Ostfalia
Hochschule für angewandte
Wissenschaften

Wolfsburg

Fakultät Wirtschaft

# WWP

# Wolfsburg Working Papers

# No. 21-02

# The Tragedy of Algorithm Aversion

Ibrahim Filiz, Jan René Judek, Marco Lorenz and Markus Spiwoks, February 2021

# The Tragedy of Algorithm Aversion

Ibrahim Filiz, Jan René Judek, Marco Lorenz and Markus Spiwoks

**Keywords:** Algorithm aversion, technology adoption, framing, behavioral economics, experiments.

**JEL classification:** D81, D91, G41, O33

**Abstract:** Algorithms already carry out many tasks more reliably than human experts. Nevertheless, some subjects have an aversion towards algorithms. In some decision-making situations an error can have serious consequences, in others not. In the context of a framing experiment we examine the connection between the consequences of a decision-making situation and the frequency of algorithm aversion. This shows that the more serious the consequences of a decision are, the more frequently algorithm aversion occurs. Particularly in the case of very important decisions, algorithm aversion thus leads to a reduction of the probability of success. This can be described as the tragedy of algorithm aversion.

**Ibrahim Filiz**, Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 160 3344 078, E-Mail: ibrahim.filiz@ostfalia.de

**Jan René Judek,** Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 5361 892 225 420, E-Mail: ja.judek@ostfalia.de

**Marco Lorenz,** Georg August University Göttingen, Faculty of Economic Sciences, Platz der Göttinger Sieben 3, D-37073 Göttingen, Germany, Tel.: +49 1522 6672 503, E-Mail: marco.lorenz@stud.uni-goettingen.de

**Markus Spiwoks**, Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 5361 892 225 100, E-Mail: m.spiwoks@ostfalia.de

## 1. Introduction

Automated decision-making or decision aids, so-called algorithms, are becoming increasingly significant for many people's working and private lives. The progress of digitalization and the growing significance of artificial intelligence in particular mean that efficient algorithms have now already been available for decades (see, for example, Dawes, Faust & Meehl, 1989). These algorithms already carry out many tasks more reliably than human experts. However, only a few algorithms are completely free of errors. Some areas of application of algorithms have serious consequences in the case of a mistake – such as autonomous driving (cf. Shariff, Bonnefon & Rahwan, 2017), making medical diagnoses (cf. Majumdar & Ward, 2011), or support in criminal proceedings (cf. Simpson, 2016). On the other hand, algorithms are also used for tasks which do not have such severe consequences in the case of an error, such as dating service (cf. Brozovsky & Petříček, 2007), weather forecasts (cf. Sawaitul, Wagh & Chatur,  2012) and the recommendation of recipes (cf. Ueda, Takahata & Nakajima, 2011).

Some subjects have a negative attitude towards algorithms. This is usually referred to as algorithm aversion (for an overview of algorithm aversion see Burton, Stein & Jensen, 2020). Many decision-makers thus tend to delegate tasks to human experts or carry them out themselves. This is also frequently the case when it is clearly recognizable that using algorithms would lead to an increase in the quality of the results.

Previous publications on this topic have defined the term algorithm aversion in quite different ways (Table 1). These different understandings of the term are reflected in the arguments put forward as well as in the design of the experiments carried out. From the perspective of some researchers, it is only possible to speak of algorithm aversion when the algorithm recognizably provides the option with the highest quality result or probability of success (cf. Burton, Stein & Jensen, 2020; Köbis & Mossink, 2020; Castelo, Bos & Lehmann, 2019; Ku, 2020; Dietvorst, Simmons & Massey, 2015). However, other authors consider algorithm aversion to be present as soon as subjects exhibit a fundamental disapproval of an algorithm in spite of its possible superiority (cf. Efendić, Van de Calseyde & Evans, 2020; Niszczota & Kaszás, 2020; Horne et al., 2019; Logg, Minson & Moore, 2019; Rühr et al., 2019; Yeomans et al., 2019; Prahl & Van Swol, 2017).

Another important aspect of how the term algorithm aversion is understood is the question of whether and possibly also how the subjects hear about the superiority of the algorithm. Differing approaches were chosen in previous studies. Dietvorst, Simmons and Massey (2015) focus on the gathering of experience in dealing with an algorithm in order to be able to assess its probability of success in comparison to one's own performance. In a later study, Dietvorst, Simmons and Massey (2018) specify the average error of the algorithm. Alexander, Blinder and Zak (2018) provide exact details on the probability of success of the algorithm, or they refer to the rate at which other subjects used the algorithm in the past.

