Cognitive Biases Undermine Consensus on Definitions of Intelligence and Limit Understanding

Dagmar Monett[1,2*], **Luisa Hoge**[1] and **Colin W. P. Lewis**[2]
[1]Computer Science Dept., Berlin School of Economics and Law, Germany
[2]AGISI.org
{dagmar.monett, colin.lewis}@agisi.org, s_hoge@stud.hwr-berlin.de

Abstract

There are several reasons for the lack of a consensus definition of (machine) intelligence. The constantly evolving nature and the interdisciplinarity of the Artificial Intelligence (AI) field, together with a historical polarization around what intelligence means, are among the most widely discussed rationalizations, both within the community and outside it. These factors are aggravated by the presence of cognitive biases in subjective reasoning by experts on the definition of intelligence, as we have found in a recent study of experts' opinions across multiple disciplines. In this paper, we show how different cognitive biases can undermine consensus on defining intelligence, and thus how an understanding of intelligence can be substantially affected by these human traits. We also provide general recommendations for tackling these problems. An understanding of intelligence can be achieved by understanding the limits of both human expressiveness and the current discourse around definitions of intelligence within and across the concerned fields.

1 Introduction

In a recent extended report [Committee on AI, 2018] answering a call for written evidence on the current state of *"the economic, ethical and social implications of advances in artificial intelligence,"* the Select Committee on Artificial Intelligence (AI) appointed by the House of Lords in the UK concluded that *"there is no widely accepted definition of artificial intelligence. Respondents and witnesses provided dozens of different definitions."* This has been a recurrent and unwanted aspect of the AI community: since its formation as a field more than six decades ago, numerous academics have pressed for an agreed upon definition, but it has not been possible even to reach consensus on the need for one. *"When we talk of intelligence, we don't really know what we are talking about. There seems to be no generally accepted definition of what 'intelligence' is,"* writes Kugel [2002] revisiting what he thinks Alan Turing meant by attributing intelligence to computing machines. *"The problem is that we cannot yet characterize in general what kinds of computational procedures we want to call intelligent. We understand some of the mechanisms of intelligence and not others,"* pointed out McCarthy [2007], one of the founding fathers of AI, a few years later.

Intelligence is not the only fundamental concept that does not have a consensus on its definition. Similar problems have arisen for other concepts and there is a lack of well-defined or consensus definitions for several concepts in several domains. For example, in both the intelligence and counterintelligence fields, *"[t]he term 'intelligence' has far been used without clearly defining it. . . . All attempts to develop ambitious theories of intelligence have failed"* [Laqueur, 1985]. Furthermore, in the field of intelligence research Hunt and Jaeggi [2013] write *"after 100 years of research, the definition of the field is still inadequate."* In the field of computer science, the concept of *model interpretability* in Machine Learning (ML) is crucial for understanding the decision-making processes of ML models; however, this is an ill-defined concept that only a few authors have precisely articulated in the academic literature [Lipton, 2018b]. The concept of *privacy* has also been hard to define: despite many attempts having been made so far, no consensus definition has been found. There is also a lack of a concrete definition of *fairness*, due to an *"explosion of definitions in the computer science literature"* in recent years [Chouldechova and Roth, 2018]. This makes *"the detailed differences between multiple definitions [of fairness] difficult to grasp"* [Verma and Rubin, 2018].

Several factors contribute to the lack of a consensus definition. For example, many different contexts, applications, and stakeholders may deal with the same concept but from different perspectives related to their specific fields. In Bimfort's [1958] words, *"[e]ach expert tends to view the term through the spectacles of his specialty."* This is not very different from what happens with AI: *"what AI includes is constantly shifting"* [Luckin *et al.*, 2016], i.e. the field and the applications that include AI are constantly evolving, and its interdisciplinary nature might work against the development of a consensus definition [Luckin *et al.*, 2016]. In addition to this, we are dealing with a very polarized concept: *"The debate around exactly what is, and is not, artificial intelligence, would merit a study of its own"* [Committee on AI, 2018].

Other reasons include the fact that *"[c]riticism of intelligence has been partially based on exaggerated notions of what it can, and can not, accomplish"* [Laqueur, 1985]. Iron-

*Contact Author

ically, Laqueur refers here to the concept of intelligence in the intelligence and counterintelligence fields, but this also applies fully to AI. Jordan [2018] has recently warned that *"we are very far from realizing human-imitative AI aspirations. Unfortunately the thrill (and fear) of making even limited progress on human-imitative AI gives rise to levels of over-exuberance and media attention that is not present in other areas of engineering."*

However, other scientific communities have been able to acknowledge the need for a serious discussion around defining their most fundamental concepts, in order to reach consensus on the *what*, the *how*, and the *why* of these concepts [Daar and Greenwood, 2007; Gottfredson, 1997; Kaufman, 2019] and to move forward, or at least to finish or put aside fruitless debates. This has not been the case in the AI community, at least up until now.

