Artikulu honetan, adimen artifizialean (IA) dauden giza joera kognitiboen ebidentzia berrikusiko dugu, eta joera horiek IAn eta giza elkarreraginetan nola eragiten duten azalduko dugu. Argudiatzen dugu psikologiak urte askotan zehar alborapenei buruz metatu duen ezagutza erabiltzeak hobetu egin dezakeela joera horiek IAri nola eragiten dioten eta haren eragina nola minimizatu dezakegun ulertzea.

Giltza-Hitzak: Joera Kognitiboa. Psikologia. Adimen Artifiziala. IA. Alborapena. Baieztapen-Alborapena. Kausalitate-Alborapena. Kausa-Ilusioa.

En este artículo, revisamos la evidencia de los sesgos cognitivos humanos presentes en la inteligencia artificial (IA) y discutimos ejemplos de cómo estos sesgos influyen en la IA y en las interacciones humano-IA. Argumentamos que usar el conocimiento que la psicología ha acumulado sobre los sesgos durante años puede mejorar nuestra comprensión de cómo estos sesgos afectan a la IA, así como cómo podemos minimizar su impacto.

Palabras Clave: Sesgo Cognitivo. Psicología. Inteligencia Artificial. IA. Sesgo. Sesgo de Confirmación. Sesgo de Causalidad. Ilusión Causal.

Dans cet article, nous passons en revue les preuves des biais cognitifs humains présents dans l'intelligence artificielle (IA) et discutons des exemples de la façon dont ces biais influencent l'IA et les interactions personne-IA. Nous soutenons que l'utilisation des connaissances que la psychologie a déjà accumulées sur les préjugés au fil de nombreuses années peut faire progresser notre compréhension de la façon dont ces préjugés affectent l'IA, ainsi que la façon dont nous pourrions minimiser leur impact.

Mots-Clés : Biais Cognitif. Psychologie. Intelligence Artificielle. IA. Biais. Biais de Confirmation. Biais de Causalité. Illusion Causale.

Human cognitive biases present in Artificial Intelligence

Martínez, Naroa (1); Agudo, Ujué (2); Matute, Helena (3)
(1)(2) (3) Deusto University. Departmen of Psychology. Avda. de las Universidades, 24– 48007 Bilbao
(2) Bikolabs/Biko. Pamplona
(3) Jakiunde, Zientzia, Arte eta Letren Akademia. Prim, 7. E-20006 - Donostia
(1) naroa.martinez@deusto.es; (2) ujue.agudo@biko2.com; (3) matute@deusto.es

Recep.: 2022-06-11
Acept.: 2022-11-30

1. Introduction[1]

Nowadays, we delegate many of our decisions to Artificial Intelligence (AI) algorithms. We use services that, through AI algorithms, influence our decision-making by recommending, for example, news (Carlson, 2017; Diakopoulos & Koliska, 2017; Thurman & Schifferes, 2012); job offers (de Pessemier et al., 2016); health advice (Bickmore et al., 2016; Grolleman et al., 2006; Hudlicka, 2013); and even friends and partners (Slater, 2013; van Dijck et al., 2018). Public institutions and companies are also increasingly relying on algorithms to make far-reaching decisions. For example, they are using AI algorithms to evaluate and predict crime (Dressel & Farid, 2018; Kennedy et al., 2011; Lin et al., 2020; Northpointe, 2019); to grant bank loans (Bartlett et al., 2022), to assist scientific discovery (Mayer-Schonberger & Cukier, 2013), and to guide medical decisions that affect millions of people (Obermeyer et al., 2019). The reliance on algorithms in certain sectors and tasks is so intense that, for example, according to some estimates, AI algorithms are now responsible for most of the activity on Wall Street (Isidore, 2018; Patterson, 2012). In fact, some authors have begun to use the term algocracy to refer to the transformation of government bureaucracy to a bureaucracy dominated by algorithmic decisions and rules (Danaher, 2016; Lorenz et al., 2021). The central mechanism in an algocracy is the proliferation of algorithm advice and decisions based on data analysis. Thus, algorithms are progressively determining our individual and collective choices.

The term algorithm has multiple meanings according to different perspectives. This term often refers to a computer's set of rules and instructions for acting on a set of data to solve a problem (Barocas & Selbst, 2016; Lee, 2018), though it could even have a more general meaning such as, for instance, the instructions needed to bake a cake, that is, what we usually call a cooking recipe. In recent years, however, the term algorithm is generally used to refer to AI algorithms. AI algorithms are a type of technology that has the ability to learn and also to identify what should be the most appropriate rules to solve a problem given some data and a desired outcome (Duan et al., 2019; Fry & Krzywinski, 2019). Moreover, AI algorithms have dramatically improved their performance by acquiring cognitive skills more or less developed, such as perception, reasoning, and decision-making (Eiband et al., 2019). Particularly noteworthy is the ability of the so-called machine learning algorithms to learn autonomously.

The term AI algorithm is currently being discussed as a complex term that goes beyond its purely technical aspects (Elish & Boyd, 2018; Kitchin, 2017). Several authors argue that algorithms are best understood within a broader socio-technical context, which would be inseparable from the conditions in which the algorithms are developed and deployed (Geiger, 2014; Napoli, 2013; Takhteyev, 2012). Thus, some researchers propose a conceptualization that evolves over time, as a function of the social and institutional contexts in which algorithms are developed and deployed (Araujo et al., 2020). From this perspective, it is argued that to treat algorithms simply as abstract and technical processes is to lose sight of their important social and political dimensions, intrinsically framed and shaped in their production and development (Barocas & Selbst, 2016; Kitchin, 2017; MacKenzie, 2007).

Indeed, the way we define the concept of algorithms influences our perceptions of their impact as well as the way we address it. Algorithms have been traditionally studied from a technical and mathematical perspective. A study that examined the definitions of algorithms given by people found that more than 80% of the participants focused on the categories of mathematics, equation, calculation, step-by-step procedure, logic, or formula (Logg et al., 2019). Because mathematical and statistical methods are often assumed to outperform human judgment (Dawes et al., 1989), people might prefer the judgment or recommendations made by algorithms to those made by humans in some contexts (Logg et al.,

[1] Support for this research was provided by Grant IT1696-22 from the Basque Government. The funders had no role in study design, decision to publish, or preparation of the manuscript.

2018) and consider algorithms more objective, neutral, rational, and free of bias than humans (Araujo et al., 2020; Dijkstra et al., 1998; Kahneman et al., 2021; O'Neil, 2016; Sundar, 2008). This technical perspective is consistent with the image of neutrality and objectivity attributed to algorithmic processes by technological companies.

