Evaluation of the sensitivity of cognitive biases in the design of artificial intelligence

M Cazes, N Franiatte, A Delmas, J-M André, M Rodier, I Chraibi Kaadoud

Evaluation of the sensitivity of cognitive biases in the design of artificial intelligence.

M. Cazes[1 *], N. Franiatte[1*], A. Delmas[2], J-M. André[1], M. Rodier[3], I. Chraibi Kaadoud[4]

[1] ENSC-Bordeaux INP, IMS, UMR CNRS 5218, Bordeaux, France
[2] Onepoint - R&D Department, Bordeaux, France
[3] IBM - University chair "Sciences et Technologies Cognitiques" in ENSC, Bordeaux, France
[4] IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France

Abstract

The reduction of algorithmic biases is a major issue in the field of artificial intelligence. Despite the diversity of sources and temporalities at the origin of these biases, the current solutions to achieve ethical results are mostly technical. The consideration of human factors and more particularly cognitive biases remains incomplete. Nevertheless, the task of designing artificial intelligence systems is conducive to the emergence of cognitive biases. The aim of our study is to test the awareness of individuals who design artificial intelligence systems of the impact of their cognitive biases in their productions. The study focuses on conformity bias, confirmation bias and illusory correlation bias. The first results of this pre-experimentation show that these individuals believe that their decisions are subject to the cognitive biases of conformity, illusory correlation and confirmation.

Keywords

cognitive bias, artificial intelligence, decision making, human factors, human-IA interaction

Introduction

The question of biases, and their reduction, is a major issue in artificial intelligence (AI), especially since the raise of all the ethical issues related to deep learning algorithms application [39, 27, 42, 13]. These lead to more research on human-AI interaction field that studies human impact on AI systems [27, 6], and inversely, on the question of the impact of those systems on human [15, 36, 29].
Our work is in the field of human-AI interaction, not only machine learning, and questions the impact of human on AI systems during their design.
We define an AI system as a computational software, an algorithm or a scientific methodology using a set of algorithms, that assists humans to make a decision based on data analysis and data processing (mining, clustering, classification, prediction, recommandations, object recognition, etc.) whether in industrial or scientific context. We consider as an AI system any system that involves at least one of the following elements: an AI symbolic system (e.g. decision trees, expert systems) or a machine learning or deep learning algorithm (e.g. shallow and deep neural networks). From a software management point of view, the conception of an AI system goes through different steps that can be summarized as follows [3]: design, implementation of the code, test and production. We will refer in our work to all these steps as AI design. By doing so, we aim to group all the actors that can impact the product life cycle of an AI system. We will refer to these actors as AI professionals: AI researchers, AI manager products, data scientists, data analysts, data architects, data engineers and AI developers. Thus, the purpose of the current work is to propose a pre-experiment in the shape of a survey submitted to AI professionals in order to identify whether they are aware of being influenced in their work by cognitive biases classically known in human factors: conformity bias, confirmation bias, and illusory correlation bias. We therefore propose to conduct a study to test the following hypotheses (Figure 5):

- H0: Actors involved in the design of AI systems are aware that the biases of illusory correlation, conformity and confirmation influence their task.
- H1: Feedback collected during the qualitative interviews shows that AI systems design choices are influenced by these cognitive biases.

We adopt a different approach for works in literature by using human factors knowledge to address the issue of the presence of biases in AI systems by focusing first on the AI professionals' sensitivity of the cognitive biases. We will test H0 in this pre-experiment. This work is thus a first step in a larger question of the evaluation of cognitive biases on AI systems from a human factor perspective.
As the subject of biases in AI systems is much debated and studied within the AI community, we are concerned about clarifying the contribution and position of the current work: we do not aim to establish any causal effect of relation between cognitive biases and algorithmic biases.

*these authors contributed equally to this work

These concepts are clearly distincts. However there will be cited in our work since they both intervene in the larger research field of human-AI interaction. Table 1 provides definitions and non-exhaustive examples of the two distinct concepts according to the literature.

	Cognitive bias	Algorithmic bias
Definitions	"Distortion (systematic deviation from a norm) that information undergoes when entering or leaving the cognitive system. In the first case, the subject selects the information, in the second, he selects the answers" **[26]**	"Algorithmic bias can occur when the data used to train a machine learning algorithm reflects the implicit values of the humans involved in collecting, selecting, or using that data" **[38]**
	"Systematic pattern of deviations from the norm or rationality in judgment" **[16]**	"Problems related to the gathering or processing of data that might result in prejudiced decisions on the bases of demographic features such as race, sex, and so forth" **[39]**
Examples	Confirmation bias **[49]** Conformity bias **[20]** Illusory correlation bias **[8]**	Omitted variable bias **[5]** Data selection bias **[31]**