2 Reasons for a Consensus Definition

There are several pressing reasons for a consensus definition of machine (or artificial) intelligence, including the following:

Transparency, Understanding, Sustainability: If one of the goals is to develop algorithms and machines that improve the well-being of individuals, then since many of these systems are increasingly using data and information on these individuals to aid in decision-making, it is of utmost importance that they know and understand how these systems work with their data, process it, and make decisions that can potentially affect their lives. However, *"[t]he public knowledge and understanding on AI . . . is suffering from a lack of transparency as to capabilities and thus impacts of AI"* [Nemitz, 2018]. Hence, *"to achieve sustainable change towards socially just and transparent AI development beyond a framing of data ethics as competitive advantage . . . , it is paramount to consider [among other points, that] we need a clear picture of AI"* [Sloane, 2018].

Governance, Regulation: The quickly evolving and transformative character of AI algorithms and systems in several spheres of modern life are increasingly demanding a balance between innovation and regulation, without similar precedents. The question of how to guarantee that these algorithms and systems are researched, developed, and deployed in ways that not only advance but also protect humanity against possible harm implies also thinking about their governance [Dafoe, 2018; Gasser and Almeida, 2017]. Thus, *"having a usable definition of AI–and soon–is vital for regulation and governance because laws and policies simply will not operate without one"* [Lea, 2015] because *"AI cannot and will not serve the public good without strong rules in place"* [Nemitz, 2018].

Media, Hype: Misleading media coverage raises false expectations of real progress in AI and creates ambiguity in funding situations. As Lipton [2018a] emphasizes, *"[t]he lack of specificity allows journalists, entrepreneurs, and marketing departments to say virtually anything they want."* The hyped tone not only misinforms the general public but also diverts important research into monolithic thinking about what AI is. AI is not only deep learning,[1] and is not even only ML! This has caused a negative view of AI and its applications by the public, *"which in their view had largely been created by Hollywood depictions and sensationalist, inaccurate media reporting . . . concentrating attention on threats which are still remote, such as the possibility of 'superintelligent' artificial general intelligence, while distracting attention away from more immediate risks and problems"* [Committee on AI, 2018].

Documenting: Even for documenting the evolution of AI as a field, defining it and its goals is crucial. Some recent works, such as [Martínez-Plumed *et al.*, 2018], have used AI to shed light on its evolution, but *"a lack of clarity in terms of definitions and objectives seems to have plagued the [AI] field right back to its origins in the 1950s. This makes tracing [its] evolution . . . a difficult task"* [Committee on AI, 2018].

Understanding, Development: The lack of a clear definition of intelligence is a perceived stumbling block to the pursuit of understanding intelligence and building machines that replicate and exceed human intelligence [Brooks, 1991]. As is the case in the current discourse, the confusing use of concepts such as AI, ML and deep learning, for example, is not only problematic but also *"prevents more productive conversations about the abilities and limits of such technologies"* [Sloane, 2018].

Achieving a consensus definition is not straightforward. When asked about the possibility of reaching agreement on a definition of artificial intelligence, almost 60% of respondents to the AGISI research survey on defining intelligence [Monett and Lewis, 2018] believed that it would be possible to reach consensus, compared to one-third of the respondents who believed the opposite. Nevertheless, the view that a definition of intelligence is not self-evident was supported by more than 80% of these participants.

If a concept is ill-defined, it cannot be well understood. We believe that a definition of intelligence based on concepts that are themselves well-defined is a fundamental milestone that must be reached prior to understanding this concept. This is also important in understanding its limits:[2] as we show in the next sections, different cognitive biases can undermine the consensus on definitions of intelligence, and thus its understanding can be substantially affected by these human traits.

3 Dissecting Written Opinions on Intelligence

We analyzed a corpus of more than 4,000 expert opinions, which was obtained from the survey on defining intelligence referenced above. The diversity of opinions reflects the diversity of the respondents, and different research fields, who originated from 57 countries and more than 184 different institutions around the world. They worked mainly in academia ($N = 441$, 79.3%) and industry ($N = 114$, 20.5%), and their primary roles were researchers ($N = 424$, 76.3%) and/or educators ($N = 193$, 34.7%), as described in [Lewis and Monett, 2018; Monett and Lewis, 2018].

[1] We recommend to read Darwiche's insightful paper *Human-Level Intelligence or Animal-Like Abilities?* [Darwiche, 2018].

[2] And thereby the limits, and also the risks, of intelligent technologies, as pointed out by Leetaru in [Leetaru, 2018].

Participants to the survey were presented with different definitions of machine and human intelligence from the literature, with nine definitions in each group: *MI1* to *MI9* for machine intelligence, and *HI1* to *HI9* for human intelligence. They are presented in Table 1. These definitions were first provided in historical (published date) order and then in alphabetical order, starting with the surname of the first cited author. Literature references and any information about the authors were deliberately omitted. Respondents were asked to rate their level of agreement with each definition by selecting an option from a five-point Likert rating scale ranging from *"1=Strongly disagree"* to *"5=Strongly agree."* They then had the option of arguing or justifying their selection by providing an open-ended answer.