However, far from this image of neutrality and objectivity, several studies have highlighted several potential problems associated with the use of algorithms, particularly the fact that they can be biased (Bolukbasi et al., 2016; Caliskan et al., 2017; Sweeney, 2013). Biases are systematic errors in judgments and decision-making that occur in the processing and interpretation of information. Many of these biases that are currently being detected in AI, are actually discriminative social biases, such as racism and sexism. Indeed, AI has been particularly criticized on augmenting and being affected by discriminative social biases, such as gender bias (e.g., French, 2018; Lambrecht & Tucker, 2016; Rodger & Pendharkar, 2004; Schwemmer et al., 2020) and racial bias (e.g., Angwin, 2017; Angwin et al., 2017; Buolamwini & Gebru, 2018; Coley et al., 2021; Obermeyer et al., 2019). However, since the existence of those biases in AI is already well-known and is already being criticized not only in academic forums but also in the media, we will not focus our discussion on those discriminative biases, but on more basic and general principles of cognition and cognitive biases that may be affecting our relationship with AI in more general ways.

Thus, the aim of this article is to review evidence of human cognitive biases present on AI. First, we explain some possible contributions from psychology to the study of AI. Then, we offer a definition of human cognitive biases and describe some consequences and problems associated to them. Finally, we provide some examples of human cognitive biases and their impact on AI. The main advantage of adopting this approach is that what it is known from the rigorous and long-term research tradition in psychology about cognitive biases can be usefully incorporated into the study of biases present in AI.

2. Possible Contributions from Psychology to the Study of AI

Given the ubiquitous use of AI algorithms in our societies, scientific research on the behavior of algorithms becomes crucial. However, the study of the behavior of AI algorithms comes today almost exclusively from AI developers and technological companies. Indeed, a Web of Science search (March 16, 2022) with the terms "algorithm behavior" yielded 51% of the references within the area of Computer Science, whereas studies from behavioral sciences, and in particular from psychology, were scarce (7 %). Given the current narrow approach to the study of AI behavior, some researchers have highlighted the urgent need to conduct studies from a multidisciplinary perspective (Martínez, Viñas, & Matute, 2021). The need for collaborative expert networks and multidisciplinary teams (Barocas & Selbst, 2016; Howard & Borenstein, 2017; Kremin et al., 2003), and the value of having diverse teams in the workplace are also increasingly recognized (OSTP-OPM, 2016). Similar to the study of the behavior of human intelligent agents (where we rely on a variety of disciplines about individual and collective behavior, such as biology, psychology, economics, and ethics, to mention just a few), the study of AI behavior requires complementary perspectives to properly advance knowledge, diagnose problems and devise potential solutions (Rahwan & Cebrian, 2018). In addition, if we would like to know how the use of biased AI algorithms can affect human behavior and decisions, we also need to conduct rigorous psychological experiments in which humans interact with AI agents (e.g., Agudo & Matute, 2021). Thus, some of these areas such as psychology and computer science are becoming increasingly integrated.

In addition, psychology has long made use of computer models to understand human cognition, sometimes under the term computational psychology (Anderson et al., 2008; Eysenck & Brysbaert, 2018; Sun, 2001). This approach relies on experimental findings regarding human cognition and the development of computer models that make predictions for new experiments. For instance, the process of learning can be studied through computer simulations, that is, using computer programs that attempt to simulate the learning of humans and animals (see e.g., Alonso, Mondragón, & Fernández, 2012; Musca et al., 2010; Rescorla & Wagner, 1972; Sutton & Barto, 1981; Vadillo et al., 2016). Computational modeling usually differs from AI in that computational modeling consists of building computer programs to simulate or model some aspects of human or animal cognitive functioning (including their biases, shortcuts, and limitations), whereas AI aims to produce results that may seem intelligence but generally involve processes which are different from those used by humans and other animals (Eysenck & Brysbaert, 2018). However, an emerging field in AI called cognitive technology or cognitive computing is becoming more similar to computational psychology. Cognitive technology is a new type of computing which has the ability to perform tasks that traditionally required human cognitive skills (e.g., perception and learning), focusing on emulating some of the key cognitive elements of humans and animals (Herrmann, 2004; IBM, 2020).

The field of AI is therefore using many different methods and is progressively improving its ability to process implicit, ambiguous and, complex information through an understanding of context, reasoning, and learning from experience. For example, some applications of cognitive technology and current AI research are recognition of handwriting, face identification, voice recognition technology, pattern recognition in data (e.g., image processing for various applications like biomedical imaging), computer vision and natural language processing (e.g., extracting meaning from complex, multi-format documents, and creating multi-lingual systems). Nowadays, cognitive technology is a leading and very promising AI field, and it allows the handling of large amounts of data and complex decision-making. This flourishing field of computer science based on human cognition has a huge challenge ahead related to how to minimize the development of biases in AI (e.g., Cappelli et al., 2019).

Therefore, an additional and highly significant contribution from psychology to AI could come from its knowledge on cognitive biases. Although the sources, mechanisms and potential strategies to reduce biases in AI and humans may differ, there is a constant interaction between AI and humans. Thus, cognitive biases are a mature area of research in cognitive psychology that can help address the current challenge that humanity faces in interacting with AI.

3. Human Cognitive Biases

Psychology has an extensive background in the study of cognitive biases and has made enormous progress in investigating the impact of biases on human judgments and decisions (e.g., Gilovich et al., 2002; Kahneman et al., 2021; Kahneman & Tversky, 1972; Krueger & Funder, 2004; Matute et al., 2015, 2019; Stanovich & West, 2000). There is also literature in psychology related to how to reduce and prevent cognitive biases (Arkes, 1991; Barberia et al., 2013; 2018; Larrick, 2008; Lilienfeld et al., 2009). Indeed, it has been argued that one of the most relevant contributions that psychology could make to humanity is the development of successful debiasing strategies (Lilienfeld et al., 2009). Therefore, the contribution from cognitive psychology could be also important to the study of how to reduce biases that impact AI through human intervention.

There is considerable research that has shown that human thought is prone to a biased interpretation of reality (Dawes, 2001; Dawes et al., 1989; Gilovich et al., 2002; Kahneman & Tversky, 1996; Nisbett

& Ross, 1980; Stanovich & West, 2000). For example, visual (i.e., not optical but visual) illusions are cognitive biases that make people erroneously interpret visual information. A very classic visual illusion occurs when seeing a line with the arrowheads outwards, <----> which seems to be longer than another line of the same size with the arrowheads inwards, >----< (Muller-Lyer, 1889). Indeed, the two lines are of identical length, but a visual illusion takes place when people estimate their length. Similarly, we all show cognitive illusions and biases in our judgments and decisions (Shepperd & Koch, 2005; Tversky & Kahneman, 1974). As mentioned above, cognitive biases are systematic and predictable errors that occur in all of us when we process and interpret information (Kahneman, 2003; Kahneman & Tversky, 1973; Stanovich & West, 2000; Tversky & Kahneman, 1974). The list of known cognitive biases is continually growing, though not all authors agree on all the examples. For example, Krueger & Funder (2004) made considerable progress in classifying biases, compiling a non-exhaustive list of 42 different biases, though not all of them are equally robust phenomena. A recent book edited by Pohl (2022) presents an excellent review of most common cognitive illusions and the evidence supporting each of them. Thus, it is not our aim to present an additional review of all possible cognitive biases, but to present some examples of how human biases interact and impact AI algorithms, and some of the problems that this interaction may cause on human wellbeing.