TABLE 1 – Differences between cognitive and algorithmic bias: extraction of definitions from scientific studies and non-exhaustive examples [26, 38, 16, 39, 49, 20, 8, 5, 31]

1 The biases of artificial intelligence

[42] showed that Amazon's facial recognition AI was very effective in performing facial recognition for light-skinned men (almost zero error rate) but had an error rate of 31.4% for dark-skinned women. This is one of the studies that highlights ethical issues related to discriminatory practices of AI. Their social consequences are concrete [39] as shown by the results of the NGO ProPublica's survey showing that the software used to predict crimes in the USA (The Correctional Offender Management Profiling for Alternative Sanctions, COMPAS) had a 'racist bias' [13]. The explanation of these discriminatory practices requires an understanding of the emergence of bias in the results of AI algorithms. We consider algorithmic bias, as defined in Table 1, including for example 'Problems related to the gathering or processing of data that might result in prejudiced decisions on the bases of demographic features such as race, sex, and so forth' [39] . The sources of algorithmic biases are of various natures and occur at different temporalities. Algorithmic biases may appear upstream (selection of learning data), during (inclusion of certain variables) and downstream (interpretation bias) of the learning phase [44]. Moreover, it is essential to qualify the nature of the source of these biases: technical bias (omitted variable bias, database bias, selection bias...) or 'society bias' (emotional or cognitive biases) [31]. However, despite the identification of the diversity of the sources of algorithmic biases, the correction of these biases is considered only from a statistical and algorithmic performance point of view i.e. in relation to certain metrics such as the percentage of prediction error or the accuracy in machine learning [39] (Figure 2). Indeed, [39] denounces the sole search for technical performance, which is neither relevant nor desirable to deal entirely with algorithmic biases. The influence of the human characteristics of designers on programmes is known [36], as recommendations are made regarding the need to incorporate diversity in the profiles of AI professionals [29, 34] (Figure 2, bottom right). However, there are no empirical studies investigating how human factors - and especially cognitive biases - can be concretely illustrated in the AI design process [40].

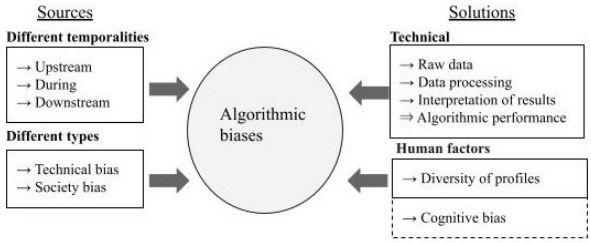

FIGURE 1 – Sources of algorithmic biases and solutions implemented [34] [31] [44]

We note that the debates around algorithmic biases bring into play problems of taking into account human factors in the actors of the design of artificial intelligence systems. We propose to carry out a study on this topic and more particularly the issue of the cognitive biases of the actors of this design chain.

2 Artificial intelligence design and cognitive biases

The conception of an AI system is a complex task. [32] develops the concept of complexity: a phenomenon is complex when 'the whole is more than the sum of its parts' i.e. there is an emergence of information. The conception of an AI system is a question of manipulating phenomena whose complexity is not mastered [1, 33]. In particular concerning the neural networks which make information emerge from which it is difficult to directly apprehend the chain of causality. In industry, designing AI systems is a task subject to constraints of time, means, objectives and expectations on the part of customers. These elements are part of a complex and uncertain context that is conducive to the use of heuristics on the part of designers and developers. [21] defined a heuristic as 'a simple procedure that makes it possible to find adequate, although often imperfect, answers to difficult questions'. These heuristics reduce the complexity of the task but can lead to the appearance of cognitive biases. [23] wrote as early as 1974 : 'In general, these heuristics are very useful, but sometimes they lead to severe and systematic errors'. The 'severe and systematic errors' considered in the quotation from Kahneman and Tversky [23] correspond to what will be referred to later as cognitive biases. It seems to us that they can be defined in this way as a systematic deviation of logical and rational thinking from reality.

Some authors have sought to model the functioning of cognitive biases by considering the theory of dual processes [48, 21] (Figure 2). The latter is based on the dichotomy between two main modes of reasoning: unconscious, rapid and automatic processes on the one hand, and slow and deliberate processes on the other. These two modes are colloquially referred to as 'intuitive' and 'analytical' but are

also known as System 1 (intuitive) and System 2 (analytical). According to [21], these two systems should be 'seen as agents with their own capabilities, limitations and functions'. Nevertheless, these 'agents' interact with each other and thus divide up certain tasks related to information processing. For example, while system 1 will handle automatic activities such as 'orienting towards the source of a sudden noise' or 'making a face of disgust at a horrible image', system 2 will handle tasks such as 'focusing on the voice of a particular person in a crowded and noisy room' or 'checking the validity of a complex logical argument'.