A total of 4,041 respondents' opinions were collected this way, constituting a corpus with 2,424 opinions on the definitions of machine intelligence and 1,617 on the definitions of human intelligence extracted from a total of 556 survey responses. Nine comments were not considered for processing, since they had a URL as their only content. Not all of the respondents provided their reasons for or against the definitions from the literature and not all definitions were commented alike; some definitions polarized respondents more than others and the length of the comments varied significantly from respondent to respondent.

In the following subsections, the different cognitive biases that might be present in the collected respondents' opinions are analyzed together with their possible explanations. This is the main focus of this paper. Thus, other survey results and analyses are out of scope here for space limit reasons; they are included in separate papers.

3.1 Anchoring Effect

For the first 220 responses (39.6% of the total of responses that were collected), the percentages of positive agreement (i.e. the ratings of *"Strongly agree"* or *"Agree"*) with the definitions of machine intelligence show a decreasing trend line in a linear approximation with the definitions from the literature that were presented for agreement (see the darkest trend line in Figure 1).

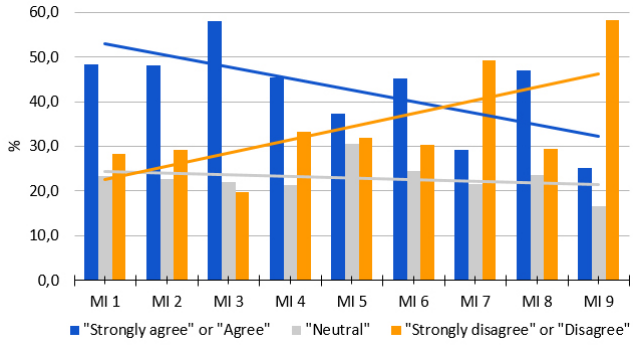

Figure 1: Level of agreement with the definitions of machine intelligence for the first 220 responses.

All of the definitions had a fixed position on the list: *MI1* was presented at position 1, *MI2* at position 2 and so on. There was therefore the possibility of a strong dependence between the percentage of positive opinions and the position of a definition on the list. Furthermore, the percentage of negative agreement shows an increasing trend line, opposite to that for positive agreement. The percentages of neutral answers remained quite stable.

It appears that respondents tended to rely heavily on the first definitions (the *anchors*) that were presented. This is a cognitive bias known as *anchoring*, or the *anchoring effect*, which is present when *"different starting points yield different estimates, which are biased toward the initial values"* or anchors [Tversky and Kahneman, 1974].

That the percentages of both positive and negative agreement might depend on the position of the definition in the list was first noticed after a partial analysis of the responses from the first 220 participants, as mentioned above. A reordering of the positions of the definitions was used from then on: the definitions were shuffled after every 56 responses on average (this varied depending on the flux of responses) with the hope that all of them would have the same probability of being anchors. A total of six random shuffles were made before the survey was closed, and the last 336 responses (60.4% of the total) were collected in this way.

The results were as expected: seven of the nine definitions of machine intelligence benefited from this shuffling (see Figure 2).

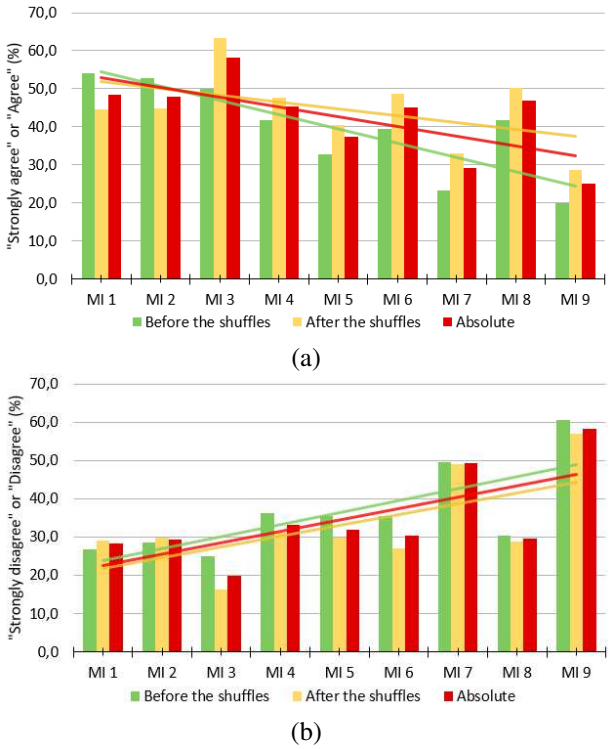

Figure 2: (a) Positive and (b) negative levels of consensus with the definitions of machine intelligence before and after reordering ($N = 556$).

Both the percentages of positive agreement after shuffling

Table 1: Definitions of machine and human intelligence that were presented to the survey participants.