**Table 1:** Definitions of algorithm aversion in the literature

| **Authors** | **Definition of algorithm aversion** |
|---|---|
| Dietvorst, Simmons & Massey, 2015 | "Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion (…)*" |
| Prahl & Van Swol, 2017 | "The irrational discounting of automation advice has long been known and a source of the spirited “clinical versus actuarial” debate in clinical psychology research (Dawes, 1979; Meehl, 1954). Recently, this effect has been noted in forecasting research (Önkal et al., 2009) and has been called algorithm aversion (Dietvorst, Simmons, & Massey, 2015)." |
| Dietvorst, Simmons & Massey, 2018 | "Although evidence-based algorithms consistently outperform human forecasters, people often fail to use them after learning that they are imperfect, a phenomenon known as *algorithm aversion*." |
| Castelo, Bos & Lehmann, 2019 | "The rise of algorithms means that consumers are increasingly presented with a novel choice: should they rely more on humans or on algorithms? Research suggests that the default option in this choice is to rely on humans, even when doing so results in objectively worse outcomes." |
| Commerford, Dennis, Joe & Wang, 2019 | “(…) *algorithm aversion* – the tendency for individuals to discount computer-based advice more heavily than human advice, although the advice is identical otherwise.” |
| Horne, Nevo, O’Donovan, Cho & Adali, 2019 | “For example, Dietvorst et al. (Dietvorst, Simmons, and Massey 2015) studied when humans choose the human forecaster over a statistical algorithm. The authors found that aversion of the automated tool increased as humans saw the algorithm perform, even if that algorithm had been shown to perform significantly better than the human.<br>Dietvorst et al. explained that aversion occurs due to a quicker decrease in confidence in algorithmic forecasters over human forecasters when seeing the same mistake occur (Dietvorst, Simmons, and Massey 2015).” |
| Ku, 2019 | “(…) “algorithm aversion”, a term refers by Dietvorst et al. (Dietvorst et al. 2015) means that humans distrust algorithm even though algorithm consistently outperform humans.” |
| Leyer & Schneider, 2019 | “In the particular context of the delegation of decisions to AI-enabled systems, recent findings have revealed a general algorithmic aversion, an irrational discounting of such systems as suitable decision-makers despite objective evidence (Dietvorst, Simmons and Massey, 2018)” |
| Logg, Minson & Moore, 2019 | "(…) human distrust of algorithmic output, sometimes referred to as “algorithm aversion” (Dietvorst, Simmons, & Massey, 2015).[1] “; Footnote 1: "while this influential paper [of Dietvorst et al.] is about the effect that seeing an algorithm err has on people's likelihood of choosing it, it has been cited as being about how often people use algorithms in general." |
| Önkal, Gönül & De Baets, 2019 | “(…) people are averse to using advice from algorithms and are unforgiving toward any errors made by the algorithm (Dietvorst et al., 2015; Prahl & Van Swol, 2017).” |
| Rühr, Streich, Berger & Hess, 2019 | "Users have been shown to display an aversion to algorithmic decision systems [Dietvorst, Simmons, Massey, 2015] as well as to the perceived loss of control associated with excessive delegation of decision authority [Dietvorst, Simmons, Massey, 2018]." |
| Yeomans, Shah, Mullainathan & Kleinberg, 2019 | "(...) people would rather receive recommendations from a human than from a recommender system (...). This echoes decades of research showing that people are averse to relying on algorithms, in which the primary driver of aversion is algorithmic errors (for a review, see Dietvorst, Simmons, & Massey, 2015)." |
| Berger, Adam, Rühr & Benlian, 2020 | “Yet, previous research indicates that people often prefer human support to support by an IT system, even if the latter provides superior performance – a phenomenon called algorithm aversion.” (…) “These differences result in two varying understandings of what algorithm aversion is: unwillingness to rely on an algorithm that a user has experienced to err versus general resistance to algorithmic judgment.” |
| Burton, Stein & Jensen, 2020 | "(…) algorithm aversion—the reluctance of human forecasters to use superior but imperfect algorithms—(…)" |
| De-Arteaga, Fogliato & Chouldechova, 2020 | “*Algorithm aversion*–the tendency to ignore tool recommendations after seeing that they can be erroneous (…)” |
| Efendić, Van de Calseyde & Evans, 2020 | "Algorithms consistently perform well on various prediction tasks, but people often mistrust their advice."; “However, repeated observations show that people profoundly mistrust algorithm-generated advice, especially after seeing the algorithm fail (Bigman & Gray, 2018; Diab, Pui, Yankelevich, & Highhouse, 2011; Dietvorst, Simmons, & Massey, 2015; Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009).” |

| | |
|---|---|
| Erlei, Nekdem, Meub, Anand & Gadiraju, 2020 | "Recently, the concept of algorithm aversion has raised a lot of interest (see (Burton, Stein, and Jensen 2020) for a review). In their seminal paper, (Dietvorst, Simmons, and Massey 2015) illustrate that human actors learn differently from observing mistakes by an algorithm in comparison to mistakes by humans. In particular, even participants who directly observed an algorithm outperform a human were less likely to use the model after observing its imperfections." |
| Germann & Merkle, 2020 | "The tendency of humans to shy away from using algorithms even when algorithms observably outperform their human counterpart has been referred to as algorithm aversion." |
| Ireland, 2020 | "(…) some researchers find that, when compared to humans, people are averse to algorithms after recording equivalent errors." |
| Jussupow, Benbasat & Heinzl, 2020 | "(…) literature suggests that although algorithms are often superior in performance, users are reluctant to interact with algorithms instead of human agents – a phenomenon known as algorithm aversion" |
| Kawaguchi, 2020 | "The phenomenon in which people often obey inferior human decisions, even if they understand that algorithmic decisions outperform them, is widely observed. This is known as algorithm aversion (Dietvorst et al. 2015)." |
| Köbis & Mossink, 2020 | "When people are informed about algorithmic presence, extensive research reveals that people are generally averse towards algorithmic decision makers. This reluctance of "human decision makers to use superior but imperfect algorithms" (Burton, Stein, & Jensen, 2019; p.1) has been referred to as algorithm aversion (Dietvorst, Simmons, & Massey, 2015). In part driven by the belief that human errors are random, while algorithmic errors are systematic (Highhouse, 2008), people have shown resistance towards algorithms in various domains (see for a systematic literature review, Burton et al., 2019)." |
| Niszczota & Kaszás, 2020 | "When given the possibility to choose between advice provided by a human or an algorithm, people show a preference for the former and thus exhibit algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015, 2016; Longoni et al., 2019)." |
| Wang, Harper & Zhu, 2020 | "(…) people tend to trust humans more than algorithms even when the algorithm makes more accurate predictions." |

In addition, when dealing with algorithms, the way in which people receive feedback is of significance. Can subjects (by using their previous decisions) draw conclusions about the quality and/or success of an algorithm? Dietvorst, Simmons and Massey (2015) merely use feedback in order to facilitate experience in dealing with an algorithm. Prahl and Van Swol (2017) provide feedback after every individual decision, enabling an assessment of the success of the algorithm. Filiz et al. (2021) also follow this approach and use feedback after every single decision in order to examine the decrease in algorithm aversion over time.

Other aspects which emerge from the previous definitions of algorithm aversion in the literature are the reliability of the algorithm (perfect or imperfect), the observation of its reliability (the visible occurrence of errors), access to historical data on how the algorithmic forecast was drawn up; the setting (algorithm vs. expert; algorithm vs. amateur; algorithm vs. subject) as well as extent of the algorithm's intervention (does the algorithm act as an aid to decision-making or does it carry out tasks automatically?).

In our view, the superiority of the algorithm (higher probability of success) and the knowledge of this superiority are the decisive aspects. We only speak of algorithm aversion when subjects are clearly aware that not using the algorithm reduces the expected value of their utility and they do not deploy it nevertheless. A decision against the use of an algorithm which is known to be superior reduces the expected value of the subject's pecuniary utility and thus has to be viewed as a behavioral anomaly (cf. Frey, 1992; Kahneman & Tversky, 1979; Tversky & Kahneman, 1974).

In decision-making situations which lead to consequences which are not so serious in the case of an error, a behavioral anomaly of this kind does not have particularly significant effects. In the case of a dating service, the worst that can happen is meeting with an unsuitable candidate. In the case of an erroneous weather forecast, unless it is one for seafarers, the worst that can happen is that unsuitable clothing is worn, and if the subject is the recommendation of recipes, the worst-case scenario is a bland meal. However, particularly in the case of decisions which have serious consequences in the case of a mistake, diverging from the rational strategy would be highly risky. For example, a car crash or a wrong medical diagnosis can, in the worst case, result in someone's death. Being convicted in a criminal case can lead to many years of imprisonment. In these serious cases, it is important to be sensible and use an algorithm when it is superior in terms of its probability of success. Can algorithm aversion be overcome in serious situations in order to make a decision which maximizes utility and which, at best, can save a life?