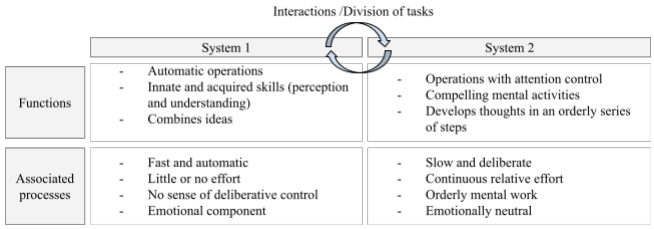

FIGURE 2 – Functions and processes associated with Systems 1 and 2, based on the work of Daniel Kahneman and Amos Tversky [23] [21]

Also, it appears that the notion of expertise is associated with the development of better performance of system 1. However, this 'intuitive expertise' is only valid when the task to be performed is part of a predictable environment in which the individual can learn from regularities [22]. In the case of developers of AI systems, the result of their actions on the neural network involves an element of uncertainty. In other words, their understanding of the impact of the modification of certain parameters is not complete. For example, [24] emphasised the fact that it seems unlikely that a single programmer would have a complete understanding of a large complex system designed by many teams around the world. Thus, despite the sense of expertise and mastery, the system of intuitive thinking cannot be fully effective.

In the field of computer science, the concrete bias impacts on developers has already been proved. Studies have shown that developers' cognitive confirmation bias has an influence on the quality of their programs. Indeed, [7] demonstrate that developers do unit testing - a software testing method to determine whether the written code is fit for use [19] - to show that their system works rather than testing it. More precisely, in the field of AI, research highlighted the concrete impacts of cognitive biases of the people involved in the design and implementation of AI programs. For example, the choice of the set of training variables for the creation of AI recruitment software has an important role in the quality and results of the program. However, this choice is made by managers and developers who are subject to their own cognitive biases such as generalization bias [45].

Finally, the task of designing an AI system has characteristics conducive to the emergence of cognitive biases. There are a multitude of cognitive biases, and it will be a matter of identifying the most relevant ones in our context.

3 Impossible and irrelevant identification of all cognitive biases

There are multiple typologies that classify cognitive biases according to pre-established criteria [18, 4, 12]. From the seminal work of Kahneman and Tversky [23] to more recent research [43], there does not seem to be a consensus on the comprehensiveness and classification of cognitive biases. Moreover, cognitive biases raise several epistemological questions, which are important to consider as they directly influence the applications of research results [28]. In order to qualify 'diversion' i.e. a bias, a rational behaviour must be established, i.e. a norm from which the behaviour deviates [47]. But how can behaviour be described as irrational ? Similarly, the nature of the task seems to influence the cognitive mechanism at work. [9] was thus able to demonstrate the existence of content effects for Wason's task: human beings use reasoning strategies appropriate to the nature of the problem they are facing. Cognitive biases would therefore not be 'mysterious irrationalities' but adaptations of the mind [16]. But how can cognitive processing be described as biased if it is context-dependent ? These elements show the complexity surrounding the notion of cognitive bias, which often makes it difficult to identify and evaluate them for a specific task. [5] already identified cognitive biases that could have impacts during the design and implementation of an AI system. In accordance with this work, we have chosen to focus on the study of three cognitive biases: conformity bias, confirmation bias, and the bias of illusory correlation.

4 Conformity bias, confirmation bias, illusory correlation bias

According to the codex of cognitive bias [30], the conformity bias and illusory correlation are due to the problem of the lack of sense of the individual's environment. In order to create meaning, individuals extrapolate 'attributes on the basis of stereotypes, generalities or antecedents' (translation of [30]). This led, among other reasons, to the development of the explainable AI and interpretable AI fields. Indeed, the neural networks used create a combinatorial system can be difficult to understand for the human cognitive system [1] and by extent to understand how AI systems lead to the given result. The difficulty of apprehending the internal functioning of neural networks can thus lead to a lack of understanding on the part of the AI professionals [1, 33] on the behavior or inner mecanisms of their AI systems. Consequently, it is possible to assume that this task takes place in an environment conducive to the expression of the biases mentioned above (Figure 3). Moreover, designing an AI program is mainly by selecting the data that will be used as a basis for learning the neural network. Data from various sources, types and formats must be chosen so that AI can produce quality information [14]. Therefore, it is ne-

cessary to select information in an environment that is saturated with it. Indeed, according to the codex typology of cognitive biases, too much information leads to cognitive biases [30]. Faced with the need to filter this information, individuals are attracted by what confirms their own convictions. This mechanism is at the origin of confirmation bias. The design of AI systems would therefore be an environment conducive to the emergence of confirmation bias (Figure 3).

