


# The Winograd Schema Challenge: Are You Sure That We Are on the Right Track?

Nicos Isaak<sup>1,2</sup> 

<sup>1</sup>Open University of Cyprus, Cyprus

<sup>2</sup>Computational Cognition Lab, Cyprus

**Keywords:** Winograd Schema Challenge, WSC Framework, Knowledge Representation and Reasoning.

**Abstract:** In the past few years, the Winograd Schema Challenge (WSC), the task of resolving ambiguities in carefully-structured sentences, has received considerable interest. According to Levesque, what matters when it comes to the WSC is not a good semblance of intelligent behavior but the behavior itself. In this regard, the WSC has been proposed to understand human behavior as a challenge that could lead to the endowment of machines with commonsense reasoning abilities. Here, we argue that most systems developed so far have typically been designed and evaluated without considering the challenge's purpose, emphasizing the semblance of intelligence rather than understanding human behavior itself. At the same time, we present an overview of systems developed so far along with a novel developmental-evaluation framework (WSC-Framework 01). The WSC-Framework offers guidelines on what we might need to do to move the field towards the endowment of machines with commonsense reasoning to tackle Winograd schemas the way humans do.


## 1 INTRODUCTION

The Winograd Schema Challenge (WSC), a novel litmus test for machine intelligence, has been proposed as an alternative to the well-known Turing Test. Unlike the Turing Test, which is based on short free-form conversations where a machine attempts to imitate humans, machines passing the WSC are expected to demonstrate the ability to think without having to pretend to be somebody else (Levesque et al., 2012).

Although over the last years the AI community has made progress regarding the tackle of specific Winograd schemas, most systems developed so far have typically been designed and evaluated without considering the challenge's purpose, emphasizing the semblance of intelligence rather than understanding human behavior itself (Kocijan et al., 2020). According to Levesque (2014), *what matters when it comes to the science of AI is not a good semblance of intelligent behavior at all, but the behavior itself, what it depends on, and how it can be achieved*. However, it seems that the technology of AI gets all the attention, meaning that most of the research uses techniques that have little to do with what we intuitively imagine human intelligence to be. We believe that there is a dif-

ferent point of view where we should focus on that views the brain as processing information, not strictly as patterns of words or data (Levesque, 2014). This has to do with the problem that the science of AI has faced in the last decades, according to which we still do not have a system that can read a news story and tell you who did what to whom, when, where, and why (Marcus and Davis, 2019). Put simply, we do not know how to transfer humans commonsense reasoning ability to machines, the kind of knowledge that we take for granted and expect ordinary people to possess.

Regarding the WSC, the current state of affairs might point to the need for a novel framework to move the field towards the endowment of machines with commonsense reasoning to tackle Winograd schemas the way humans do. According to Kinzler and Spelke (2007), humans are endowed with a small number of systems that stand at the foundation of all our beliefs and values and that new skills and knowledge build on these foundations. In this sense, these separable systems could be potentially developed using various tools and techniques from good-old classical and modern AI. In this regard, here we discuss usability issues that should be considered and addressed in designing systems that aim to tackle the WSC. The general idea is to handle people's behavior on Wino-

<sup>a</sup>  <https://orcid.org/0000-0003-2353-2192>

grad questions as a natural phenomenon to be explained—not as another competition that we need to tackle just for the sake of the tackle. As Levesque mentioned, *even a single example can tell us something important about how people behave, however insignificant statistically*. Reasoning about how these different kinds of actions interrelate while answering WSC questions, we propose a novel but straightforward WSC-Framework for the design and evaluation of future developed WSC systems.

## 2 THE CHALLENGE

The Winograd Schema Challenge (WSC) was proposed in 2012 (Levesque et al., 2012) to serve as the means to understand human behavior. The WSC requires resolving pronouns in schemas where shallow parsing techniques seem not to be directly applicable, where the use of world knowledge and the ability to reason seem necessary. Winograd schemas consist of pairs of halves, with each half consisting of a sentence, a question, two possible pronoun targets (answers), and the correct pronoun target (Levesque et al., 2012).

Given just one of the halves, the aim is to resolve the definite pronoun through the question to one of its two co-referents. To avoid trivializing the task, the co-referents are of the same gender, and both are singular or plural. The two halves differ in a special word or small phrase that critically determines the correct pronoun target—e.g., “Sentence: The city councilmen refused the demonstrators a permit because they **feared/advocate** violence. Question: Who **feared/advocate** violence? Pronoun Targets: The city councilmen, The demonstrators”.

It is believed that the WSC can provide a more meaningful measure of machine intelligence when compared to the Turing Test, exactly because of the presumed necessity of reasoning with commonsense knowledge to identify how the special word or phrase affects the resolution of the pronoun. As expected from its reliance on commonsense knowledge, English-speaking adults have no difficulty with the challenge, meaning they can easily pass it with an average score of 92% (Bender, 2015; Isaak and Michael, 2021b). In this regard, the WSC should be viewed as a task to examine basic forms of intelligent behavior, meaning answering certain ad-hoc questions posed in English by paying attention to the behavior itself, what it depends on, and how it can be achieved. Therefore, schemas can be designed to test for problem-solving skills, for an ability to visualize, or to reason (Levesque, 2014).

