive proposition. Performance time by demographic category does not vary much, and finally, the time spent on the experiment influences very little the success rate (Figure 5).

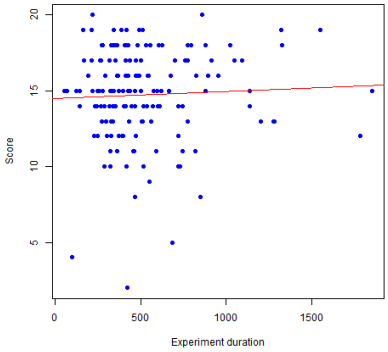

Figure 5: 2D plot time spent x success rate.

Subjects feedback. Subjects showed a variety of reactions. Some explained their reasoning: "I only looked at sex and class, taking more things into account was too confusing", “The pattern is pretty obvious. First class was a high chance of survival. Female is a high chance of survival. Children did better than adults. I didn't trust that the algorithm would do better than saying survive for 1st class and female.”.

Others detailed their frustration at various constraints we had voluntarily set for the task: "I wish the algorithm had provided me with some explanation about its recommendations. Typically, when I disagreed with the recommendation, I would have loved to ask "why do you recommend this?"", "If the algorithm is only 80% accurate, why show us the algorithm answer before we make our decision?", "I would have needed a few explanations on how the algorithm works before the experiment and while doing it some feedback on how it decided would have been helpful.", "I would have preferred to answer with probabilities rather than a binary choice.", “I would be interested to know the AI learning method”.

Finally, many showed an understanding and appreciation for the study, noting that wrong recommendations could indeed affect their judgment: "Interesting experiment. Would love to see the end study!", "Funny (not topic, itself) and interesting", "AI = Random?", "I am very curious to understand the analyses process and the results.

Would it be possible to receive the paper when published? Thanks". In debriefing interviews, subjects indicated that the test had made them aware of the complexity of the thought process at play when deciding to trust an algorithmic recommendation, which indeed, was the primary motivator of this study.

7 Discussion

7.1 Limitations

Our study results align with [43], [24], [54] and [1]. Taken together, they suggest that decision aids are useful only when the “algorithm alone” success is within a certain range, which we can roughly estimate as [70%..human success rate + constant]. Still, our measurements on this task may not be generalizable to other decision tasks. The causes of the uncertainty in a decision-making task may vary widely, the bias patterns may equally vary. Our contribution lies in the metrics of collaboration, automation bias and resistance, how to assess them and put them in production for our specific context, it is not a contribution to Decision Theory.

The task we have evaluated is not an expert task, it is more comparable to a routine managerial decision than to a complex decision such as medical diagnosis. Even though the metrics we have defined can be applied to those more complex contexts, experimental setup and access to many experts should make this very difficult, justifying more longitudinal approaches such as [12]. As mentioned in the introduction, our focus is in providing generic decision aids in the context of business decisions, and our task matches this context.

Finally, the effects we have observed, while statistically significant, may seem quite modest. We have shown a significant advantage of augmented decision-making in certain circumstances, and a significant difference in several presentation modes to contain, or, on the contrary, increase, automation bias and resistance. *The small amplitude of those effects is only a consequence of the narrow window in which augmented decision-making has its usefulness.* As our results suggest, when the algorithm clearly outperforms humans, it is probably better to rethink the role of the human. Conversely, when the algorithmic aid is of poor relevance, it is likely not useful, and, on the contrary, may lead to unwanted propagation of algorithmic biases through authority or complacency cognitive biases. This, by the way, has been confirmed a posteriori by the collaborative design sessions of the fraud detection workflow that inspired our study.

7.2 Towards a methodology and embedded measures of performance in augmented decision making

Now that our model has shown its explanatory and discriminatory power, our next goal is to apply the model we have defined to real usage scenarios, first in the use case described before: fraud alert detection. Another important direction is to apply

our methodology to other decision aids, such as nearest-neighbors' methods, that shows the data and outcomes of cases closely related to the present case.

In the longer run, we envision that decision aids will be generalized in business decision systems, provided they are instrumented with tooling that continuously assesses their relevance and performance so as to mitigate risks while improving the productivity and quality of decision-making.