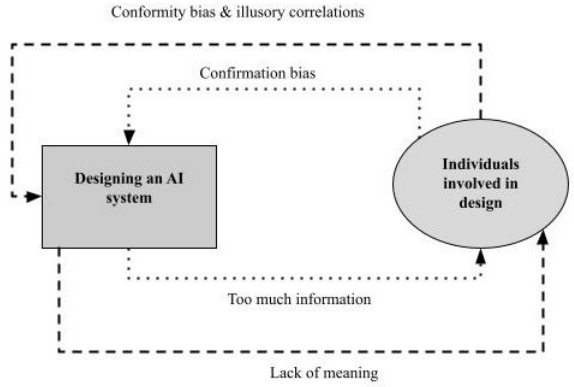

FIGURE 3 – Schematic representation of our hypothesis: link between elements of the context of the task and the cognitive biases they could encourage to appear

Finally, the characteristics of the AI systems design task seem to create an environment conducive to the emergence of the following cognitive biases: illusory correlation, conformity and confirmation bias. In order to understand more precisely how these biases operate, the following section will detail the cognitive mechanisms underlying their appearance.

4.1 Conformity Bias

Conformity bias can be defined as the 'modification of an individual's behavior or judgment to bring it into harmony with the behavior or judgment of the majority' (translation of [20]). As Asch's pioneering experience shows, it is a powerful mechanism in decision-making [2]. The consequences of conformity bias are ambivalent and contextual. It is, for example, one of the factors at the origin of collective intelligence phenomena [35] but also of 'collective conservatism' [46].

4.2 Confirmation bias

[49] was one of the first to point out the existence of a confirmation bias. [4] state that in forming a judgment, a large majority of individuals reason by trial and error on the basis of previous judgments, and tend to confirm the accuracy of their initial hypotheses. Thus, confirmation bias can be defined as the tendency to explain facts with consistent stories and neglect facts that contradict them [37, 11]. Confirmation bias can have different consequences depending on the context in which it is embodied. For example, in the run-up to the US presidential election, [25] studied the evolution of political books, highlighting a confirmation bias that shows that people who buy these books do so not for information, but for confirmation of their opinion.

4.3 The illusory correlation bias

The first descriptions and explanations of the phenomenon of illusory correlation come from social psychology, and more particularly from the studies of the researchers Chapman and Chapman [8]. When two events are correlated, individuals attribute cause and effect relationships to one of them. The meaning of this correlation may be erroneous, or even totally illusory [41].

The three biases of conformity, confirmation and illusory correlation are therefore the focus of our study. After having identified them, we will transpose them into the context that actors in AI systems development may encounter.

5 Experimentation

5.1 Purpose and hypothesis

We focus our study on the impact of cognitive biases in the AI systems design process. Indeed, the nature and conditions of the AI systems design task are, in our opinion, clues to suppose that the biases of illusory correlation, conformity and confirmation would be at the origin of weaknesses in AI systems creation. There is a lack of empirical studies on how cognitive biases act precisely on decision making in an AI system design context. We therefore propose to conduct a study to test the following hypotheses (Figure 4). As a reminder, our assumptions are as follows:

— H0: Actors involved in the design of AI systems are aware that the biases of illusory correlation, conformity and confirmation influence their task.
— H1: Feedback collected during the qualitative interviews shows that AI systems design choices are influenced by these cognitive biases.

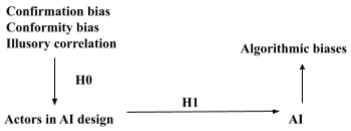

FIGURE 4 – Hypothesis on the impact of cognitive biases in AI systems design

5.2 Methodology for testing the H0 hypothesis

To test the H0 hypothesis, an online questionnaire was carried out. The purpose of this questionnaire is to perform a preliminary study to obtain the self-assessments of AI professionals of the impact of cognitive biases in their work, and to recruit target people to have more accurate feedback. The online tool FramaForms was used, which anonymizes the data and removes it permanently after six months. The

questionnaire was sent via the bull-i3 mailing list on January 4, 2021, in the BAIA newsletter on January 10, 2021 and in the RISC newsletter on January 26, 2021. We have blocked the recording of the answers to the questionnaire on February 17, 2021. Bull-i3 is a mailing list of the IRIT (Toulouse Institute of Computer Science Research) which gathers all the members (industrialists, researchers, professors, PhD students,...) of the Information, Intelligence and Interaction communities concerned by the issues of these research fields. BAIA is the newsletter of the Bordeaux Artificial Intelligence Alliance and RISC is the information relay on cognitive sciences. These three mailing lists are French. This is a methodological choice. We wanted to avoid bias related to the translation of the questionnaire or cultural bias. The questionnaire is addressed to what we call AI professionals as defined above. The exclusion criteria set concern people whose expertise is not related to AI e.g. the answer of a journalist was discarded because his profile did not correspond to an AI professional.
The questionnaire is structured in four main parts, each of which includes several multiple-choice questions: I) Sensitivity to cognitive bias ; II) Cognitive biases and your profession ; III) Your profile. The questions were originally in French (Figure 5).