There is consensus in the literature that cognitive biases reflect adaptive processes which are usually called heuristics (Gigerenzer & Todd, 1999; Shepperd & Koch, 2005; Tversky & Kahneman, 1974). Heuristics are cognitive shortcuts that allow people to make quick and undemanding judgments and decisions under conditions of uncertainty, and they most of the time result effective, adaptive, and necessary (Gigerenzer & Goldstein, 1996). However, when the conditions change, people may keep using the same shortcut, and in that case, the result could be maladaptive, erroneous, and problematic. In those cases, they are called biases. That is, heuristics and biases are the two sides of the same phenomenon. Heuristics are adaptive responses to uncertain conditions, but they sometimes result in biases, that is, systematic errors that occur in most people under certain conditions.

Nonetheless, it is also noteworthy that cognitive biases can also lead sometimes to positive emotions such as a sense of control, and to the prevention of anxiety and depressed moods (see e.g., Alloy & Clements, 1992; Blanco, 2017; Langer, 1975; Matute, 1994, 1996; Taylor & Brown, 1988). However, most often cognitive biases lead to wrong or sub-optimal decisions and can produce undesirable outcomes. Cognitive biases are associated to multiple threats to human welfare, such as social stereotypes (Crocker, 1981; Hamilton & Gifford, 1976; Murphy et al., 2011), discrimination in hiring and promotion (Krieger & Fiske, 2006), ideological extremism (Lilienfeld et al., 2012; Tavris & Aronson, 2007), social intolerance and war (Johnson, 2004; Lilienfeld et al., 2009), hostile driving behaviour (Stephens & Ohtsuka, 2014), financial bubbles (e.g., Malmendier & Tate, 2005), paranormal (Blanco et al., 2015), superstitions (Griffiths et al., 2019), pseudoscientific beliefs (Lewandowsky et al., 2012; Matute et al., 2011; Torres et al., 2020), and public health problems such as medical diagnostic errors (Croskerry, 2013; Phua & Tan, 2013) and the use of alternative and complementary medicine (Blanco et al., 2014; Yarritu et al., 2015), among others. In addition, cognitive biases may contribute to psychopathological conditions such as social phobia (Bar-Haim et al., 2007), gambling (Orgaz et al., 2013), eating disorders (Brooks et al., 2011), depression (de Raedt & Koster, 2010; Peckham et al., 2010), and the development of psychotic experiences (Gawęda et al., 2015).

4. Some Examples of Human Cognitive Biases and their Impact on AI

Today, human cognitive biases have also a negative impact on AI algorithms. These biases can arise in different moments of the development of AI where humans are involved, from developers to users.

Humans intervene at various phases in the AI lifecycle, such as in data collection, AI model development, and AI model deployment (IBM, 2020; Xu & Doshi, 2019). The phase of data collection includes activities with the data that will be used to train the AI (identification, collection, and curation of data). The phase of AI model development can include the formulation, programming, training, and testing the AI model. The phase of AI model deployment refers to the interaction of the AI with the market, that is, the final users and contexts, and can include further learning and further development. In this phase AI models take the most advantage of context by interacting with the end user and getting feedback and new inputs so that they can keep learning and adapting to their new context (Garcia et al., 2018). The introduction of human cognitive biases in some or all of these phases of the AI lifecycle promotes the perpetuation and amplification of human biases in AI algorithms (Ashmore et al., 2019; Baeza-Yates, 2018; Garcia et al., 2018; O'Neil, 2016).

Below, we provide some examples of studies that help understand the impact of human cognitive biases in AI. We would like to highlight that the biases we discuss in this article (as any other biases) are present in all humans, not only in humans who develop or interact with AI. Given that the empirical literature in this area is scarce and disconnected across disciplines, we will try to connect the available research (which is usually conducted in computer science and has been almost exclusively studied from a technical perspective) with the psychological literature on cognitive biases.

4.1. Representativeness Bias

Representativeness bias occurs when assessing which example members are representative of a category based on their similarity to the prototype exemplar for that category, rather than on their actual likelihood of membership (Tversky & Kahneman, 1974). For example, this bias could explain why sparrows are more easily associated with the category of birds than penguins, although both belong to it (Rosh, 1975).

This human cognitive bias can impact AI at the time of training. For instance, it has been noted that the training stage for most AI algorithms used on face recognition applications makes use of pictures from people from developed countries as their training data, assuming that they are representative of the world population. This would be an example of a sampling error produced by the representativeness bias. Given that most AI agents are trained in developed countries, the selection of the sample responds more to the similarity of the exemplars to the prototype of people from developed countries (Barocas & Selbst, 2016). This bias is not a problem related only to AI developers, but is related, as any other biases, to all humans. For instance, the existence of this bias has also been noted in the literature of behavioral sciences, which relies mostly on experiments with WEIRD samples, an acronym for White, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010), even though these samples are often taken as representative of the general population.

This representativeness bias in the selection of data with which to train AI algorithms (also known in the field of AI as sampling bias) has led to certain minority groups suffering the effects of inaccurate algorithmic predictions. For example, representativeness bias in facial datasets has a strong effect on the accuracy of algorithms. One study showed that some datasets used to train commercial gender classification algorithms are composed of up to 86% white subjects and there is a 34% risk of misclassifying black women (Buolamwini & Gebru, 2018). Another example, (Devries et al., 2019) found that some current object recognition systems (such as Google Cloud Vision, Amazon Rekognition, or IBM Watson) are less effective at recognizing objects when photographs show household objects in non-Western countries and low-income communities, as these AI algorithms were trained on non-representative photographic datasets. In their study, the accuracy of the algorithms was approximately

15% higher when tested on photographs from the United States relative to photographs from Somalia or Burkina Faso.

4.2. Confirmation Bias

The confirmation bias refers to the tendency to seek, remember and confirm information consistent with one's own hypothesis and point of view, ignoring or reinterpreting information that contradicts them (Garb, 1998; Nickerson, 1998; Tavris & Aronson, 2007; Wason, 1966). This bias can lead us to draw distorted conclusions and often predisposes us to persevere with a belief even after receiving strong evidence that contradicts it (Ross et al., 1975). An example of confirmation bias is that of a supporter of a political party who only reads newspapers that support that political ideology.