Tversky & Kahneman (1981) show that decisions can be significantly influenced by the context of the decision-making situation. The story chosen to illustrate the problem influences the salience of the information, which can also lead to an irrational neglect of the underlying mathematical facts. This phenomenon is also referred to as the framing effect (for an overview see Cornelissen & Werner, 2014). Irrespective of the actual probability of success, subjects do allow themselves to be influenced. Algorithm aversion can be more or less pronounced in different decision-making contexts. It is possible that subjects who have to decide on the use of an algorithm also take the consequences of their decision into account. This study therefore uses a framing approach to examine whether subjects are prepared to desist from their algorithm aversion in decision-making situations which can have severe consequences. We thus consider whether there are significantly different frequencies of algorithm aversion depending on whether the decision-making situations can have serious consequences or not.

## 2. Experimental design and hypotheses

In order to answer the research question we carry out an economic experiment in which the subjects assume the perspective of a businessperson who offers a service to his/her customers. A decision has to be made on whether this service should be carried out by specialized algorithms or by human experts.

In this framing approach, three decision-making situations with potentially serious consequences (Treatment A) and three decision-making situations with significantly less serious effects are compared (Treatment B). In Treatment A it concerns the following services: (1) Driving services with the aid of autonomous vehicles (algorithm) or with the aid of drivers (2), The evaluation of MRI scans with the help of a specialized computer program (algorithm) or with the aid of doctors, and (3) The evaluation of files on criminal cases with the aid of a specialized computer program (algorithm) or with the help of legal specialists. In Treatment B it concerns the following services: (1) A dating site providing matchmaking with the aid of a specialized computer program (algorithm) or with the support of staff trained in psychology, (2) The selection of recipes for cooking subscription boxes with the aid of a specialized computer program or the help of staff trained as professional chefs, and (3) The drawing up of weather forecasts with the help of a specialized computer program (algorithm) or using experienced meteorologists (Table 2).

**Table 2:** Treatments and decision-making situations

| **Decision-making situation** | **Treatment** |
|---|---|
| Autonomous driving<br>Evaluation of MRI scans<br>The assessment of criminal case files | A (possibly serious consequences) |
| Dating service<br>Selection of recipes<br>Drawing up weather forecasts | B (no serious consequences) |

The decision-making situations are selected in such a way that the subjects should be familiar with them from public debates or from their own experience. In this way, it is easier for the subjects to immerse themselves in the respective context. Detailed descriptions of the decision-making situations can be viewed in Appendix 3.

The study has a between-subjects design. Each subject is only confronted with one of a total of six decision-making situations. All six decision-making situations have the same probability of success: the algorithm carries out the service with a probability of success of 70%. The human expert carries out the service with a probability of success of 60%. The payment structure is identical in both treatments. The participants receive a show-up fee of €2, and an additional payment of €4 is made if the service is carried out successfully. It is only the contextual framework of the six decision-making situations which varies.

First of all, the subjects are asked to assess the gravity of the decision-making situation on a scale from 0 (not serious) to 10 (very serious). This question has the function of a manipulation check - in this way it can be seen whether the subjects actually perceive the implications of the decision-making situations in Treatment A as more serious than those in Treatment B. In the case of autonomous vehicles and the evaluation of MRI scans, it could be a matter of life and death. In the evaluation of documents in the context of criminal cases, it could lead to serious limitations of personal freedom. The three decision-making situations in Treatment A can thus have significant consequences for third parties if they end unfavorably. The situation is different in the case of matchmaking, selecting recipes and drawing up weather forecasts. Even when these tasks cannot be accomplished in a satisfactory way sometimes, the consequences are usually not very severe. A date might turn out to be dull, or one is disappointed by the taste of a lunch, or you are out without a jacket in the rain. None of those things would be pleasant, but the implications in Treatment B are far less serious than those in Treatment A.

A *homo oeconomicus* (a person who acts rationally in economic terms) must – regardless of the context – prefer the algorithm to human experts, because it maximizes his or her financial utility. Every decision in favor of the human experts has to be considered algorithm aversion.

Algorithm aversion is a phenomenon which can occur in a wide range of decision-making situations (cf. Burton, Stein & Jensen, 2020). We thus presume that the phenomenon can also be observed in this study. Although the decision-making situations offer no rational grounds for choosing the human experts, some of the participants will do precisely this. Hypothesis 1 is: Not every subject will select the algorithm. Null hypothesis 1 is therefore: Every subject will select the algorithm.

Castelo, Bos & Lehmann (2019) show that framing is suited to influencing algorithm aversion. A dislike for algorithms appears to various degrees in different contexts. Nonetheless, in this study, the algorithm was not recognizably the most reliable alternative, and there is also no performance-related payment for the subjects. In Castelo, Bos & Lehmann (2019), algorithm aversion is therefore not modeled as a behavioral anomaly.

However, we expect that the frame will have an influence on algorithm aversion if the financial advantage of the algorithm is clearly recognizable. Hypothesis 2 is: The proportion of decisions made in favor of the algorithm will vary significantly between the two treatments. Null hypothesis 2 is therefore: The proportion of decisions made in favor of the algorithm will not vary significantly between the two treatments.

In the literature there are numerous indications that framing can significantly influence the decision-making behavior of subjects (cf. Tversky & Kahneman, 1981). If subjects acted rationally and maximized their utility, neither algorithm aversion nor the framing effect would arise. Nonetheless, real human subjects – as the research in behavioral economics frequently shows – by no means act like *homo oeconomicus.* Their behavior usually tends to correspond more to the model of bounded rationality put forward by Herbert A. Simon (1959). Human beings suffer from cognitive limitations – they fall back on rules of thumb and heuristics. But they do try to make meaningful decisions – as long as this does not involve too much effort. This kind of 'being sensible' – which is often praised as common sense – suggests that great efforts have to be made when decisions can have particularly severe consequences. The founding of a company is certainly given much more thought than choosing which television program to watch on a rainy Sunday afternoon. And much more care will usually be invested in the selection of a heart surgeon than in the choice of a pizza delivery service.