Id.	Definition	How published
MI1	*"Artificial Intelligence is . . . the study of the computations that make it possible to perceive, reason, and act."*	Winston, P. H. (1992). Artificial Intelligence. Third Edition, Addison-Wesley Publishing Company.
MI2	*"Intelligence measures an agent's ability to achieve goals in a wide range of environments."*	Legg, S. and Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. Minds and Machines, 17(4):391-444, Springer.
MI3	*"The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets "intelligence" as a form of "relative rationality." "*	Wang, P. (2008). What Do You Mean by "AI"? In P. Wang, B. Goertzel, and S. Franklin (eds.), Artificial General Intelligence 2008, Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications, 171:362-373. IOS Press Amsterdam, The Netherlands.
MI4	*"The goal is to build computer systems that exhibit the full range of the cognitive capabilities we find in humans. . . . The ability to pursue tasks across a broad range of domains, in complex physical and social environments. [A human-level intelligence] system needs broad competence. It needs to successfully work on a wide variety of problems, using different types of knowledge and learning in different situations, but it does not need to generate optimal behavior."*	Laird, J. E., Wray, R. E., and Langley, P. (2009). Claims and Challenges in Evaluating Human-Level Intelligent Systems. In B. Goertzel, P. Hitzler, and M. Hutter (eds.), Proceedings of the Second Conference on Artificial General Intelligence. Atlantis Press.
MI5	*"Pragmatic general intelligence measures the capability of an agent to achieve goals in environments, relative to prior distributions over goal and environment space. Efficient pragmatic general intelligences measures this same capability, but normalized by the amount of computational resources utilized in the course of the goal-achievement."*	Goertzel, B. (2010). Toward a Formal Characterization of Real-World General Intelligence. In E. B. Baum, M. Hutter, and E. Kitzelmann (eds.), Artificial General Intelligence, Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010, Lugano, Switzerland, March 5-8, 2010 pp. 19-24. Advances in Intelligent Systems Research 10. Amsterdam: Atlantis.
MI6	*"Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."*	Nilsson, N. J. (2010). The Quest for Artificial Intelligence. A History of Ideas and Achievements. Cambridge University Press.
MI7	*"Intelligence is concerned mainly with rational action. Ideally, an intelligent agent takes the best possible action in a situation."*	Russell, S. J. and Norvig, P. (2010). Artificial Intelligence: A Modern Approach, Third Edition. Prentice Hall.
MI8	*"Machines matching humans in general intelligence that is, possessing common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains."*	Bostrom, N. (2014). Superintelligence. Paths, Dangers, Strategy. Oxford University Press.
MI9	*"Machine Intelligence is the ability of an agent to provide rational, unbiased guidance and service to humans so as to help them achieve optimal outcomes in a range of circumstances."*	Lewis, C. W. P. and Monett, D. (2017). A Theory on Understanding Human Intelligence and a Persuasive Definition of Machine Intelligence for the Benefits of Humanity (working paper, unpublished).
HI1	*"Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment. It is global because it characterizes the individual's behavior as a whole; it is an aggregate because it is composed of elements or abilities which, though not entirely independent, are qualitatively differentiable."*	Wechsler, D. (1939). The measurement of adult intelligence (p. 3). Baltimore: Williams & Wilkins.
HI2	*"Intelligence is an individual's ability to respond to a given situation by anticipating the possible consequences of his actions."*	Bigge, M. L. (1976). Learning Theories for Teachers. Third Edition, London: Harper & Row Publishing.
HI3	Intelligence is an individual's *"entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature (problem solving skills) that [is] available at any one period of time."*	Humphreys, L. G. (1984). General Intelligence. In C. R. Reynolds and R. T. Brown (eds.), Perspectives on bias in mental testing (p. 243), Springer.
HI4	*"Intelligence is . . . a quality of behavior. Intelligent behavior is essentially adaptive, insofar as it represents effective ways of meeting the demands of a changing environment."*	Anastasi, A. (1986). Intelligence as a quality of behavior. In R. J. Sternberg and D. K. Detterman (eds.), What is intelligence?: Contemporary viewpoints on its nature and definition (pp. 19-21). Norwood, NJ: Ablex.
HI5	*"Intelligence is mental self-government. . . . The essence of intelligence is that it provides a means to govern ourselves so that our thoughts and actions are organized, coherent, and responsive to both our internally driven needs and to the needs of the environment."*	Sternberg, R. J. (1986). Intelligence is mental self-government. In R. J. Sternberg and D. K. Detterman (eds.), What is intelligence? Contemporary viewpoints on its nature and definition. Norwood, N.J: Ablex.
HI6	Intelligence is *"the ability to see relationships and to use this ability to solve problems."*	Fontana, D. (1988). Psychology for Teachers. Second Edition, London: Macmillan.
HI7	*"Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather it reflects a broader and deeper capability for comprehending our surroundings "catching on," "making sense" of things, or "figuring out" what to do."*	Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence, 24:13-23.
HI8	*"Intelligence is clearly a combination of the ability to "figure things out on the spot" and the ability to retain and repeat things that have been figured out in the past."*	Deary, I. J., Penke, L., and Johnson, W. (March 2010). The neuroscience of human intelligence differences. Nature Reviews, Neuroscience, 11:201-211.
HI9	To think and behave *"rationally means taking the appropriate action given one's goals and beliefs and holding beliefs that are commensurate with available evidence."* Intelligence is thus: *"optimal behavior in the domain of practical affairs. The optimization of the individual's goal fulfilment."*	Stanovich, K. E. (February 2014). Assessing Cognitive Abilities: Intelligence and More. Journal of Intelligence. 2(1):8-11.