A confirmation bias can occur when the AI model is being developed, particularly when humans test the model. Some empirical studies have reported confirmation bias in software testers, which is particularly evident in their tendency to design tests that confirm rather than question the correct functioning of the algorithm (Çalikli & Bener 2013; Salman, 2016; Salman et al., 2019; Teasley et al., 1994). Indeed, the available evidence on confirmation bias in software testing reveals that it had an impact on the quality of the software, causing a high number of software deficits (Calikli & Bener, 2013). In their study, Calikli & Bener (2013) conducted an empirical study addressing the influence of human testers' confirmation bias on software deficits using five datasets obtained from two companies. They found that the confirmation bias predicted 60-93% of the software deficits.

4.3. Primacy Effect

The primacy effect is a cognitive bias that refers to an individual's tendency to better recall information presented first in a sequence than information presented later (Asch, 1946; Bruce & Papay, 1970; Murdock, 1962). In contrast, the recency effect is the tendency to recall the last information presented (the most recently presented) best. For example, if we were asked to remember a long list of words, such as a shopping list, we would be more likely to remember the first and last items on the list better than those presented in the middle of the list. For illustrative purposes, we provide below some examples of how the primacy effect can affect the interaction of humans with AI.

An example of the primacy effect on users interacting with AI algorithms is observed in users when they search the Internet. People scan results in the order in which they are presented (e.g., Guan & Cutrell, 2007; Lorigo et al., 2008, for studies using eye-tracking technology). People click more and spend more time on first results than on later results (Insights, 2013; Donnini, 2013). For example, the study conducted by Insights (2013) showed that 91% of clicks occurred on the first page of the search results. Indeed, there is abundant research showing that the ranking of search engine results has a critical impact on users' attitudes, preferences, and behavior (e.g., Insights, 2013; Epstein & Robertson, 2015; Guan & Cutrell, 2007; Lorigo et al., 2008; Pan et al., 2007). There is evidence, for example, on the impact of Internet search ranking even on political preferences in democratic elections. In a series of experiments, Epstein & Robertson (2015) studied the voting preferences of 4,556 undecided voters of the United States and India by presenting to participants a Google-like search engine ranking of results to assess whether the order of information displayed could have an effect on participants. Using the 2014 Indian elections as the context for the experiment, the task consisted on participants obtaining information about the candidates in a search engine in which the order of the results was manipulated so that the firstly presented pages favored one or the other candidate. According to the authors, the participants considered more relevant those search results that were presented in the first

positions, changing the voting preferences by 20%. In addition, up to 85% of the participants showed no awareness of the manipulation. The authors refer to this effect as a search engine manipulation effect. Given that people show this primacy effect when interacting with AI, AI itself can be affected by primacy bias. Since AI uses data on user behavior for its learning and this behavior may reflect higher consumption of the top positions of a piece of content, the AI may learn that this content is more successful and relevant, which in turn increases the likelihood that the AI will show it in the top positions in the future. Developer and activist O'Neil (2016) has referred to this phenomenon as a *biased feedback loop.*

4.4. Anchoring Bias

The anchoring bias refers to the tendency to rely excessively on the first information that we learn (the anchor) and to generate adjustments in our evaluations from this starting point (Tversky & Kahneman, 1974). For example, when we go shopping and some item is on sale, the previous price (usually shown crossed out) acts as an anchor to evaluate whether the current price is appealing.

In relation to AI, this anchoring bias may affect users who use shopping platforms with AI. For example, a study examining 28 million Amazon ratings on more than 1.7 million products found that the product's published rating was used as a starting point (anchor), and then the users adjusted their ratings accordingly (Wang & Wang, 2014). In this case, AI amplifies the impact of the user's bias by interacting with thousands of users affected by the anchoring bias who are evaluating the same product.

The anchoring bias can occur unintentionally in the layout of the information in the user interface, but it can also be intentionally promoted by the company owning the AI platform (Yeung, 2018). This is the case described by Eslami et al. (2017), who observed that the rating algorithm of Booking.com biased low hotel scores upward by increasing the anchor value, particularly for low and medium-quality hotels. To this end, the company asked users to rate the different aspects of the hotels (location, cleanliness...) on a scale where the lowest possible score was 2.5 points, although users searching for hotels are actually led to believe that they can search for hotels on the platform with scores lower than that (e.g., 1 point). Presumably, this is a bias purposely introduced by the company that owns this AI. Like with the primacy bias, however, the AI algorithms interact and learn from the biased user behavior, so there is a risk that AI will end up inheriting and even amplifying this bias by creating, once again, a biased feedback loop (O'Neil, 2016).

4.5. Causal Illusion

Causal learning is the cognitive process of inferring causal relationships from available information. It is an efficient and rapid mechanism that has conferred important survival advantages to humans and other animals throughout evolution (Blanco & Matute, 2018; Shanks & Dickinson, 1987; Wasserman, 1990). But it is not error-free. The illusion of causality, or causality bias, occurs when people believe that there is a causal relationship between two events that often occur together but are not causally related (see Matute et al., 2015; 2022 for review). For example, this illusion occurs when a student believes that wearing an amulet to an exam causes getting a good grade, or when a patient believes that a pseudotherapy with no effects has curative properties.

This illusion is particularly common when the desired effect occurs frequently, as is the case, for instance, in some diseases that involve frequent spontaneous remissions (e.g., back pain). In those cases, people tend to associate the remissions of their painful crises with the remedy that they just

took. Sometimes this bias of causality can occur when there is not even a correlation between the two events; sometimes correlation occurs and is confounded with causality. Indeed, when both the potential cause and the potential effect are frequent and co-occur frequently (even by mere chance), the probability that people will associate them to each other and infer an illusory causal relationship between them increases (Blanco et al., 2013).

In some of these cases, the illusion of causality between two events is favored by the existence of a third, hidden, unobserved variable (Matute et al., 2019). For example, a causal illusion may occur when people infer that the increase in a child's height causes a greater acquisition of knowledge, instead of realizing that both height and knowledge increase with age. A similar problem can also occur in AI when developers use proxies in the development of the AI model. For instance, a proxy variable is one that correlates with the inferred cause and is therefore used as the main predictor variable. However, as we mentioned above, the fact that two variables seem to be related does not imply that there is either causality or correlation between them. Such apparent correlation might be due to a third, hidden variable.

An important research that clearly addresses this problem is the one conducted by Obermeyer et al. (2019). They found that an algorithm used to guide the medical care of approximately 200 million people each year was biased against black people. In particular, the results showed that the algorithm assigned the same level of risk to white patients and black patients, when actually black patients were sicker. The bias occurred because the model used the proxy of history of health care costs to predict who would need additional medical care. However, the health care costs reflect racial biases in the historical data because less money has been spent historically on caring black patients than white patients. In this case, a third hidden variable, namely, race, was behind the relationship between the medical cost and needs.