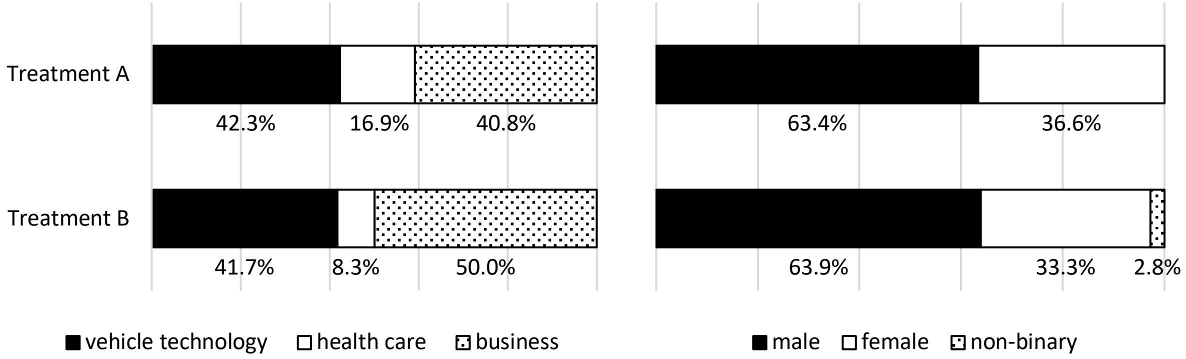
This everyday common sense, which demands different levels of effort for decision-making situations with different degrees of gravity, could contribute towards the behavioral anomaly of algorithm aversion appearing more seldom in Treatment A (decisions with possible serious consequences) than in Treatment B (decisions with relatively insignificant effects). Hypothesis 3 is thus: The greater the gravity of a decision, the more seldom the behavioral anomaly of algorithm aversion arises. Null hypothesis 3 is therefore: Even when the gravity of a decision-making situation increases, there is no reduction in algorithm aversion.

## 3. Results

This economic experiment is carried out between 2-14 November 2020 in the Ostfalia Laboratory of Experimental Economic Research (OLEW) of Ostfalia University of Applied Sciences in Wolfsburg. A total of 143 students of the Ostfalia University of Applied Sciences take part in the experiment. Of these, 91 subjects are male (63.6%), 50 subjects are female (35%) and 2 subjects (1.4%) describe themselves as non-binary. Of the 143 participants, 65 subjects (45.5%) study at the Faculty of Economics and Business, 60 subjects (42.0%) at the Faculty of Vehicle Technology, and 18 subjects (12.6%) at the Faculty of Health Care. Their average age is 23.5 years.

Of the participants, 71 subjects are in a decision-making situation which has been assigned to Treatment A, while 72 subjects are presented with a decision-making situation which has been assigned to Treatment B. The distribution of the subjects to the two treatments has similarities to their distribution among the faculties as well as their gender. In Treatment A (respectively Treatment B), 42.3% (41.7%) of the subjects belong to the faculty of vehicle technology, while 16.9% (8.3%) belong to the faculty of health care, and 40.8% (50.0%) belong to the faculty of business. In Treatment A (respectively Treatment B), 63.4% (63.9%) of the subjects are male, and 36.6% (33.3%) are female, and 0% (2.8%) are non-binary (Figure 1).

**Figure 1:** Proportions of the subjects belonging to a faculty and/or gender in Treatments A and B.



The experiment is programmed with z-Tree (cf. Fischbacher, 2007). Only the lottery used to determine the level of success when providing the service is carried out by taking a card from a pack of cards. In this way we want to counteract any possible suspicion that the random event could be manipulated. The subjects see the playing cards and can be sure that when they choose the algorithm there is a probability of 70% that they will be successful (the pack of cards consists of seven +€4 cards and three ±€0 cards). In addition, they can be sure that if they choose a human expert their probability of success is 60% (the pack of cards consists of six +€4 cards and four €±0 cards) (see Appendix 4).

The time needed for reading the instructions of the game (Appendix 1), answering the test questions (Appendix 2) and carrying out the task is 10 minutes on average. A show-up fee of €2 and the possibility of a performance-related payment of €4 seem appropriate for the time spent - it is intended to be sufficient incentive for meaningful economic decisions, and the subjects do actually give the impression of being concentrated and motivated.

The results of the manipulation check show that the subjects perceive the gravity of the decision-making situations significantly differently (Table 3 and Figure 2). In the decision-making situations with serious consequences (Treatment A), the average of the perceived gravity is 9.0 with a standard deviation of 1.37. The box extends from the 1st quartile $x_{0,25}$ = 8 (lower limit) to the 3rd quartile $x_{0,75}$ = 10 (upper limit). The median has a value of 10. In the evaluation of the gravity of the decision-making situations of Treatment B, there is an average of 6.54 with a standard deviation of 2.53. The box extends from the 1st quartile $x_{0,25}$ = 5 (lower limit) to the 3rd quartile $x_{0,75}$ = 8 (upper limit). The median is 7 and is thus far below the median in Treatment A.

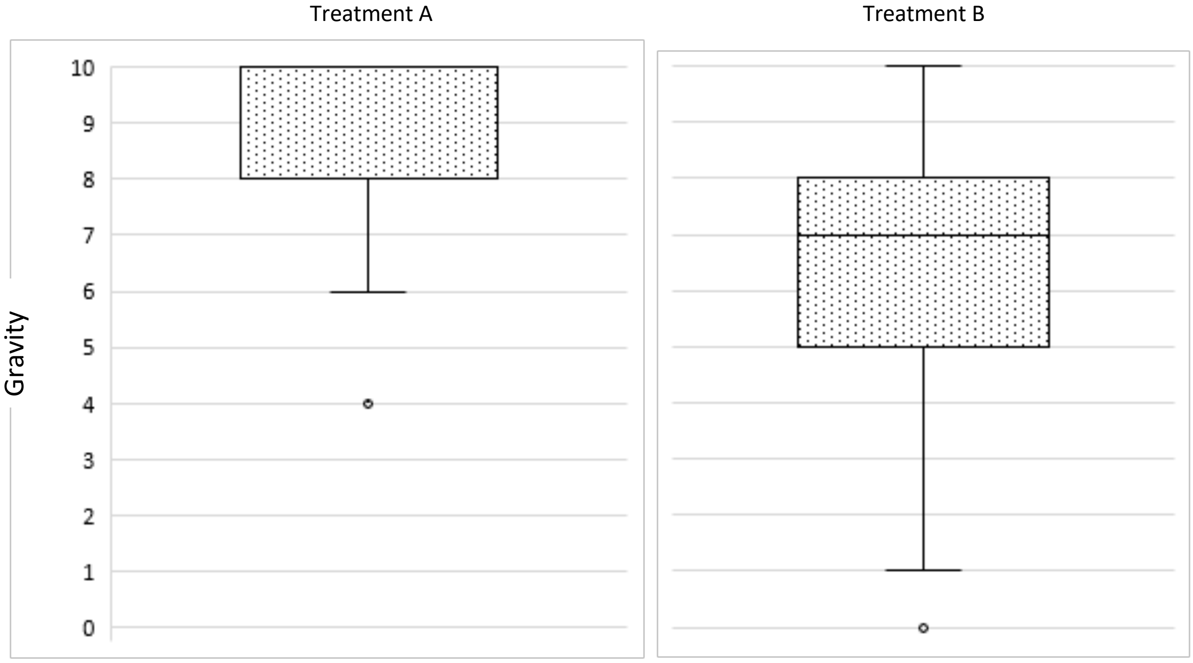
**Table 3:** Evaluation of gravity in Treatments A and B

| | Treatment A | Treatment B |
|---|---|---|
| First quartile | 8 | 5 |
| Third quartile | 10 | 8 |
| Median | 10 | 7 |
| Average | 9.00 | 6.54 |
| Standard deviation | 1.37 | 2.53 |