and the absolute values (i.e. also counting all ratings given from the first response on) improved considerably. The only definitions for which the percentage values worsened were the original first two definitions from the fixed list, *MI1* and *MI2*. The percentages of negative agreement also changed: again, the same seven definitions of machine intelligence benefited from the shuffles and received, on average, fewer negative ratings in the last 336 responses. The impact of these changes was less evident for the negative agreement as for the positive agreement, however. This suggested that the gains in positive agreement after reordering were mainly from potentially undecided people. A closer look at the variation in the percentages of neutral selections seems to confirm this: these responses also changed, and to a greater extent than the percentages of negative agreement.

With regard to the definitions of *human* intelligence, the trends were similar: the percentages of responses showing positive, negative, and neutral agreement with the definitions of human intelligence before and after shuffling show the same trends as for the definitions of machine intelligence analyzed above.

Overall, the definitions of machine and human intelligence that benefited the most were *MI3* and *HI7*, respectively. Both definitions were the most accepted definitions from the collection, especially *HI7*, the undisputed overall winner. The definitions that had a clear disadvantage with respect to the percentages of positive agreement were the first ones from their respective lists when the lists were fixed, since the shuffling markedly diminished their anchoring effect.

3.2 Other Cognitive Biases when Arguming about Intelligence

The corpus containing 4,041 opinions on the definitions of machine and human intelligence is now analyzed in more detail, with regard to how many comments were provided in relation to the level of agreement, whether respondents commented more or less when they disagreed, and how many comments were provided versus the level of agreement.

It was observed that respondents tended to comment more when justifying why they did not agree with the definitions of intelligence from the literature, and tended to comment less when justifying why they did agree. When comments were provided, the percentage of positive agreement of those responses was much lower than the percentage of negative agreement, i.e. for ratings of *"Strongly agree"* and *"Agree"* combined, the total number of comments provided was lower, at almost half of the total number for ratings of *"Strongly disagree"* and *"Disagree"* (see Figure 3 (a)). Furthermore, when people did not comment at all, the number of responses with positive agreement and no comment was more than double that of those with a negative rating and no comment (see Figure 3 (b)).

Corresponding hypothesis tests were carried out and the results show that there is a correlation between the number of comments and the level of agreement with both types of definitions of intelligence (see Figure 4).

One possible explanation for these relationships might again be the presence of cognitive biases. The results are consistent with research in argumentative theory: people reason proactively from the perspective of having to defend their opinions and the main function of reasoning is to produce arguments to convince others [Mercier and Sperber, 2011]. Furthermore, the reasoning used to produce arguments exhibits a strong *confirmation bias*.[3] n general, *"rejecting what we are told generally requires some justification"* [Mercier and Sperber, 2011].

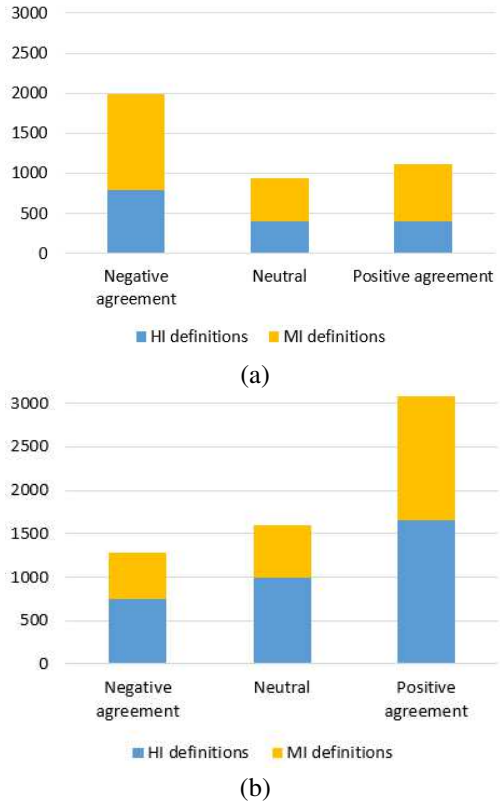

Figure 3: Level of agreement with the definitions of human and machine intelligence for responses (a) with and (b) without comments to justify the selection.

Moreover, when people disagree with the conclusion of an argument, they often spend more time evaluating it, as Mercier and Sperber [2011] show in their work on human reasoning. These authors also point out that polarization increases with the time spent thinking about an item. This is again the case for the comments provided by respondents to the survey on definitions of intelligence: the disagreement increased with time.