8 Conclusion

We have presented a model, as a set of metrics to include in A/B testing of decision aids used in business decision tasks to assess their usefulness and control the biases a particular decision aid and its presentation may induce on the decision maker. Applied to a simple decision task, our metrics can be used to show the possibility of human-system collaboration (72% of success for humans alone, vs 76% for a human assisted by an algorithm that gives the correct answer 75% of the time). We have also defined measurements of automation biases and resistance to this bias. Applied to our decision task, we found a significant effect of wrong recommendations on the rationality of subjects (-5%), indicating that underlying algorithmic biases in the decision aid may be propagated in the decision process instead of compensated by the human.

Testing several presentation variants, we found that a technique that presenting the recommendation only on request (optional) was effective in increasing the resistance of the subjects, all the while preserving the performance of decision-making.

Our measurement system is meant to be embedded in A/B testing of generic decision support system used in many contexts of business decision management systems and business processes. *Augmented decision systems shift part of the responsibility of the decision maker to the system designer. Introducing systems that may induce authority, complacency or other cognitive biases creates liabilities for the system designers. Therefore, tools to objectivize the performance and impact of biases are needed to alleviate this liability, replacing impressions and subjective design guidelines with objective measures.* We believe furthering this work is important to shape the future of augmented business decision-making.

Acknowledgments

We thank Grégoire Colombet and François Jaquin for introducing us to their client's decision problems that led to this research work. We also thank Pranivan Baudouin and Christopher Dolloff for their precious assistance along this project.

References

1. Alberdi E., Strigini L., Povyakalo A.A., Ayton P. (2009) Why Are People's Decisions Sometimes Worse with Computer Support?. In: Buth B., Rabe G., Seyfarth T. (eds) Com-

puter Safety, Reliability, and Security. SAFECOMP 2009. Lecture Notes in Computer Science, vol 5775. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-04468-7_3

2. Veronika Alexander, Collin Blinder, Paul J. Zak, Why trust an algorithm? Performance, cognition, and neurophysiology, Computers in Human Behavior, Volume 89, 2018, Pages 279-288, ISSN 0747-5632, DOI: 10.1016/j.chb.2018.07.026.
3. Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. DOI: 10.1145/3290605.3300233
4. Anderson, Christopher (2003). "The Psychology of Doing Nothing: Forms of Decision Avoidance Result from Reason and Emotion". Psychological Bulletin. 129 (1): 139–167. DOI:10.1037/0033-2909.129.1.139. PMID 12555797. SSRN 895727
5. J. Elin Bahner, Anke-Dorothea Hüper, Dietrich Manzey, Misuse of automated decision aids: Complacency, automation bias and the impact of training experience, International Journal of Human-Computer Studies, Volume 66, Issue 9, 2008, Pages 688-699, ISSN 1071-5819, DOI: 10.1016/j.ijhcs.2008.06.001.
6. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. and Nagar, S., 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
7. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 377, 1–14. DOI: 10.1145/3173574.3173951
8. Silvia Bonaccio, Reeshad S. Dalal, Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences, Organizational Behavior and Human Decision Processes, Volume 101, Issue 2, 2006, Pages 127-151, ISSN 0749-5978, DOI: 10.1016/j.obhdp.2006.07.001.
9. Burton, JW, Stein, M-K, Jensen, TB. A systematic review of algorithm aversion in augmented decision making. J Behav Dec Making. 2020: 33: 220– 239. DOI: 10.1002/bdm.2155
10. Business Rules Journal, "A Brief History of the Business Rule Approach, 3rd ed." Business Rules Journal Vol. 9, No. 11, (Nov. 2008) : http://www.brcommunity.com/a2008/b448.html
11. Cabitza F. (2019) Biases Affecting Human Decision Making in AI-Supported Second Opinion Settings. In: Torra V., Narukawa Y., Pasi G., Viviani M. (eds) Modeling Decisions for Artificial Intelligence. MDAI 2019. Lecture Notes in Computer Science, vol 11676. Springer, Cham. DOI: 10.1007/978-3-030-26773-5_25
12. Carrie J. Cai et al. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*). Association for Computing Machinery, New York, NY, USA, Paper 4, 1–14. DOI: 10.1145/3290605.3300234
13. Castelo, N., Bos, M.W. and Lehmann, D.R., 2019. Task-dependent algorithm aversion. Journal of Marketing Research, 56(5), pp.809-825.
14. Chalmers (ed) The Extended Mind, Philosophy of mind: classical and contemporary readings, Oxford University Press, 2002.

15. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018;378(11):981-983. DOI: 10.1056/NEJMp1714229
16. Coglianese, Cary and Lehr, David, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017). Faculty Scholarship at Penn Law. 1734. https://scholarship.law.upenn.edu/faculty_scholarship/1734
17. Dijksterhuis, A., Bos, M.W., Nordgren, L.F. and Van Baaren, R.B., 2006. On making the right choice: The deliberation-without-attention effect. Science, 311(5763), pp.1005-1007.
18. Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, Steven Franconeri. Mitigating the Attraction Effect with Visualizations. IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 2019, TVCG 2019 (InfoVis 2018), 25 (1), pp.850 - 860. DOI: 10.1109/TVCG.2018.2865233. ⟨hal-01845004v2⟩
19. E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos and P. Dragicevic, "A Task-Based Taxonomy of Cognitive Biases for Information Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 2, pp. 1413-1432, 1 Feb. 2020, DOI: 10.1109/TVCG.2018.2872577.
20. E. Dimara, A. Bezerianos, and P. Dragicevic, "Conceptual and methodological issues in evaluating multidimensional visualizations for decision support," IEEE Transactions on Visualization and Computer Graphics, 2018
21. Endsley, M. R. (2017) 'From Here to Autonomy: Lessons Learned From Human–Automation Research', Human Factors, 59(1), pp. 5–27. DOI: 10.1177/0018720816681350.
22. Frees, E. (2004). Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. New York: Cambridge University Press.
23. Gerd Gigerenzer, Wolfgang Gaissmaier, Decision Making: Nonrational Theories, Editor(s): James D. Wright, International Encyclopedia of the Social & Behavioral Sciences (Second Edition), Elsevier, 2015, Pages 911-916, ISBN 9780080970875, DOI: 10.1016/B978-0-08-097086-8.26017-0.
24. Gombolay, M.C., Gutierrez, R.A., Clarke, S.G. et al. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. Auton Robot 39, 293–312 (2015). DOI: 10.1007/s10514-015-9457-9
25. Sebastian Hafenbrädl, Daniel Waeger, Julian N. Marewski, Gerd Gigerenzer,, Applied Decision Making With Fast-and-Frugal Heuristics, Journal of Applied Research in Memory and Cognition, Volume 5, Issue 2, 2016, Pages 215-231, ISSN 2211-3681, DOI: 10.1016/j.jarmac.2016.04.011.
26. Hastie, Reid and Dawes, Robyn, Rational Choice in an Uncertain World, The psychology of judgment and decision making, 2nd Edition, Sage Publications, 11/2009
27. Hirshleifer, D., Levi, Y., Lourie, B. and Teoh, S.H., 2019. Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics*, *133*(1), pp.83-98.
28. HLEG-AI. Ethics guidelines for trustworthy AI. European Commision report, April, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
29. Jeston, John; Nelis, Johan (21 January 2014). Business Process Management. Routledge. ISBN 9781136172984.
30. Kahneman, D. Klein, G. Conditions for Intuitive Expertise, A Failure to disagree. American Psychologist. 2009, Vol. 64, No. 6, 515–526 DOI: 10.1037/a0016755
31. Khenissi, Sami, "Modeling and counteracting exposure bias in recommender systems." (2019). Electronic Theses and Dissertations. Paper 3182. DOI: 10.18297/etd/3182

32. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z. et al. Explaining the user experience of recommender systems. User Model User-Adap Inter 22, 441–504 (2012). DOI: 10.1007/s11257-011-9118-4
33. Bart P. Knijnenburg, Niels J.M. Reijmer, and Martijn C. Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11). Association for Computing Machinery, New York, NY, USA, 141–148. DOI: 10.1145/2043932.2043960
34. Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*). Association for Computing Machinery, New York, NY, USA, Paper 411, 1–14. DOI: 10.1145/3290605.3300641
35. Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 1603–1612. DOI: 10.1145/2702123.2702548
36. Lacity, Mary and Willcox, Leslie. What knowledge workers stand to gain from automation, Harvard Business Review, June, 2015 https://hbr.org/2015/06/what-knowledge-workers-stand-to-gain-from-automation
37. Lemaire, Axelle, LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique. A digest available at : https://en.wikipedia.org/wiki/Loi_pour_une_R%C3%A9publique_num%C3%A9rique
38. Jennifer M. Logg, Julia A. Minson, Don A. Moore, Algorithm appreciation: People prefer algorithmic to human judgment, Organizational Behavior and Human Decision Processes, Volume 151, 2019, Pages 90-103, ISSN 0749-5978, DOI: 10.1016/j.obhdp.2018.12.005.
39. Maggi, F.M.; Di Francescomarino, F.; Dumas, M.; Ghidini, C. (2014). Predictive monitoring of business processes. Proceedings of the International Conference on Advanced Information Systems Engineering (CAiSE). Lecture Notes in Computer Science. 8484. pp. 457–472. arXiv:1312.4874. DOI:10.1007/978-3-319-07881-6_31. ISBN 978-3-319-07880-9.
40. Christoph March, 2019. "The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players," CESifo Working Paper Series 7926, CESifo.<https://ideas.repec.org/p/ces/ceswps/_7926.html>
41. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007
42. Institut Montaigne, Algorithms : mind the bias ! Report of the Institut Montaigne think-tank, March 2020. https://www.institutmontaigne.org/en/publications/algorithms-please-mind-bias
43. Linda Onnasch, Crossing the boundaries of automation—Function allocation and reliability,, International Journal of Human-Computer Studies, Volume 76, 2015, Pages 12-21, ISSN 1071-5819, DOI: 10.1016/j.ijhcs.2014.12.004.
44. Raja Parasuraman, Dietrich H. Manzey, Complacency and Bias in Human Use of Automation: An Attentional Integration HUMAN FACTORS, Vol. 52, No. 3, June 2010, pp. 381–410. DOI: 10.1177/0018720810376055.
45. Prahl, A, Van Swol, L. Understanding algorithm aversion: When is advice from automation discounted?. Journal of Forecasting. 2017; 36: 691– 702. DOI 10.1002/for.2464

46. Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Paper 103, 1–13. DOI: 10.1145/3173574.3173677
47. Romanov, Dmitry & Kazantsev, Nikolay & Edgeeva, Elina. (2019). The Presence of Order-Effect Bias in Moscow Administration. DOI: 10.1007/978-3-030-30429-4_26.
48. Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." Journal of Economic Perspectives, 19 (4): 25-42. DOI: 10.1257/089533005775196732
49. Mariarosaria Taddeo, Luciano Floridi, How AI can be a force for good. Science 24 Aug 2018: Vol. 361, Issue 6404, pp. 751-752. DOI: 10.1126/science.aat5991
50. Morris and Steven Mintz, Ethical Obligations and Decision-Making in Accounting: Text and Cases, Chapter 2: Cognitive Processes and Decision Making in Accounting. 4th Edition 2017, McGraw Hill, ISBN10: 1259543471
51. Emmanuel Tissandier, Thomas Baudel. AIDA : Automatiser la prise de décisions métier en gardant l'humain dans la boucle. *31e conférence francophone sur l'Interaction Homme-Machine (IHM 2019)*, Dec 2019, Grenoble, France. pp.2:1-6. ⟨hal-02407617⟩.
52. Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction, 12*(3), 109-124. https://doi.org/10.17705/1thci.00131.
53. VON HALLE, Barbara (2001). Business Rules Applied. Wiley. ISBN 0-471-41293-7.
54. Yeomans, M, Shah, A, Mullainathan, S, Kleinberg, J. Making sense of recommendations. J Behav Dec Making. 2019; 32: 403– 414. DOI: 10.1002/bdm.2118
55. Yetgin, E., Jensen, M., & Shaft, T. (2015). Complacency and Intentionality in IT Use and Continuance. AIS Transactions on Human-Computer Interaction, 7(1), 17-42.
56. Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 50 (November 2019), 24 pages. DOI:https://doi.org/10.1145/3359152