The graphical analysis also shows that overall, the subjects assess the gravity in Treatment A to be more serious than in Treatment B. In a direct comparison of the box plots, however, it is obvious that there is a larger range than in Treatment A, because some subjects also assess the gravity as very high in Treatment B (Figure 2).

The Wilcoxon rank-sum test (Mann-Whitney U test) (cf. Wilcoxon, 1945; Mann & Whitney, 1947) shows that the gravity of the decision-making situations in Treatment A is assessed as being significantly higher than that of the decision-making situations in Treatment B (z = 6.689; p = 0.000).

**Figure 2:** Box plot for the assessment of the gravity of the decision-making situations



Overall, only 87 out of 143 subjects (60.84%) decide to delegate the service to the (superior) algorithm. A total of 56 subjects (39.16%) prefer to rely on human experts in spite of the lower probability of success. Null hypothesis 1 thus has to be rejected. The result of the t-test is highly significant (p = 0.000). On average, around two out of five subjects thus tend towards algorithm aversion (Table 4). This is a surprisingly clear result, as the decision-making situations are very obvious. The fact that preferring human experts and rejecting the algorithm reduces the expected value of the performance-related payment should really be completely clear to all of the subjects. However, the need to decide against the algorithm is obviously strong in a part of the subjects.

Furthermore, a difference in the number of decisions in favor of the algorithm between the two treatments can be observed. Whereas in Treatment A 50.7% of the subjects trust the algorithm, this figure rises to 70.83% in Treatment B. Karl Pearson's $\chi^2$ test (cf. Pearson, 1900) reveals that null hypothesis 2 has to be rejected (p = 0.014). The frequency with which algorithm aversion occurs is influenced by the implications involved in the decision-making situation. The framing effect has an impact.

**Table 4:** Decisions for and against the algorithm

| | | | Decisions for the algorithm | | Decisions against the algorithm | |
|---|---|---|---|---|---|---|
| | | n | Number | Percent | Number | Percent |
| Treatment A | (serious) | 71 | 36 | 50.70% | 35 | 49.30% |
| Treatment B | (not serious) | 72 | 51 | 70.83% | 21 | 29.17% |
| Σ | | 143 | 87 | 60.84% | 56 | 39.16% |

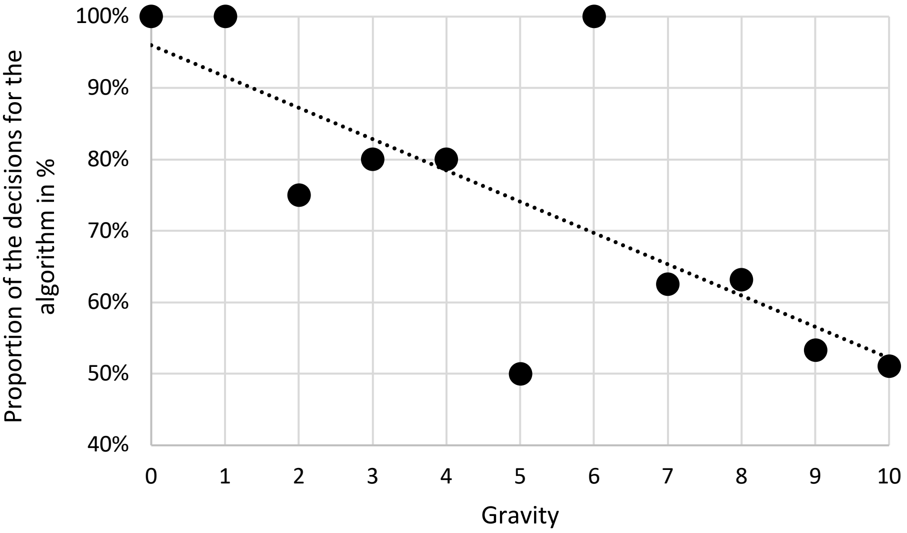
A framing effect sets in, but not in the way one might expect. Whereas in Treatment A (possibly serious consequences) 49.3% of the subjects do exhibit the behavioral anomaly of algorithm aversion, this is only the case in 29.17% of the subjects in Treatment B (no serious consequences) (Table 4). Null hypothesis 3 can therefore not be rejected.

There may be situations in which people like to act irrationally at times. However, common sense suggests that one should allow oneself such lapses in situations where serious consequences must not be feared. For example, there is a nice barbecue going on and the host opens a third barrel of beer although he suspects that this will lead to hangovers the next day among some of his guests. In the case of important decisions, however, one should be wide awake and try to distance oneself from reckless tendencies. For example, if the same man visits a friend in hospital whose life would be acutely threatened by drinking alcohol after undergoing a complicated stomach operation, he would be wise to avoid bringing him a bottle of his favorite whisky. This comparison of two examples illustrates what could be described as common sense, and would be approved of by most neutral observers.

Nevertheless, the results of the experiment point in the opposite direction. In the less serious decision-making situations (Treatment B) the tendency towards algorithm aversion is much less marked than in the serious situations (Treatment A).

This result is confirmed by a regression analysis which demonstrates the relationship between algorithm aversion and the perceived gravity of the decision-making situation. For the possible assessments of the consequences (from 0 = not serious to 10 = very serious), the respective average percentage of the decisions in favor of the algorithm is determined. The decisions of all 143 subjects are included in the regression analysis. Differentiation between the two treatments does not play a role here (Figure 3).