With regard to the smaller numbers of comments justifying a positive agreement or even no comments at all, the results were also consistent with research in argumentative theory: accepting what we are told generally does not require justification, because *"[a] good argument is an argument that is not refuted"* [Mercier and Sperber, 2011].

[3]Nickerson [1998] defines confirmation bias as *"[s]eeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand."*

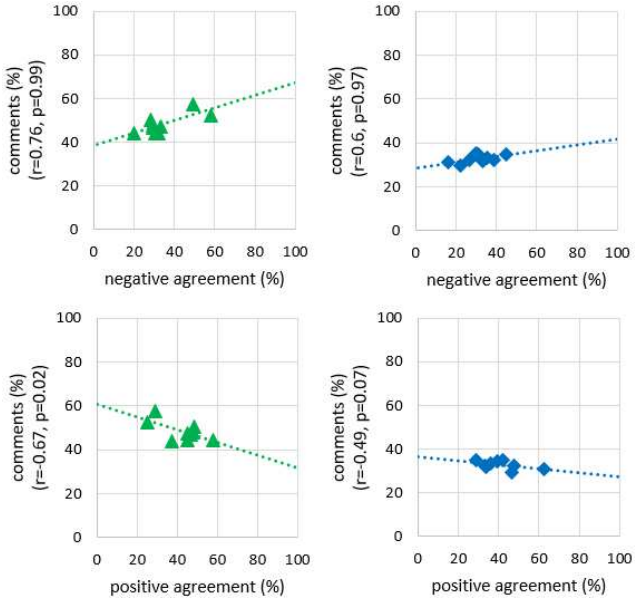

Figure 4: Scatter plots of different hypothesis tests for opinions on definitions of machine (plots on the left) and human intelligence (plots on the right).

3.3 Focalism (Again)

We analyzed not only the number of comments provided to justify the level of agreement but also the most often commented definitions of the survey, to explore why people argued more about those than about other definitions.

The most commented definition of intelligence was Russell and Norvig's; this was a definition from their well-known book *Artificial Intelligence: A Modern Approach*, which is used in more than 1,300 universities in over 110 countries worldwide.[4] This definition, *MI7*, received a total of 320 comments (57.6% of respondents commented) and was the second least accepted definition in the survey, receiving only 29.1% positive agreement. The second most commented definition was the least accepted definition of machine intelligence.

To explore why these definitions were the most commented but the least accepted, we took a closer look at their structure and both the terminology and language they used to determine whether some explanation might be possible. Russell and Norvig's definition, for example, is short and for this reason it may be missing important aspects when defining intelligence. However, this is not expected to be a reason for commenting more, since other definitions from the list were even shorter.

Nevertheless, there were arguments that included the words *"rational"* and *"best"* in at least 129 (40.3%) and 122 (38.1%) comments, respectively, out of all those provided for Russell and Norvig's definition. Four other definitions from the collection also used the words *rational*, *rationality*, or *rationally* in their texts but received many fewer comments. These are concepts that have received much attention when defining intelligence, since humans sometimes make irrational decisions that may not seem intelligent [Stanovich, 2015], and therefore the reason for the polarization over Russell and Norvig's was not obvious.

A possible explanation might be the presence of a cognitive bias called *focalism*,[5] also known as the *focusing effect* or *focusing illusion* [Kahneman *et al.*, 2006], which is the tendency to place too much importance on one aspect of an event. It may be that the respondents tended to place too much importance on the word *"rational,"* overlooking the word *"mainly"* (intelligence is concerned *mainly*, but not exclusively with rational action), and on the words *"best possible action"* while overlooking *"ideally"* (*ideally*, but not in every situation or always).

Another possible explanation might be the presence of other cognitive biases. For example, respondents may have been reflecting less on the definitions they were evaluating than on how to defend their opinion [Mercier and Sperber, 2011], which had already been expressed in terms of a negative level of agreement with a definition before they started describing why.[6] This is known as *attitude polarization*. Alternatively, it could also be associated with *bolstering* [McGuire, 1964], which is a bias arising from the pressure to justify an opinion rather than moving away from it, because the respondent has already stated before what his or her opinion is. This and other possible biases that might be present are considered in more detail in [Mercier and Sperber, 2011].

4 Automated Search for Cognitive Biases

Understanding natural language is one of the oldest research topics in the field of AI. Giving machines the ability to process and analyze information by looking at its meaning is not only considered a very difficult task, but has also attracted broad commercial attention in recent years, in terms of both investments and applications. However, although there exist a myriad of algorithms and tools that analyze the different semantic aspects of written speech, there is still no automated (or semi-automated) tool that can detect the cognitive biases present in natural language. This is a much more complex task that will require a human component for the foreseeable future.

One example of the tools that use machine learning algorithms to analyze written speech is the Perspective API created by Google and its subsidiary Jigsaw,[7] which was released in September 2017. It identifies *"whether a comment could be perceived as 'toxic' to a discussion"* and scores comments accordingly by assigning a toxicity score. Google defines *toxic* as *"a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."* The creators of the Perspective API do not recommend its use in the automated

[4] As claimed by the authors on the website http://aima.cs.berkeley.edu/ (Last accessed: July 11, 2019).