In a similar vein, several proxy variables are often used as if they had psychometric properties even though they lack such properties. An example is offered by Duportail (2019) who explains in her book how the Tinder algorithm may be using the length of sentences and words people type on the dating platform as a measure to infer the intelligence of candidates. This measure of intelligence implies a reduction of a variable as complex as intelligence to a simplistic proxy without relying on evidence-based psychological principles about intelligence (e.g., see Wechsler et al., 2008, for a widely used intelligence test).

In sum, causal learning is today a complex and vibrant research area both in humans and AI, and it will not be easy to overcome causal biases in either agent. With respect to human biases, which are the main topic of this article, some evidence-based debiasing strategies have been developed and published in recent years, and it appears that, at least under controlled laboratory (see Matute et al., 2015, 2019, for reviews) and classroom conditions (e.g., Barberia et al., 2013; 2018), progress is being made. Nonetheless, much research is still necessary to generalize these findings to larger populations and more naturalistic settings.

5. Discussion

The goal of this review was twofold. First, to contribute to the study of human cognitive biases present in AI. And second, to make use of knowledge that psychology has accumulated about biases to advance our knowledge of how these biases could impact AI, and thus, be minimized. To this end, in this review

we describe some examples of human cognitive biases that impact AI as well as some possible contributions of psychology in the investigation of these biases.

As we discussed above, today many of our decisions are delegated to AI algorithms. The term AI algorithm can be approached from different perspectives, the most frequent being the purely technical perspective. Most of the published studies on AI come only from the area of Computer Science. This technical perspective also affects how people define AI (generally as a mathematical and logical process; Logg et al., 2019), how AI is perceived (people tend to consider algorithms as more objective, neutral, rational, and free of bias than humans; e.g., Araujo et al., 2020) and what decisions humans and AI should make (with a greater preference in some contexts for the judgment or recommendations of AI over those of humans; Logg et al., 2018). However, it is currently being pointed out that the conceptualization of AI should start from a broader socio-technical perspective (Elish & Boyd, 2018; Kitchin, 2017) including the conditions in which it is developed and deployed (Geiger, 2014). In this regard, the need to conduct studies to examine the behavior of algorithms from a multidisciplinary approach has also been highlighted. Furthermore, in contrast to the image of objectivity and neutrality, there is growing evidence that AI can contain a number of important biases with serious consequences (Bolukbasi et al., 2016; Caliskan et al., 2017; Sweeney, 2013).

Therefore, given that a perspective that is not purely technical in the study of AI is necessary, that multidisciplinary research is recommended, and that there is extensive literature on the biases exhibited by AI, we suggest that psychology can contribute significantly to research on the behavior of AI and on how we humans interact with AI. Among the possible contributions of psychology, we especially highlighted two of them. On the one hand, the background of psychological research in areas which are very close to the emerging field in AI, such as computational psychology (Anderson et al., 2008; Eysenck & Brysbaert, 2018; Sun, 2001). These research areas make use of computer models to understand human cognition. On the other hand, we discussed the extensive psychological literature on the study of human cognitive biases (e.g., Kahneman & Tversky, 1972; Kahneman et al., 2021; Gigerenzer & Todd, 1999), in our attempt to contribute to the study of human biases present in AI due to human intervention and interaction with AI.

We also presented some examples of the impact of human cognitive biases on AI, in particular, the confirmation bias, the primacy effect, the representativeness bias, the anchoring bias, and some problems related with causality. These cognitive biases are introduced in different phases of the AI development where humans are involved because people in general (not only those humans who develop the AI) tend to be biased (e.g., Dawes, 2001).

In this article, we have discussed only how human cognitive biases may impact AI. Thus, we did not discuss discriminatory biases (such as racial and gender biases) directly in this article. Instead, we focused our discussion on more basic principles of cognition and cognitive biases that may affect our interactions with AI, because they are more general and are, quite possibly, at the base of the discriminative biases as well. As an example, the racial and gender biases present in AI that often cause such a large impact at the media level, as we mentioned in the introduction, could be in part the result of the representativeness bias described above. The reason is that most AI models are developed and trained in western countries by teams composed predominantly by white men who use data coming predominantly from the datasets easily available in western countries to train their models. We suggest that understanding cognitive biases in humans and reducing their impact in AI should also contribute to the reduction of discriminatory social biases present in AI.

As we already mentioned, human cognitive biases are associated to various threats to human welfare (e.g., Crocker, 1981; Hamilton & Gifford, 1976; Murphy et al., 2011), but today, the presence of human

biases in AI is augmenting this problem and has a very negative impact on our societies. Indeed, beyond reflecting the biases we have as individuals, the biases presented in the algorithms tend to be systematically amplified (e.g., Kay et al., 2015), due to the large amount of data that the algorithms handle and their widespread use. Moreover, AI also can amplify the impact of these biases, since sometimes an AI model is reused as a basis for developing other algorithms. If this base AI model is affected by biases, the other algorithms which are developed after it will inherit its biases, thereby amplifying its negative impact. This was the case experienced firsthand by AI researcher Buolamwini (2016). When she was a computer science student at Georgia Tech, the social robot she worked with could not identify her face because of the color of her skin. Sometime later, while participating in a demonstration of another social robot in Hong Kong, the machine learning algorithm recognized all the participants in the demo except her. The robot used the same generic facial recognition software as the robot in Georgia. As she said, the "algorithmic bias can travel as quickly as it takes to download some files off of the Internet". Biases present in AI can now spread rapidly and globally through the Internet.

In addition, decisions being made, by default, by algorithms that have inherited human biases can affect severely millions of people, due to the increased acceptance and use of algorithms by the general public, companies, and governments for high-impact decisions (e.g., Bartlett et al., 2022), for bank loans; (Obermeyer et al., 2019), for medical decisions; (Ferrara et al., 2016), and for the identification of suspicious profiles (e.g., Dressel & Farid, 2018), among others. In addition, biases present in AI may also involve the added risk of perpetuation of human biases when AI uses biased user information or behavior in its learning and feedback processes. This situation would favor the perpetuation of such biases (O'Neil, 2016). An example of this perpetuation of biases is the software used by the Pennsylvania police to predict crime, PredPol. This model may appear free of bias because it uses geography as a proxy for the probability of crime, ignoring the individual's race and ethnicity. However, PredPol can be configured to take into account in its predictions not only violent crimes, but also minor crimes, such as vagrancy or the selling small amounts of drugs. These types of crimes, endemic to many impoverished neighborhoods, feed the algorithm and cause police to be sent to these geographic areas more frequently, which in turn increases the likelihood that more crimes will be recorded there and more new data will be generated that will reinforce the algorithm's future predictions in this way, perpetuating the bias in this way (O'Neil, 2016).