**Figure 3:** Decisions in favor of the algorithm depending on the gravity of the decision-making situation



If the common sense described above would have an effect, the percentage of decisions for the algorithm from left to right (in other words with increasing perceived gravity of the decision-making situation) would tend to rise. Instead, the opposite can be observed. Whereas in the case of only a low level of gravity (zero and one) 100% of decisions are still made in favor of the algorithm, the proportion of decisions for the algorithm decreases with increasing gravity. In the case of very serious implications (nine and ten), only somewhat more than half of the subjects decide to have the service carried out by an algorithm (Figure 3). If the perceived gravity of a decision increases by a unit, the probability of a decision in favor of the algorithm falls by 3.9% (t: -2.29; p = 0.023). Null hypothesis 3 can therefore not be rejected. In situations which have serious consequences in the case of an error, algorithm aversion is actually especially pronounced.

These results are very surprising, given that common sense would deem – particularly in the case of decisions which have serious consequences – that the option with the greatest probability of success should be chosen. If subjects allow themselves to be influenced by algorithm aversion to make decisions to their own detriment, they should only do so when they can take responsibility for the consequences with a clear conscience. In cases where the consequences are particularly severe, maximization of the success rate should take priority. But the exact opposite is the case. Algorithm aversion appears most frequently in cases where it can cause the most damage. To this extent it seems necessary to speak of the tragedy of algorithm aversion.

The decisive advantage of a framing approach is that the influence of a factor can be clearly identified. There is only one difference between the decision-making situations in Treatment A and Treatment B: the gravity of the possible consequences. It is needless to say that these are consequences which might have to be borne by third parties. It would be possible to continue this line of research by giving up the framing approach and modeling a situation where the subjects are directly affected. In this case, different incentives would have to be introduced into the two treatments. Success in Treatment A (possible serious consequences) would then have to be rewarded with a higher amount than in Treatment B (no serious consequences). However, we presume that our results would also be fully confirmed by an experiment based on this approach, given that it is a between-subjects design in which every subject is only presented with one of the six decision-making situations. Whether one receives €4 or €8 for a successful choice in Treatment A will probably not have a notable influence on the results. Nonetheless, the empirical examination of this assessment is something which will have to wait for future research efforts.

## 4. Summary

Many people decide against the use of an algorithm even when it is clear that the algorithm promises a higher probability of success than a human mind. This behavioral anomaly is referred to as algorithm aversion.

The subjects are placed in the position of a businessperson who has to choose whether to have a service carried out by an algorithm or by a human expert. If the service is carried out successfully, the subject receives a performance-related payment. The subjects are informed that using the respective algorithm leads to success in 70% of all cases, while the human expert is only successful in 60% of all cases. In view of the recognizably higher success rate, there is every reason to trust in the algorithm. Nevertheless, just under 40% of the subjects decide to use the human expert and not the algorithm. In this way they reduce the expected value of their performance-related payment and thus manifest the behavioral anomaly of algorithm aversion.

The most important objective of the study is to find out whether decision-making situations of varying gravity can lead to differing frequencies of the occurrence of algorithm aversion. To do this, we choose a framing approach. Six decision-making situations (three of which have potentially serious effects and another three which could have not very serious consequences) have an identical payment structure. The differing consequences of the decision-making situations do not affect the subjects themselves, but possibly have implications for third parties. Against this background there is no incentive or reason to act differently in each of the six decision-making situations. It is a between-subjects approach – this means that each subject is only presented with one of the six decision-making situations.

The results are clear. In the three decision-making situations with potentially serious consequences for third parties (Treatment A), just under 50% of the subjects exhibit algorithm aversion. In the three decision-making situations with not very serious consequences for third parties (Treatment B), however, less than 30% of the subjects exhibit algorithm aversion.

This is a really surprising result. If a framing effect were to occur, it would have been expected to be in the opposite direction. In cases with implications for freedom or even danger to life (Treatment A), one should tend to select the algorithm as the option with a better success rate. Instead, algorithm aversion shows itself particularly strongly here. If it is only a matter of arranging a date, creating a weather forecast or offering recipes (Treatment B), the possible consequences are quite clear. In a situation of this kind, one can still afford to have irrational reservations about an algorithm. Surprisingly, however, algorithm aversion occurs relatively infrequently in these situations.

One can call it the tragedy of algorithm aversion because it arises above all in situations in which it can cause particularly serious damage.

**List of references**

Alexander, V., Blinder, C. & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology, Computers in Human Behavior, 89(2018), 279-288.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn, *Business & Information Systems Engineering*, 1-14.

Brozovsky, L. & Petříček, V. (2007). Recommender System for Online Dating Service, ArXiv, abs/cs/0703042.

Burton, J., Stein, M. & Jensen, T. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making, *Journal of Behavioral Decision Making*, 33(2), 220-239.

Castelo, N., Bos, M. W. & Lehmann, D. R. (2019). Task-dependent algorithm aversion, *Journal of Marketing Research*, 56(5), 809-825.

Commerford, B. P., Dennis, S. A., Joe, J. R., & Wang, J. (2019). Complex estimates and auditor reliance on artificial intelligence, http://dx.doi.org/10.2139/ssrn.3422591.

Cornelissen, J. & Werner, M. D. (2014). Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature, *The Academy of Management Annals*, 8(1), 181-235.

Dawes, R., Faust, D. & Meehl, P. (1989). Clinical versus actuarial judgment, *Science*, 243(4899), 1668-1674.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 509, 1-12.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them, *Management Science*, 64(3), 1155-1170.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err, *Journal of Experimental Psychology: General*, 144(1), 114-126.

Efendić, E., Van de Calseyde, P. P. & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions, *Organizational Behavior and Human Decision Processes*, 157(C), 103-114.

Erlei, A., Nekdem, F., Meub, L., Anand, A. & Gadiraju, U. (2020). Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 43-52.

Filiz, I., Judek, J. R., Lorenz, M. & Spiwoks, M. (2021). Reducing Algorithm Aversion through Experience, *Wolfsburg Working Papers*, 21-01.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics*, 10(2), 171-178.

Frey, B. S. (1992). Behavioural Anomalies and Economics, in: *Economics As a Science of Human Behaviour*, 171-195.

Germann, M., & Merkle, C. (2019). Algorithm Aversion in Financial Investing, http://dx.doi.org/10.2139/ssrn.3364850.

Horne, B. D., Nevo, D., O'Donovan, J., Cho, J. & Adali, S. (2019). Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone?, *ArXiv*, abs/1904.01531.

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors, *Journal of Crime and Justice*, 43(2), 174-192.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion, *Proceedings of the 28th European Conference on Information Systems (ECIS)*, https://aisel.aisnet.org/ecis2020_rp/168.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica* 47(2), 263-291.

Kawaguchi, K. (2020). When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business, *Management Science, Articles in Advance*, https://doi.org/10.1287/mnsc.2020.3599.