[5] A cognitive bias studied in Social Psychology which is a type of *anchoring*.

[6] Respondents were asked first to rate their level of agreement and then to justify why.

[7] See https://www.perspectiveapi.com for more (Last accessed: July 11, 2019).

moderation of conversations but as an assistant to humans in their work.

We used the Perspective API to analyze the corpus of experts' opinions on the definitions of intelligence in order to develop guidelines on how the AI community could contribute with constructive and objective feedback when discussing intelligence. This is one of our long-term goals.

The experts' comments justifying their level of agreement with the definitions of *machine* intelligence received an average toxicity score of 9.6%, which is lower than the average score obtained for *human* intelligence (10.3%). The highest toxicity values were assigned to single opinions commenting on machine intelligence, however. As analyzed above, the definitions of machine intelligence were more polarized and received many more comments although they were "less toxic" in general.

The results are not satisfactory, however; it is questionable as to how we can rely, even partially, on the use of automated tools. Comments such as *"Intelligence not originating from a human being"* were rated by the Perspective API with a toxicity level of 46%, for instance. Much work remains to be done in this respect. This is why we advise against using automated tools for the detection of cognitive biases or semantic information in written natural language; their current state of development is still strongly dependent on narrow domains, and needs much improvement.

5 Conclusions

Cognitive biases form part of people's judgment and cannot be always avoided. They affect how humans reason about and interpret not only concepts and phenomena but also other humans' opinions. There is an extensive body of research on cognitive biases, mainly in Psychology and other related fields. We show that they are also present when definitions are judged; especially, definitions of intelligence.

As Kelley [2014] has suggested in his work on Logic and Critical Thinking, *"it is not a good idea to include controversial information in a definition."* If a definition can be thought of as a neutral framework for providing a common understanding of the concept that is defined, then defining this well is crucial for interpretations of the concept by all parties, even opposing ones, and for reaching consensus on what is defined. However, even definitions that exclude controversial information are not exempt from biased judgment.

We also show that, although most cognitive biases cannot be kept away from human reasoning and evaluations, shuffling the definitions (of intelligence, but this conclusion could also be extended to other concepts) not only helps to counteract an anchoring effect that might arise but also means that people tend to be less unsure about making a decision when this happens. Furthermore, they take sides more often, at least on average and when rating definitions of machine and human intelligence from the literature.

The results presented in this paper could inform not only AI researchers and practitioners but also marketers and developers, for example when they present products or solutions to problems based on intelligent algorithms to users: what matters is not only the vocabulary that is used to describe "how intelligent" these artifacts are but also the ordering of the information that is presented. Similarly, other implications of the same kind may be expected in situations where people are asked to evaluate solutions, concepts, items, topics, etc. derived from or related to intelligent systems.

In general, when seeking feedback from users (including experts) about the definitions of intelligence already published in the scientific literature (and this can be generalized to systems, products, and other aspects that are judged), we should not expect users to provide their opinions when they agree with what is presented but rather to do so after a negative impression or discordance with the item that is being evaluated. The results presented here for definitions of intelligence are also consistent with findings in other areas [Walz and Ganguly, 2015].

Cognitive biases undermine an understanding of intelligence, and are a product of human subjective reasoning that in most cases cannot be avoided. However, knowing that cognitive biases are present in experts' opinions is a first step in helping to improve the definition of intelligence. In our opinion, it is very important to make all stakeholders aware of the cognitive biases that might be present when they define intelligence, in particular, or interact with or develop intelligent systems, in general, because this could also have an impact on the way human reasoning is modeled or automated. This is why we believe that each of the pressing rationale for a consensus definition of machine intelligence we discussed at the beginning of this paper are not more than a supporting statement of the need for understanding. Mercier [1912] in his empirical work on Logic stated more than one hundred years ago that *"[j]ust as not everything can be demonstrated, so not everything can be defined."* Our thesis is that if intelligence can be defined better, then this may also contribute to understanding it well.

References

[Bimfort, 1958] M. T. Bimfort. A definition of intelligence. *Center for the Study of Intelligence*, 2:75–78, 1958.

[Brooks, 1991] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[Chouldechova and Roth, 2018] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. arXiv e-prints, arXiv:1810.08810, 2018.

[Committee on AI, 2018] Select Committee on AI. AI in the UK: ready, willing and able? Report of Session 201719, HL Paper 100, The Authority of the House of Lords, UK, 2018.

[Daar and Greenwood, 2007] A.S. Daar and H.L. Greenwood. A proposed definition of regenerative medicine. *Journal of Tissue Engineering and Regenerative Medicine*, 1:179–184, 2007.

[Dafoe, 2018] A. Dafoe. AI governance: A Research Agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford. https://bit.ly/2PAb5Gp, 2018.