Our main conclusion is that human cognitive biases are present in many AI agents at the moment since we humans are involved in their life cycle and are prone to a biased interpretation of reality. This problem can go unnoticed because most people have difficulty recognizing their own biases (Pronin, Lin, & Ross, 2002) and because given the lack of transparency of most AI agents, most human biases present in algorithms are detected a posteriori. Given that algorithms are intrinsically framed and shaped in a socio-technical context (Geiger, 2014; Napoli, 2013; Takhteyev, 2012), AIs should be transparent and reliable, and should guarantee that they meet the goals and ethical demands of the society they are designed to serve.

11. References

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. *PLoS ONE 16*(4): e0249454. https://doi.org/10.1371/journal.pone.0249454

Alloy, L. B., & Clements, C. M. (1992). Illusion of control: invulnerability to negative affect and depressive symptoms after laboratory and natural stressors. *Journal of abnormal Psychology, 101*(2), 234-245.

Alonso, E., Mondragón, E., & Fernández, A. (2012). A Java simulator of Rescorla an' Wagner's prediction error model and configural cue extensions. *Computer Methods and Programs in Biomedicine, 108*(1), 346-355.

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences, 12*(4), 136–143. https://doi.org/https://doi.org/10.1016/j.tics.2008.01.006

Angwin, J. (2017, November 29). Facebook to temporarily block advertisers from excluding audiences by race. *ProPublica*. https://www.propublica.org/article/facebook-to-temporarily-block-advertisers-from-excluding-audiences-by-race

Angwin, J., Tobin, A., & Varner, M. (2017, November 21). Facebook (still) letting housing advertisers exclude users by race. *ProPublica*. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Araujo, T., Helberger, N., Kruikemeier, S., Vreese, C. H. de, & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society, 35*(3), 1–13. https://doi.org/10.1007/S00146-019-00931-w

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*(3), 486–498. https://doi.org/10.1037/0033-2909.110.3.486

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 258–290. https://doi.org/10.1037/H0055756

Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ArXiv*. http://arxiv.org/abs/1905.04223

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61. https://doi.org/10.1145/3209581

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*(1), 1–24. https://doi.org/10.1037/0033-2909.133.1.1

Barberia, I., Blanco, F., Cubillas, C. P., & Matute, H. (2013). Implementation and assessment of an intervention to debias adolescents against causal illusions. *PLoS ONE, 8*(8), e71303. https://doi.org/10.1371/journal.pone.0071303

Barberia, I., Tubau, E., Matute, H., & Rodríguez-Ferreiro, J. (2018). A short educational intervention diminishes causal illusions and specific paranormal beliefs in undergraduates. *PloS ONE, 13*(1), e0191907. https://doi.org/10.1371/journal.pone.0191907

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Bickmore, T. W., Utami, D., Matsuyama, R., & Paasche-Orlow, M. K. (2016). Improving access to online health information with conversational agents: A randomized controlled experiment. *Journal of Medical Internet Research, 18*(1). https://doi.org/10.2196/jmir.5239

Blanco, F. (2017). Positive and negative implications of the causal illusion. *Consciousness and Cognition, 50*, 56–68. https://doi.org/10.1016/J.CONCOG.2016.08.012

Blanco, F., Barberia, I., & Matute, H. (2014). The lack of side effects of an ineffective treatment facilitates the development of a belief in its effectiveness. *PLoS ONE, 9*(1), e84084. https://doi.org/10.1371/JOURNAL.PONE.0084084

Blanco, F., Barberia, I., & Matute, H. (2015). Individuals who believe in the paranormal expose themselves to biased information and develop more causal illusions than nonbelievers in the laboratory. *PLoS ONE, 10*(7), e0131378. https://doi.org/10.1371/journal.pone.0131378

Blanco, F., & Matute, H. (2018). The illusion of causality: A cognitive bias underlying pseudoscience. *Pseudoscience: The Conspiracy against Science*, 45–76. https://doi.org/10.7551/mitpress/9780262037426.003.0003

Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning and Behavior, 41*(4), 333–340. https://doi.org/10.3758/S13420-013-0108-8

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Brooks, S., Prince, A., Stahl, D., Campbell, I. C., & Treasure, J. (2011). A systematic review and meta-analysis of cognitive bias to food stimuli in people with disordered eating behaviour. *Clinical Psychology Review, 31*(1), 37–51. https://doi.org/10.1016/j.cpr.2010.09.006

Bruce, D., & Papay, J. P. (1970). Primacy effect in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 9*(5), 473–486. https://doi.org/10.1016/S0022-5371(70)80090-1

Buolamwini, J. (2016). How I'm fighting bias in algorithms [Video]. TED Conferences. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Calikli, G., & Bener, A. B. (2013). Influence of confirmation biases of developers on software quality: An empirical study. *Software Quality Journal, 21*(2), 377–416. https://doi.org/10.1007/S11219-012-9180-0

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cappelli, P., Tambe, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3263878

Carlson, M. (2017). Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism: *New Media & Society, 20*(5), 1755–1772. https://doi.org/10.1177/1461444817706684

Coley, R. Y., Johnson, E., Simon, G. E., Cruz, M., & Shortreed, S. M. (2021). Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatry,* 78(7), 726–734. https://doi.org/10.1001/jamapsychiatry.2021.0493

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*(2), 272–292. https://doi.org/10.1037/0033-2909.90.2.272

Croskerry, P. (2013). From mindless to mindful –actice--cognitive bias and clinical decision making. *The New England Journal of Medicine, 368*(26), 2445–2448. https://doi.org/10.1056/NEJMP1303712

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology, 29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Dawes, R. M. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally.* Westview.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/SCIENCE.2648573

de Pessemier, T., Vanhecke, K., & Martens, L. (2016). Scalable, high-performance algorithm for hybrid job recommendations. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2987538.2987539

de Raedt, R., & Koster, E. H. W. (2010). Understanding vulnerability for depression from a cognitive neuroscience perspective: A reappraisal of attentional factors and a new conceptual framework. *Cognitive, Affective & Behavioral Neuroscience, 10*(1), 50–70. https://doi.org/10.3758/CABN.10.1.50

Devries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? *ArXiv.* https://doi.org/10.48550/arXiv.1906.02659

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism, 5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour and Information Technology, 17*(3), 155–163. https://doi.org/10.1080/014492998119526

Donnini, G. (2013, July 22). The value of Google result positioning. *Search Engine Journal*. https://www.searchenginejournal.com/the-value-of-google-result-positioning/65176/

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Duportail, J. (2019). *El algoritmo del amor: Un viaje a las entrañas de Tinder.* Contra.

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). When people and algorithms meet. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 96–106. https://doi.org/10.1145/3301275.3302262

Elish, M. C., & Boyd, D. (2018, November 13). Don't believe every AI you see. *The Ethical Machine.* https://ai.shorensteincenter.org/ideas/2018/11/12/dont-believe-every-ai-you-see-1

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; Things can be worse than they – appear" - Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM* 2017, 62–71.