* Mandatory response

Questions
I) Sensitivity to cognitive bias
Have you ever heard about the 3 cognitive biases mentioned in the email (conformity, confirmation, illusory correlation)? *
If yes, in what context?
II) Cognitive biases and your profession
Do you think that some decision making in your professional environment has already been influenced by one of the cognitive biases mentioned above? *
If yes, which one(s)?
III) Your profile
In which sector do you work? *
If private, how big is your business?
What is the field of activity of your company / laboratory / other? *
What is the title of your profession? *
How old are you?
How do you identify yourself?
What is your nationality? *

FIGURE 5 – Questions in the questionnaire sent

5.3 Results

5.3.1 Profile of sampled respondents

39 people responded to our questionnaire. Among them, 11 women (28.2%) and 28 men (71.8%). Regarding age (in years), 18 people surveyed (46.2% of the people surveyed) were in the 35-55 age category. Next, in order of importance, come the 18-25 age group (10 people or 25.6%), then the 25-35 age group (7 people or 17.9%) and finally the 55+ age group (4 people or 10.3%).

5.3.2 Professional Profile

36 out of 39 people surveyed (92.3%) are in either the digital or artificial intelligence sector, or both. Four people combine one of these sectors with a specialty or do not belong to either one (i.e. 'Administration, data mining project', 'Social Robotics', 'Geomatics' and one person looking for work). The respondents are equally divided between the private sector (19 people) and the public sector (19 people), with one person not belonging to either of these two sectors ('In search of employment'). Of the 19 people working in the private sector, 9 (47.4%) belong to a large group, 6 (31.2%) to a start-up, 3 (15.8%) to a PME (Petite ou Moyenne entreprise in french, Small or Medium Factory) and 1 (5.3%) to neither of these categories. Finally, as regards the professions practiced by the respondents: 35.9% of them qualify as researchers/professors, the remainder are divided between student/doctoral students (17.9%), data related professions (25.6%) and other professions (including for example an application architect, a research engineer and an innovation manager) (Figure 6).

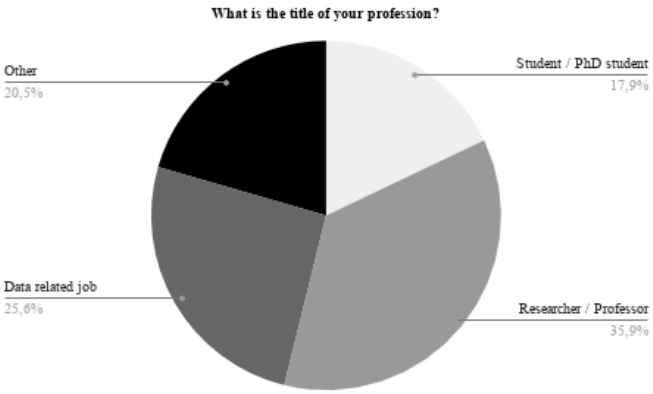

FIGURE 6 – Respondents' responses regarding the title of their professional designation.

5.3.3 Sensitivity to cognitive biases

In the questionnaire sent out, we briefly defined the biases of conformity, confirmation and illusory correlation. 66.7% of the people surveyed told us that they were aware of at least one of these three cognitive biases. Among the latter, this knowledge comes in 100% of cases from personal interest and/or professional training.

5.3.4 Potential impacts of cognitive biases on tasks in the work environment

Finally, all respondents believe that some decisions in their professional environments have already been influenced by at least one of the three cognitive biases. Among the answers obtained concerning the frequency of this influence : 56.4% answered 'Yes, regularly' and 43.6% 'Yes, rarely'.

According to the people surveyed, conformity bias seems to be the most frequent cognitive bias, followed by confirmation bias and then illusory correlation bias (Figure 7).

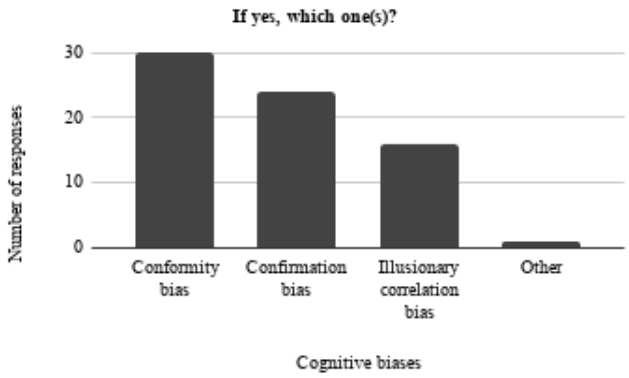

FIGURE 7 – Biases chosen by respondents who answered positively to the question 'Do you think that some decision making in your professional environment has already been influenced by one of the previously mentioned cognitive biases' (Question originally in French)

5.4 Discussion

First and foremost, the gender figures (28.2% women and 71.8% men) seem to reflect the current context, as women are still in the minority in the digital and artificial intelligence sectors. Indeed, according to INSEE, in 2017, the share of women in IT professions was 28%. Also, a study by the World Economic Forum and LinkedIn (2018) reports that only 22% of jobs in the field of artificial intelligence are held by women, and even less by senior managers.