[Darwiche, 2018] A. Darwiche. Human-level intelligence or animal-Like abilities? *Communications of the ACM*, 61:56–67, 2018.

[Gasser and Almeida, 2017] U. Gasser and V. A. F. Almeida. A layered model for AI governance. *IEEE Internet Computing*, 21:58–62, 2017.

[Gottfredson, 1997] L. S. Gottfredson. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24:13–23, 1997.

[Hunt and Jaeggi, 2013] E. Hunt and S. M. Jaeggi. Challenges for research on intelligence. *Journal of Intelligence*, 1:36–54, 2013.

[Jordan, 2018] M. Jordan. Artificial intelligenceŁ–The Revolution Hasn't Happened Yet. Medium. https://bit.ly/2HzhHEC, 2018.

[Kahneman *et al.*, 2006] D. Kahneman, A. B. Krueger, D. Schkade, N. Schwarz, and A. A. Stone. Would you be happier if you were richer? A focusing illusion. *Science*, 312:19081910, 2006.

[Kaufman, 2019] S. Kaufman. The neuroscience of creativity: A Q&A with Anna Abraham: The latest state of the field of the neuroscience of creativity. Beautiful Minds, Scientific American. http://bit.ly/2T51AAW, 2019.

[Kelley, 2014] D. Kelley. *The Art of Reasoning: An Introduction to Logic and Critical Thinking*. W. W. Norton & Company, New York, NY, fourth edition, 2014.

[Kugel, 2002] P. Kugel. Computing Machines Can't Be Intelligent (...and Turing Said So). *Minds and Machines*, 12:563–579, 2002.

[Laqueur, 1985] W. Laqueur. *A World of Secrets: The Uses and Limits of Intelligence*. Basic Books, New York, NY, 1985.

[Lea, 2015] G. Lea. Why we need a legal definition of artificial intelligence. World Economic Forum. https://bit.ly/2Qcv7Gy, 2015.

[Leetaru, 2018] K. Leetaru. Does AI Truly Learn And Why We Need to Stop Overhyping Deep Learning. Forbes. https://bit.ly/2ApBBgl, 2018.

[Lewis and Monett, 2018] C. W. P. Lewis and D. Monett. Text analysis of unstructured data on definitions of intelligence. In *Proceedings of The 2018 Meeting of the International Association for Computing and Philosophy, IACAP 2018*, Warsaw, Poland, 2018.

[Lipton, 2018a] Z. C. Lipton. From AI to ML to AI: On Swirling Nomenclature & Slurried Thought. Approximately Correct: Technical and Social Perspectives on Machine Learning. https://bit.ly/2QWDs6t, 2018.

[Lipton, 2018b] Z. C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue Machine Learning*, 16:1–27, 2018.

[Luckin *et al.*, 2016] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier. *Intelligence Unleashed: An argument for AI in Education*. Pearson, London, 2016.

[Martínez-Plumed *et al.*, 2018] F. Martínez-Plumed, B. S. Loe, P. A. Flach, S. Ó hÉigeartaigh, K. Vold, and J. Hernández-Orallo. The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 5180–5187, Stockholm, Sweden, 2018.

[McCarthy, 2007] J. McCarthy. What is artificial intelligence? Basic Questions. Computer Science Department, Stanford University. https://stanford.io/2lSo373, 2007.

[McGuire, 1964] W. J. McGuire. Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, 1:191–229, 1964.

[Mercier and Sperber, 2011] H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34:57–111, 2011.

[Mercier, 1912] C. Mercier. *Elements of Logic*. The Manhattanville Press, New York, NY, third edition, 1912.

[Monett and Lewis, 2018] D. Monett and C. W. P. Lewis. Getting clarity by defining artificial intelligence–A Survey. In V. C. Müller, editor, *Philosophy and Theory of Artificial Intelligence 2017*, volume SAPERE 44, pages 212–214. Springer, Berlin, 2018.

[Nemitz, 2018] P. Nemitz. Constitutional democracy and technology in the age of artificial intelligence. *Phil. Trans. R. Soc. A*, 376:20180089, 2018.

[Nickerson, 1998] R. S. Nickerson. Confirmation bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2:175–220, 1998.

[Sloane, 2018] M. Sloane. Making artificial intelligence socially just: why the current focus on ethics is not enough. The London School of Economics and Political Sciences. https://bit.ly/2NF2amR, 2018.

[Stanovich, 2015] K. E. Stanovich. Rational and irrational thought: The Thinking That IQ Tests Miss. Why smart people sometimes do dumb things. *Scientific American Mind*, 23, 2015.

[Tversky and Kahneman, 1974] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science, New Series*, 185:1124–1131, 1974.

[Verma and Rubin, 2018] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the 2018 ACM/IEEE International Workshop on Software Fairness, FairWare 2018*, pages 1–7, Gothenburg, Sweden, 2018.

[Walz and Ganguly, 2015] A. Walz and R. Ganguly. Apptentive 2015 consumer survey: The Mobile Marketer's Guide to App Store Ratings & Reviews. Apptentive, 2015.