Köbis, N. & Mossink, L. D. (2020). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry, *Computers in Human Behavior*, 114(2021), 1-13.

Ku, C. Y. (2020). When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives, *Információs Társadalom*, 19(4), 61-76.

Leyer, M., & Schneider, S. (2019). Me, You or Ai? How Do We Feel About Delegation, *Proceedings of the 27th European Conference on Information Systems (ECIS)*, 1-17.

Logg, J., Minson, J. & Moore, D. (2019). Algorithm appreciation: People prefer algorithmic to human judgment, *Organizational Behavior and Human Decision Processes*, 151 (C), 90-103.

Majumdar, A. & Ward, R. (2011). An algorithm for sparse MRI reconstruction by Schatten p-norm minimization, *Magnetic resonance imaging*, 29(3), 408-417.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics*, 18(1), 50-60.

Mill, J. S. (1836). On the definition and method of political economy, *The philosophy of economics*, 41-58.

Niszczota, P. & Kaszás, D. (2020). Robo-investment aversion, *PLoS ONE*, 15(9), 1-19.

Önkal, D., Gönül, M. S., & De Baets, S. (2019). Trusting forecasts, *Futures & Foresight Science*, 1(3-4), 1-10.

Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175.

Persky, J. (1995). The Ethology of Homo Economicus, *Journal of Economic Perspectives*, 9(2), 221-231.

Prahl, A. & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?, *Journal of Forecasting*, 36(6), 691-702.

Rühr, A., Streich, D., Berger, B. & Hess, T. (2019). A Classification of Decision Automation and Delegation in Digital Investment Systems, *Proceedings of the 52$^{nd}$ Hawaii International Conference on System Sciences*, 1435-1444.

Sawaitul, S. D., Wagh, K. & Chatur, P.N. (2012). Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach, *International Journal of Emerging Technology and Advanced Engineering*, 2(1), 110-113.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles, *Nature Human Behaviour*, 1(10), 694-696.

Simon, H. A. (1959). Theories of Decision-Making in Economics and Behavioral Science, *The American Economics Review*, 49(3), 253-283.

Simpson, B. (2016). Algorithms or advocacy: does the legal profession have a future in a digital world?. *Information & Communications Technology Law*, 25(1), 50-61.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science*, 211(4481), 453-458.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases, *Science*, 185(4157), 1124-1131.

Ueda, M., Takahata, M. & Nakajima, S. (2011). User's food preference extraction for personalized cooking recipe recommendation, *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation*, 781, 98-105.

Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 684, 1-14.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, 1(6), 80-83.

Yeomans, M., Shah, A. K., Mullainathan, S. & Kleinberg, J. (2019). Making Sense of Recommendations, *Journal of Behavioral Decision Making*, 32(4), 403-414.

**Appendix 1:** Instructions for the game

### The game

You are a businessperson and have to decide whether you want a service you are offering for the first time carried out solely by an algorithm or solely by human experts. You are aware that the human experts carry out the task with a probability of success of 60%. You are also aware that the algorithm carries out the task with a probability of success of 70%.

### Procedure

After reading the instructions and answering the test questions the decision-making situation is presented to you. This specifies the service which your company offers. First of all, you are asked to assess the gravity of the decision-making situation from the perspective of your customers. Then you decide whether the service should be carried out by human experts or by an algorithm.

### Payment

You receive a show-up fee of €2 for taking part in the experiment. Apart from this, an additional payment of €4 is made if the service is carried out successfully.

### Information

- Please remain quiet during the experiment
- Please do not look at your neighbor's screen
- Apart from a pen/pencil and a pocket calculator, **no** aids are permitted (smartphones, smart watches etc.)

**Appendix 2:** Test questions

**Test question 1:** Which alternatives are available to you to carry out the service?

a) I can provide the service myself or have it done by an algorithm.
b) I can provide the service myself or have it done by human experts.
c) I can have the service carried out via human experts or by an algorithm. (correct)

**Test question 2:** For how many newly-offered services do you need to make a choice?

a) None
b) One (correct)
c) Two

**Test question 3:** How much is the bonus payment for carrying out the task successfully?

a) €1
b) €2.50
c) €4 (correct)

**Test question 4:** How much is the bonus payment if you carry out the task wrongly?

a) -€2.50
b) €0 (correct)
c) €2.50

**Appendix 3**: Decision-making situations in Treatments A and B

**Treatment A:** Rather serious decision-making situations

**Decision-making situation A-1:** Autonomous driving

You are the manager of a public transport company and have to decide whether you want to transport your 100,000 passengers solely with autonomous vehicles (algorithm) or solely with vehicles with drivers (human experts). The task will be considered to have been successfully completed when all of your customers have reached their destination safely. In an extreme case, a wrong decision could mean the death of a passenger.

I choose: O Autonomous vehicles (algorithm)

O Drivers (human experts)

**Decision-making situation A-2:** MRI scan

You are the manager of a large hospital and have to decide whether the MRI scans of your 100,000 patients with brain conditions should be assessed solely by a specialized computer program (algorithm) or solely by doctors (human experts). The task will be considered to have been successfully completed when all life-threatening symptoms are recognized immediately. In an extreme case, a wrong decision could mean the death of a patient.

I choose: O Specialized computer program (algorithm)

O Doctors (human experts)

**Decision-making situation A-3:** Criminal cases

You are the head of a large law firm and have to decide whether the analysis of the case documents of your 100,000 clients should be carried out exclusively by a specialized computer program (algorithm) or solely by defense lawyers (human experts). The task will be considered to have been successfully completed when the penalties issued to your clients are below the national average. In an extreme case, a wrong decision could mean an unjustified long prison sentence for a client.

I choose: O Specialized computer program (algorithm)

O Defense lawyers (human experts)

**Treatment B:** Less serious decision-making situations

**Decision-making situation B-1:** Dating service

You are the manager of an online dating site and have to decide whether potential partners are suggested to your 100,000 customers solely by a specialized computer program (algorithm) or exclusively by trained staff (human experts). The task will be considered to have been successfully completed when you can improve the rating of your app in the App Store. For your customers, a wrong decision could lead to a date with a sub-optimal candidate.

I choose: O Specialized computer program (algorithm)
O Trained staff (human experts)

**Decision-making situation B-2:** Recipes

You are the manager of an online food retailer and have to decide whether your 100,000 cooking boxes – with ingredients and recipes which are individually tailored to the customers – are put together solely by a specialized computer program (algorithm) or solely by trained staff (human experts). The task will be considered to have been successfully completed when you can increase the reorder rate as a key indicator of customer satisfaction. A wrong decision could mean that the customers don’t like their meal.