This survey is a pre-experiment that establishes that all respondents believe that the cognitive biases studied in this article have (or have already had) an impact on their professional decision-making (39/39 or 100% of them). It confirms the need to take the human factors into account in the development of AI systems and to underline the relevance of the three cognitive biases mentioned in the article by Bertail et al [5] from which we drew inspiration. Also, these results allow us to note that the conformity bias is considered by respondents to be more influential in their professional tasks than the other two proposed biases (of confirmation and illusory correlation). It would therefore be possible to envisage that certain cognitive biases are more likely to influence decision making than others. However, this finding should be qualified in that the conformity bias may be better known to respondents than the other two biases. Also, another point to consider is the fact that it is generally easier for human beings to think that cognitive biases influence others before themselves (which would also bias our own analysis). This is an important point raised notably in the work of Fabrizio Butera [17]. We can summarize it here with the example of confirmation bias: 'The problem with confirmation bias is that it is itself subject to bias: people agree that confirmation bias exists in others, but find it difficult to admit that it exists in them as well. A meta-bias if you like.' This notion of 'meta-bias' should therefore be taken into account in our research and should lead us to interpret any results very cautiously.

Besides, during the study, we realized the semantic - even conceptual - ambiguity that can exist around the term 'bias', whether cognitive or algorithmic. Initially, the connotation of the latter seemed to us rather negative, often leading us to wonder about the potential means of 'debiaising' the human being. Nevertheless, in the course of our work, we found it interesting to also question their possible usefulness. Consequently, the term 'bias' semantically poses a confusion, in the sense that it always refers to a distortion, often interpreted as an error, whereas this distortion is also an essential means of survival for the human being. Moreover, it is difficult to find appropriate solutions to these biases (which would imply that it would be a problem). For Fabrizio Butera, cited by Hernandez [17]: 'The term bias is confusing. It refers to a tunnel of the mind. It implies that one cannot get out of it. If you are in a tunnel, you have to go from the entrance to the exit without making a detour. This implies that we are all helpless in the face of the slippery slope of our biases. Yet this is not what we observe socially.' Thus, in this article we use the terms 'cognitive bias' and 'algorithmic bias' while trying to define as well as possible what they represent for us, in order to avoid any confusion.

A second element that can be discussed is the potential social desirability bias, which results from a tendency of the individual to want to present himself favourably in the eyes of society [10]. Indeed, although our survey was anonymized, this psychological mechanism is sometimes implicitly exercised without the subject being aware of it. It would therefore be relevant to conduct another study, not related to AI systems design, in order to verify whether the awareness of cognitive biases by individuals in their professional environments shows a trend similar to the one found in this article.

Finally, the interest of this article is to question the taking into account the human factor - and more particularly cognitive biases - in the whole AI system design cycle. The questionnaire provided initial insight into the place of cognitive biases among AI designers and developers. The study of the potential impact of cognitive biases on the actors of AI system development, and their incarnations in the resulting work, represents a problem that could also be raised in future work. This is what we would like to do next, by testing hypothesis H1 through individual interviews with people who have completed our questionnaire and agreed to be contacted. The objective is to have more qualitative and targeted feedback on the cognitive biases that can occur according to the different expertises related to the AI domain.

6 Conclusion

The aim of this work is to raise the human factors and, more particularly, cognitive biases issues in the task of designing

AI systems. Indeed, the appearance of biases from different sources [31] and at all stages of designing AI systems [44] raises the question of the impact of cognitive biases during the design of an AI system. In other words, we wish to raise the question of the impact of cognitive bias in individuals at the origin of the realisation of intelligent systems. The AI system and cognitive bias relationship is an open scientific question which tends to grow in importance, particularly thanks to the growing concern about explainable AI systems intrinsically linked to the question of responsibility and ethics in AI [1].

Our objective is to test the awareness of those involved in the design of AI systems of the influence that their cognitive biases may have in their professional decision-making. Our work suggests that people involved in AI system design processes believe that their decisions are subject to the cognitive biases of conformity, illusory correlation and confirmation.