Eysenck, M. W., & Brysbaert, M. (2018). *Fundamentals of cognition.* Routledge.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717

French, M. R. (2018). *Algorithmic mirrors: An examination of how personalized recommendations can shape self-perceptions and reinforce gender stereotypes* [Doctoral dissertation, Stanford University]. http://purl.stanford.edu/wd112bf7670

Fry, H., & Krzywinski, M. (2019). *Hola mundo: Cómo seguir siendo humanos en la era de los algoritmos.* Blackie Books.

Garb, H. N. (1998). Studying the clinician: Judgment research and psychological assessment. In *Studying the clinician: Judgment research and psychological assessment.* American Psychological Association. https://doi.org/10.1037/10299-000

Garcia, R., Sreekanti, V., Yadwadkar, N., Crankshaw, D., Gonzalez, J. E., & Hellerstein, J. M. (2018). Context: The missing piece in the machine learning lifecycle. *CMI'18.* https://rlnsanz.github.io/dat/Flor_CMI_18_CameraReady.pdf

Gawęda, Ł, Prochwicz, K., & Cella, M. (2015). Cognitive biases mediate the relationship between temperament and character and psychotic-like experiences in healthy adults. *Psychiatry Research, 225*(1), 50–57. https://doi.org/10.1016/j.psychres.2014.10.006

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information Communication and Society, 17*(3), 342–356. https://doi.org/10.1080/1369118X.2013.873069

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

Griffiths, O., Shehabi, N., Murphy, R. A., & Le Pelley, M. E. (2019). Superstition predicts perception of illusory control. *British Journal of Psychology, 110*(3), 499–518. https://doi.org/10.1111/bjop.12344

Grolleman, J., van Dijk, B., Nijholt, A., & van Emst, A. (2006). Break the habit! Designing an e-therapy intervention using a virtual coach in aid of smoking cessation. In W. A. IJsselsteijn, Y. A. W. de Kort, C. Midden, B. Eggen, & E. van den Hoven (Eds.), *Persuasive Technology* (pp. 133–141). Springer Berlin Heidelberg.

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 417–420. https://doi.org/10.1145/1240624.1240691

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12*(4), 392–407. https://doi.org/10.1016/S0022-1031(76)80006-6

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Herrmann, D. J. (2004). The potential of cognitive technology. In W. R. Walker & D. J. Herrmann (Eds.), *Cognitive technology: Essays on the transformation of thought and society* (pp. 5–19). McFarland & Company.

Howard, A., & Borenstein, J. (2017). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics, 24,* 1521–1536. https://doi.org/10.1007/S11948-017-9975-2

Hudlicka, E. (2013). Virtual training and coaching of health behavior: Example from mindfulness meditation training. *Patient Education and Counseling, 92*(2), 160–166. https://doi.org/10.1016/j.pec.2013.05.007

IBM. (2020). *How to get started with cognitive technology*. https://www.ibm.com/watson/advantage-reports/getting-started-cognitive-technology.html

Insights, C. (2013).*The value of Google result positioning.* Westborough: Chitika Inc, 0-10. https://www.silesiasem.pl/wp-content/uploads/2013/07/chitikainsights-valueofgoogleresultspositioning.pdf

Isidore, C. (2018, February 6). Machines are driving Wall Street's wild ride, not humans. CNN. https://money.cnn.com/2018/02/06/investing/wall-street-computers-program-trading/index.html

Johnson, D. D. (2004). *Overconfidence and war: The havoc and glory of positive illusions.* Harvard University Press.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *The American Psychologist, 58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment.* Little, Brown Spark.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. https://doi.org/10.1037/H0034747

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–591. https://doi.org/10.1037/0033-295X.103.3.582

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,* 3819–3828. https://doi.org/10.1145/2702123.2702520

Kennedy, L. W., Caplan, J. M., & Piza, E. (2011). Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *Journal of Quantitative Criminology, 27*(3), 339–362. https://doi.org/10.1007/s10940-010-9126-2

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information Communication and Society, 20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kremin, H., Akhutina, T., Basso, A., Davidoff, J., de Wilde, M., Kitzing, P., Lorenz, A., Perrier, D., van der Sandt-Koenderman, M., Vendrell, J., & Weniger, D. (2003). A cross-linguistic data bank for oral picture naming in Dutch, English, German, French, Italian, Russian, Spanish, and Swedish (PEDOI). *Brain and Cognition, 53*(2), 243–246. https://doi.org/10.1016/S0278-2626(03)00119-2

Krieger, L. H., & Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review, 94*(4), 997–1062. https://doi.org/10.2307/20439058

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27*(3), 313–327. https://doi.org/10.1017/S0140525X04000081

Lambrecht, A., & Tucker, C. E. (2016). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2852260

Langer, E. J. (1975). The illusion of control. *Journal of personality and social psychology, 32*(2), 311-328.

Larrick, R. P. (2008). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 316–338). Wiley. https://doi.org/10.1002/9780470752937.CH16

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1). https://doi.org/10.1177/2053951718756684

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50*(1), 7–36. https://doi.org/10.1016/J.JSP.2011.09.006

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science, 4*(4), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances, 6*(7). https://doi.org/10.1126/sciadv.aaz0652

Logg, J. M., Minson, J. A., & Moore, D. A. (2018, October 26). Do people trust algorithms more than companies realize? *Harvard Business Review*. https://hbr.org/2018/10/do-people-trust-algorithms-more-than-companies-realize

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lorenz, L., Meijer, A., & Schuppan, T. (2021). The algocracy as a new ideal type for government organizations: Predictive policing in Berlin as an empirical case. *Information Polity, 26*(1), 71–86. https://doi.org/10.3233/IP-200279

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., & Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology, 59*(7), 1041–1052. https://doi.org/10.1002/asi.20794

Malmendier, U., & Tate, G. (2005). Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited. European Financial Management, 11(5), 649–659. https://doi.org/10.1111/J.1354-7798.2005.00302.X

Martínez, N., Vinas, A., & Matute, H. (2021). Examining potential gender bias in automated-job alerts in the Spanish market. *PLoS ONE 16*(12): e0260409. https://doi.org/10.1371/journal.pone.0260409

Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement in humans. *Learning and Motivation, 25*, 216-232.

Matute, H. (1996). Illusion of control: Detecting response-outcome independence in analytic but not in naturalistic conditions. *Psychological Science, 7*, 289-293. https://doi.org/10.1111/j.1467-9280.1996.tb00376.x

Matute, H., Blanco, F., & Díaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*(4), 373–389. https://doi.org/10.1037/xan0000222

Matute, H., Blanco, F., Moreno-Fernández, M.M. (2022). Causality bias. In R. F. Pohl (ed). *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory.* (3rd ed.). Routledge.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology, 6*, 1–14. https://doi.org/10.3389/fpsyg.2015.00888

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology, 102*(3), 392–405. https://doi.org/10.1348/000712610X532210

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think.* Houghton Mifflin Harcourt.