I choose: O Specialized computer program (algorithm)
O Trained staff (human experts)

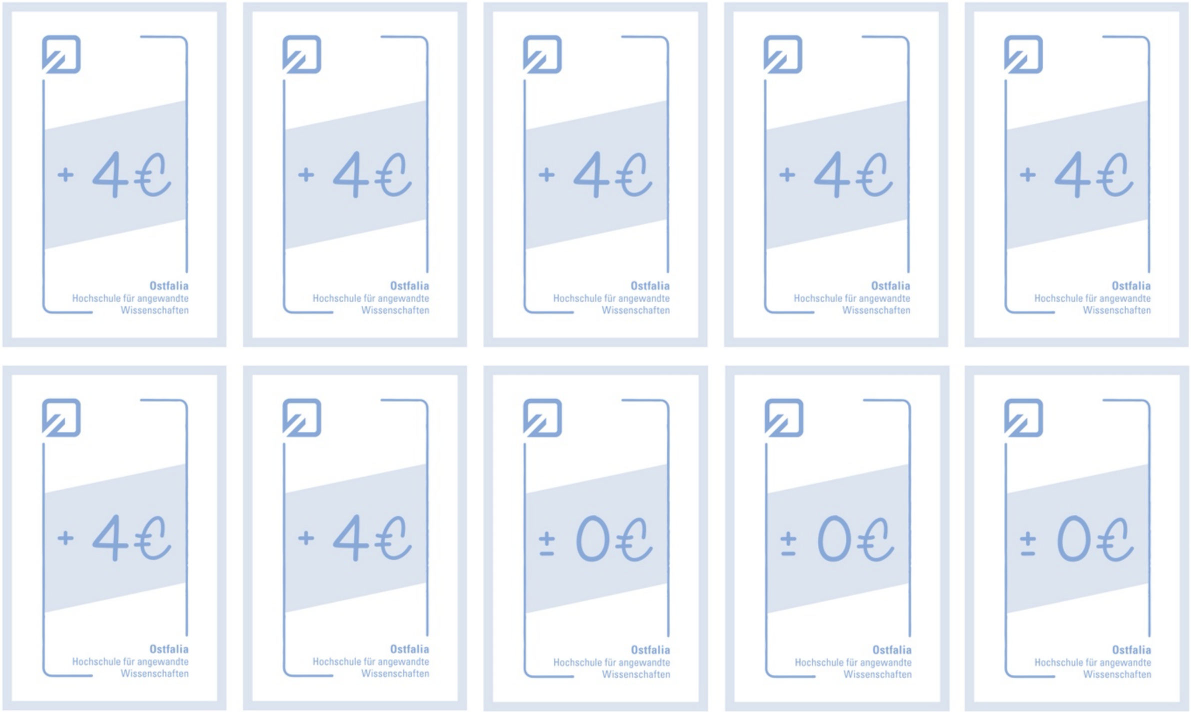
**Decision-making situation B-3:** Weather forecasts

You are the manager of a news site and have to decide whether your 100,000 daily weather forecasts for various cities are carried out solely by a specialized computer program (algorithm) or exclusively by experienced meteorologists (human experts). The task will be considered to have been successfully completed when the temperatures forecast the previous day do not diverge by more than 1 degree Celsius from the actual temperature. A wrong decision could mean that the readers of the forecasts do not dress suitably for the weather.

I choose: O Specialized computer program (algorithm)
O Experienced meteorologists (human experts)

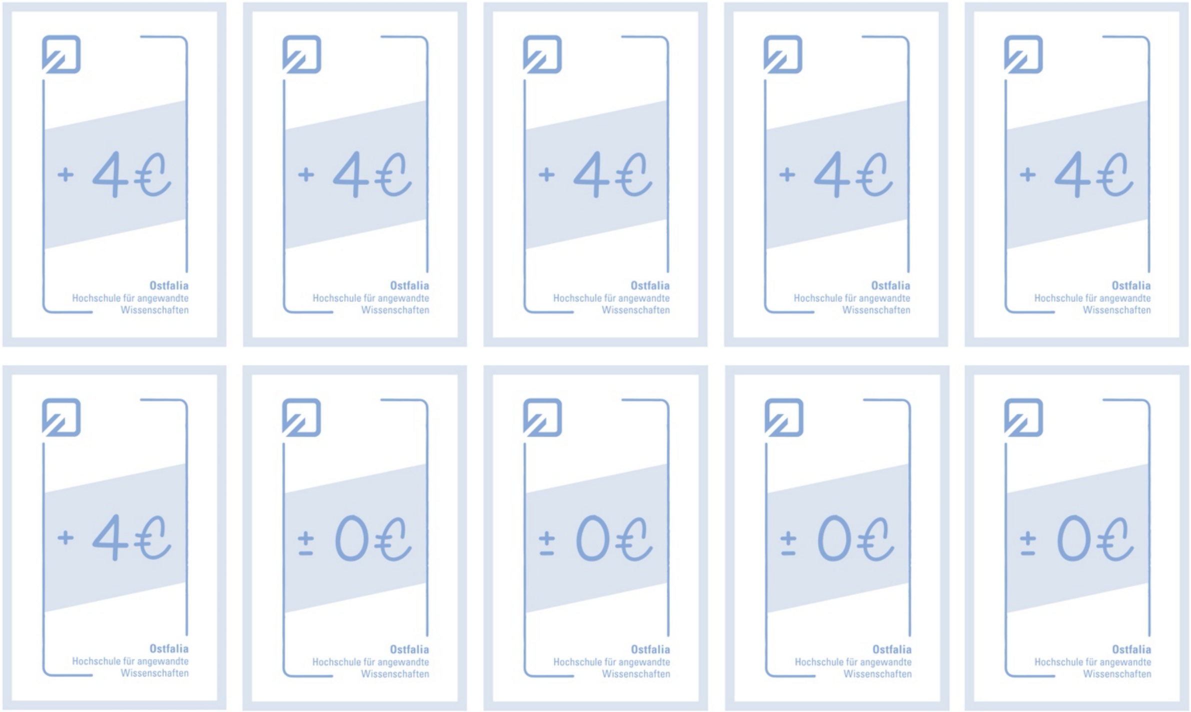
**Appendix 4:** Determination of the random event with the aid of a lottery

**Figure 3:** Pack of cards in the selection of the algorithm


+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften

Pack of cards in the selection of the algorithm: seven cards with the event +€4 and three cards with the event €±0.

**Figure 4:** Pack of cards in the selection of the human expert


+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
+ 4€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften
± 0€
Ostfalia Hochschule für angewandte Wissenschaften

Pack of cards in the selection of the human expert: six cards with the event +€4 and four cards with the event €±0.