Future work should try to understand whether there is an embodiment of these cognitive biases in the AI system, and if so, how they are embodied. This involves testing our second hypothesis (H1, Figure 4). Indeed, it is not so much a question of the cognitive biases of the individuals involved in the design of AI systems, which are natural and useful [16], but of their incarnations and therefore their consequences in artificial intelligence. Thus, it would be relevant to carry out a study using feedback, in order to evaluate how these cognitive biases are embodied, or not, in the work of individuals responsible for designing AI systems.

Authors contributions

*MC and NF contributed equally to this work. All authors contributed to manuscript revision, read and approved the submitted version.

Références

[1] A. B. ARRIETA et al. “Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI.” In : *Information Fusion* 58 (2019), p. 82-115. DOI : 10.1016/j.inffus.2019.12.012.

[2] S. E. ASCH et H. GUETZKOW. “Effects of group pressure upon the modification and distortion of judgments. ” In : *Organizational influence processes* 58 (1951), p. 295-303.

[3] Marco BARENKAMP, Jonas REBSTADT et Oliver THOMAS. “Applications of AI in classical software engineering”. In : *AI Perspectives* 2.1 (2020), p. 1-15.

[4] D.A. BAZERMAN M.H.and Moore. *The Principles of Quantum Mechanics*. John Wiley Sons., 2008.

[5] Patrice BERTAIL et al. “Algorithmes : Biais , Discrimination et Équité”. In : *HAL* (2019). DOI : hal-02077745.

[6] Aylin CALISKAN, Joanna J BRYSON et Arvind NARAYANAN. “Semantics derived automatically from language corpora contain human-like biases”. In : *Science* 356.6334 (2017), p. 183-186.

[7] Gül ÇALIKLI et Ayşe Başar BENER. “Influence of confirmation biases of developers on software quality : an empirical study”. In : *Software Quality Journal* 21.2 (2013), p. 377-416.

[8] Loren J CHAPMAN et Jean P CHAPMAN. “Genesis of popular but erroneous psycho-diagnostic observations *”. In : *Journal of Abnormal Psychology* 72.3 (1967), p. 193-204.

[9] Leda COSMIDES. “The logic of social exchange : Has natural selection shaped how humans reason ? Studies with the Wason selection task.” In : *Cognition* 31.3 (1989), p. 187-276. DOI : 10.1016/0010-0277(89)90023-1.

[10] Douglas P. CROWNE et David MARLOWE. “A new scale of social desirability independent of psychopathology”. In : *Journal of Consulting Psychology* 24.4 (1960), p. 349-354. ISSN : 00958891. DOI : 10.1037/h0047358.

[11] Erica DAWSON et Dennis T REGAN. “Motivated Reasoning and Performance on the Wason Selection Task”. In : *Personality and Social Psychology Bulletin* 28.10 (2002), p. 1379-1387. DOI : 10.1177/014616702236869.

[12] R DOBELLI. *The Art of Thinking Clearly.* 2013.

[13] Julia DRESSEL et Hany FARID. “The accuracy, fairness, and limits of predicting recidivism”. In : *Science advances* 4.1 (2018), eaao5580.

[14] Usama FAYYAD, Gregory PIATETSKY-SHAPIRO et Padhraic SMYTH. “From data mining to knowledge discovery in databases”. In : *AI magazine* 17.3 (1996), p. 37-37.

[15] Diana F GORDON et Marie DESJARDINS. “Evaluation and selection of biases in machine learning”. In : *Machine learning* 20.1-2 (1995), p. 5-22.

[16] Martie G. HASELTON, Daniel NETTLE et Paul W. ANDREWS. “The Evolution of Cognitive Bias”. In : *The Handbook of Evolutionary Psychology* (2015), p. 724-746. DOI : 10.1002/9780470939376.ch25.

[17] J. HERNANDEZ. *Biais de confirmation : pouvez-vous le contrer ?* https://www.futura-sciences.com/sante/actualites/psychologie-biais-confirmation-pouvez-vous-contrer-85040/. 2021.

[18] R.M. HOGARTH. *Judgement and choice : The psychology of decision.* John Wiley Sons., 1980.

[19] Dorota HUIZINGA et Adam KOLAWA. *Automated defect prevention : best practices in software management*. John Wiley & Sons, 2007.

[20] A. IONESCU S.and Banchet. *Psychologie sociale : Nouveau cours de psychologie, Licence.* Presses Universitaires de France - PUF, 2009.

[21] D. KAHNEMAN. *Système 1 / Système 2 : Les deux vitesses de la pensée.* Flammarion, 2011.

[22] Daniel KAHNEMAN et Gary KLEIN. "Conditions for Intuitive Expertise : A Failure to Disagree". In : *American Psychologist* 64.6 (2009), p. 515-526. ISSN : 0003066X. DOI : 10.1037/a0016755.