Mackenzie, A. (2007). Protocols and the irreducible traces of embodiment: The Viterbi algorithm and the mosaic of machine time. In R. Hassan & R. E. Purser (Eds.), 24/7: Time and temporality in the network society (pp. 89–106). Stanford, CA: Stanford University Press.

Muller-Lyer, F. C. (1889). Optische urteilstauschungen. *Archiv Fur Anatomie Und Physiologie, Physiologische Abteilung, 2*, 263–270.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482–488. https://doi.org/10.1037/h0045106

Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragón, E., & Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology, 64*(1), 24–40. https://doi.org/10.1080/17470218.2010.493615

Musca, S. C., Vadillo, M. A., Blanco, F., & Matute, H. (2010). The role of cue information in the outcome-density effect: Evidence from neural network simulations and a causal learning experiment. *Connection Science, 22*(2), 177–192. https://doi.org/10.1080/09540091003623797

Napoli, P.M. (2013). The algorithm as institution: Toward a theoretical framework for automated media production and consumption. *Fordham University Schools of Business Research Paper.* https://doi.org/10.2139/ssrn.2260923

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings in social judgement. In *Memory* (Issue 4). Prentice Hall.

Northpointe. (2019). *Practitioner's Guide to COMPAS Core*. Equivant. https://www.equivant.com/practitioners-guide-to-compas-core/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group.

Orgaz, C., Estevez, A., & Matute, H. (2013). Pathological gamblers are more vulnerable to the illusion of control in a standard associative learning task. *Frontiers in Psychology, 4.* https://doi.org/10.3389/fpsyg.2013.00306

OSTP-OPM (Office of Science and Technology Policy/ Office of Personnel Management; 2016). Reducing the impact of bias in the STEM workforce: Strengthening excellence and innovation. https://www.si.edu/content/OEEMA/OSTP-OPM_ReportDigest.pdf.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Patterson, S. (2012). *Dark pools: the rise of the machine traders and the rigging of the U.S. stock market*. Random House.

Peckham, A. D., McHugh, R. K., & Otto, M. W. (2010). A meta-analysis of the magnitude of biased attention in depression. *Depression and Anxiety, 27*(12), 1135–1142. https://doi.org/10.1002/da.20755

Pohl, R.F. (Ed.). (2022). *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (3rd ed.). Routledge.

Phua, D., & Tan, N. (2013). Cognitive aspect of diagnostic errors. *Annals of the Academy of Medicine, Singapore, 42*(1), 33–41.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Rahwan, I., & Cebrian, M. (2018). Machine behavior needs to be an academic discipline. *Nautilus*. http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). Appelton-Century-Crofts.

Rodger, J. A., & Pendharkar, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies, 60*(5–6), 529–544. https://doi.org/10.1016/j.ijhcs.2003.09.005

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General. 104*(3): 192–233.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*(5), 880–892. https://doi.org/10.1037/0022-3514.32.5.880

Salman, I. (2016). Cognitive biases in software quality and testing. 2016 *IEEE/ACM 38th International Conference on Software Engineering Companion* (ICSE-C), 823–826.

Salman, I., Turhan, B., & Vegas, S. (2019). A controlled experiment on time pressure and confirmation bias in functional software testing. *Empirical Software Engineering, 24*(4), 1727–1761. https://doi.org/10.1007/s10664-018-9668-8

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius.* https://doi.org/10.1177/2378023120967171

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 21, pp. 229– 261). Academic Press.

Shepperd, J.A., & Koch, E.J. (2005). Pitfalls in teaching judgment heuristics. *Teaching of Psychology, 32*, 43-46.

Slater, D. (2013). *Love in the time of algorithms: What technology does to meeting and mating.* Penguin.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stephens, A. N., & Ohtsuka, K. (2014). Cognitive biases in aggressive drivers: Does illusion of control drive us off the road? *Personality and Individual Differences, 68*, 124–129. https://doi.org/10.1016/j.paid.2014.04.016

Sun, R. (2001). *The Cambridge handbook of computational psychology.* Cambridge University Press. https://doi.org/10.1017/CBO9780511816772

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility, 73*–100. https://doi.org/10.1162/dmal.9780262562324.073

Sutton, R.S., & Barto, A.G. (1981). Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review, 88*, 135–170.

Sweeney, L. (2013). Discrimination in online ad delivery. *ArXiv.* http://arxiv.org/abs/1301.6822

Takhteyev, Y. (2012). *Coding places: Software practice in a South American city*. The MIT Press. https://mitpress.mit.edu/books/coding-places

Tavris, C., & Aronson, E. (2007). Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts. In *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. Harcourt.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Teasley, B. E., Leventhal, L. M., Mynatt, C. R., & Rohlman, D. S. (1994). Why software testing is sometimes ineffective: Two applied studies of positive test strategy. *Journal of Applied Psychology, 79*(1), 142–155. https://doi.org/10.1037/0021-9010.79.1.142

Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites. *Journalism Studies, 13*(5–6), 775–790. https://doi.org/10.1080/1461670X.2012.664341

Torres, M. N., Barberia, I., & Rodríguez-Ferreiro, J. (2020). Causal illusion as a cognitive basis of pseudoscientific beliefs. *British Journal of Psychology, 111*(4), 840–852. https://doi.org/10.1111/bjop.12441

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vadillo, M. A., Blanco, F., Yarritu, I., & Matute, H. (2016). Single- and dual-process models of biased contingency detection. *Experimental Psychology, 63*(1), 3–19. https://doi.org/10.1027/1618-3169/a000309

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Wang, T., & Wang, D. (2014). Why Amazon's ratings might mislead you: The story of herding effects. *Big Data, 2*(4), 196–204. https://doi.org/10.1089/big.2014.0063

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Penguin Books.

Wasserman, E. A. (1990). Detecting response- outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 27–82). Academic Press

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV technical and interpretative manual.* Pearson.

Xu, C., & Doshi, T. (2019, December 11). Fairness indicators: Scalable infrastructure for fair ML systems. *Google AI Blog*. https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html

Yarritu, I., Matute, H., Luque, D. (2015). The dark side of cognitive illusions: When an illusory belief interferes with the acquisition of evidence-based knowledge. *British Journal of Psychology, 106*, 597-608. https://doi.org/10.1111/bjop.12119

Yeung, D. (2018, October 19). When AI misjudgment is not an accident. *Scientific American*. https://blogs.scientificamerican.com/observations/when-ai-misjudgment-is-not-an-accident/