[23] Daniel KAHNEMAN et Amos TVERSKY. "Judgment under Uncertainty : Heuristics and Biases". In : *Science* 185.4157 (1974), p. 1124-1131. DOI : 10.1126/science.185.4157.1124.

[24] Rob KITCHIN. "Thinking critically about and researching algorithms". In : *Information Communication and Society* 20.1 (2017), p. 14-29. ISSN : 14684462. DOI : 10.1080/1369118X.2016.1154087.

[25] V. KREBS. *Political Polarization.* http://www.orgnet.com/divided.html. 2008.

[26] Collectif LE NY. *Grand Dictionnaire de la Psychologie*. 1991.

[27] Susan LEAVY. "Gender bias in artificial intelligence : The need for diversity and gender theory in machine learning". In : *Proceedings of the 1st international workshop on gender equality in software engineering*. 2018, p. 14-16.

[28] Jean-fabrice LEBRATY. "Biais cognitifs : quel statut dans la prise de décision assistée ?" In : *Systèmes d'information et management* 9.3 (2004), p. 1-27.

[29] Bruno LEPRI et al. "Fair, transparent, and accountable algorithmic decision-making processes". In : *Philosophy & Technology* 31.4 (2018), p. 611-627.

[30] John MANOOGIAN III et B BENSON. "Codex des biais cognitifs". In : *Penser critique.* (2016).

[31] Institut MONTAIGNE. "Algorithmes : contrôle des biais svp". In : (2020).

[32] Edgar MORIN. "Le défi de la complexité". In : *Chimères. Revue des schizoanalyses* (1988). DOI : https://doi.org/10.3406/chime.1988.1060.

[33] Shane T MUELLER et al. "Principles of Explanation in Human-AI Systems". In : *arXiv preprint arXiv :2102.04972* (2021).

[34] Sarah MYERS-WEST, Meredith WHITTAKER et Kate CRAWFORD. *Discriminating systems*. April. 2019, p. 33. ISBN : 6153295159091.

[35] Joaquin NAVAJAS et al. "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds". In : *Nature Human Behaviour* 2.2 (2018), p. 126-132. ISSN : 23973374. DOI : 10.1038/s41562-017-0273-4. arXiv : 1703.00045.

[36] Gregory S NELSON. "Bias in artificial intelligence". In : *North Carolina medical journal* 80.4 (2019), p. 220-222.

[37] Raymond S NICKERSON. "Confirmation Bias : A Ubiquitous Phenomenon in Many Guises". In : *Review of General Psychology* 2.2 (1998), p. 175-220.

[38] Helen NISSENBAUM. "How computer systems embody values". In : *Computer* 34.3 (2001), p. 120-119.

[39] Eirini NTOUTSI et al. "Bias in data-driven artificial intelligence systems—An introductory survey". In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 10.3 (2020), e1356.

[40] Tore PEDERSEN, Christian JOHANSEN et Johanna JOHANSEN. "Studying the Transfer of Biases from Programmers to Programs". In : *arXiv* 2016 (2020). arXiv : 2005.08231.

[41] Arun RAI et Charles STUBBART. "Can executive information systems reinforce biases ?" In : *Banking Technology,* 4.2 (1994), p. 87-106.

[42] Inioluwa Deborah RAJI et Joy BUOLAMWINI. "Actionable Auditing : Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products". In : *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), p. 429-435. DOI : 10.1145/3306618.3314244.

[43] Olivier SIBONY. "Comprendre et prévenir l'erreur récurrente dans les processus de décision stratégique : l'apport de la Behavioral Strategy". Thèse de doct. PSL Research University, 2017.

[44] Selena SILVA et Martin KENNEY. "Viewpoint algorithms, platforms, and ethnic bias". In : *Communications of the ACM* 62.11 (2019), p. 37-39. ISSN : 15577317. DOI : 10.1145/3318157.

[45] Melika SOLEIMANI et al. "Cognitive biases in developing biased Artificial Intelligence recruitment system". In : *Proceedings of the 54th Hawaii International Conference on System Sciences*. 2021, p. 5091.

[46] C. THALER R.and Sunstein. *Nudge*. Yale University Press., 2008.

[47] Pascal WAGNER-EGGER. "Les canons de la rationalité : essai de classification des points de vue dans le débat sur les biais cognitifs et la rationalité humaine". In : *L'Année psychologique* 111.01 (2011), p. 191. ISSN : 0003-5033. DOI : 10.4074/s0003503311001072.

[48] P C WASON, J ST et T EVANS. "Dual processes in reasoning ?* P. C. WASON". In : *Cognition.* 3.2 (1975).

[49] P. C. WASON. "On the failure to eliminate hypotheses in a conceptual task". In : *Quarterly Journal of Experimental Psychology* 12.3 (1960), p. 129-140. ISSN : 0033-555X. DOI : 10.1080/17470216008416717.