## 3 WSC APPROACHES

Developing a system that understands human behavior while tackling Winograd schemas takes more than just doing well in subsets of schemas. Although great solutions have been developed, we cannot get to the moon by climbing taller trees successively without paying attention to the unfolding human mechanisms while answering Winograd schemas. Here, we analyze the systems that were developed to tackle the challenge, which we divide into three categories: i) systems based on machine learning, ii) systems that use some kind of commonsense reasoning, and iii) theories that did not yet become actions. Given that not all systems are tested on the same subsets of schemas, the reported results should be taken with a grain of salt. Reportedly, machine learning approaches tackle schemas with an average score of 93.1% (Sakaguchi et al., 2020), whereas commonsense reasoning approaches with an average score of 70% (Sharma et al., 2015).

**Machine Learning Approaches.** Rahman and Ng (2012) system utilizes lexicalized statistical techniques to tackle schemas with an average score of 73.05%. The system finds the most probable pronoun candidate through a ranking-based approach (SVM) that combines the features derived from different resources (e.g., Web Queries).

Peng et al. (2015) system uses an Integer Linear Programming approach to acquire statistics in an unsupervised way from multiple knowledge resources (e.g., Gigaword corpus). By training a co-reference model, it achieves a score of 76%.

Emami et al. (2018), developed a Web knowledge-hunting system, which was able to tackle schemas with an average score of 57%. The system develops a set of queries to capture the predicates of each examined half and sends them to a search engine to retrieve relevant snippets.

Within a deep learning approach, Wang et al. (2019) tackle schemas with an average score of 78%. Their approach is based on the Deep Structured Similarity Model (DSSM) framework, which models semantic similarity between two texts of strings by utilizing schemas as a pairwise ranking problem.

Kocijan et al. (2019) showed that a significant improvement for tackling the WSC could be achieved by fine-tuning a pre-trained masked BERT language model. BERT, which stands for Bidirectional Encoder Representations from Transformers, randomly masks words in a particular context and predicts them. The model was able to tackle schemas with an average score of 74.7%.

Sakaguchi et al. (2020) developed a large dataset of WSC-like examples by employing crowdsourced workers. Through a fine-tuned RoBERTa language model, they gained contextualized embeddings for each example by handling it like a fill-in-the-gap problem. The authors report results on tackling schemas with an average score of 93%.

**Commonsense & Reasoning Approaches.** Sharma et al. (2015) developed a system based on Answer Set Programming (ASP). For each examined schema, the system retrieves the background knowledge directly from *Google* through fixed queries and achieves a 70% score of prediction.

Isaak and Michael (2016) developed a system that utilizes logical inferences to tackle schemas with an average score of 65.6%. The system uses the Web-sense engine (Michael, 2013), which responds to user queries with implied inferences according to the collective human knowledge found on the Web.

**Theoretical Approaches.** Bailey et al. (2015) examined two types of rhetorical relations that can be used to establish discourse coherence. Their work introduced a way for reasoning about correlation and how this could be used to justify solutions to some Winograd Schema problems. Their approach relies on the availability of axioms expressing relevant commonsense knowledge, which were manually tailored to handle specific Winograd schemas. Although this work seems to be on the right track, to the best of our knowledge, the authors did not provide evidence that their theoretical approach could be turned into a working system.

## 4 THE WSC-Framework 1.0

Every single Winograd instance is a natural phenomenon to be explained, meaning that to drive the field forward, we should focus on human behavior. In this section, we present a novel Developmental-Evaluation Framework for the WSC. The aim is twofold: The first is to guide in the design of systems that *show their work* while tackling Winograd instances. The second is to shed light on human curators' evaluation process of each system's capabilities. To the best of our knowledge, this is the first time anyone has tried to outline such a framework for the WSC.

The idea behind it is based on five observations from the literature: 1.) We know that humans blend multiple information sources in reasoning about future outcomes (Téglás et al., 2011) 2.) Show their work is not something that many systems can eas-

ily do (Marcus and Davis, 2019; Mitchell, 2019) 3.) We must combine various developed approaches into building hybrid systems that will use the best of various techniques in ways we have yet to discover (Marcus and Davis, 2019) 4.) The upshot of a complete commonsense reasoner would be surpassing for the AI community, albeit this payoff may only be recognized once a significant part of the outcome has been developed (Davis and Marcus, 2015) 5.) Currently, the best evaluators for Winograd instances are human curators, albeit an interaction with machines could amplify human and machine intelligence by combining their complementary strengths (Isaak and Michael, 2019).

### 4.1 The Developmental Part

The Developmental part of the WSC-Framework provides guidelines for designing systems that tackle Winograd instances using inferences according to the collective human knowledge. For its design, we consider how the human mind is organized where, along with existing AI approaches, it offers a guideline on building systems that focus on the unfolding human mechanisms while tackling Winograd instances.

The Developmental part, divided into five models, can be used as the basis for understanding any drawn inference for the correct resolving of any Winograd instance, as all of the results should and would be examined by human curators (see WSC-Framework 1.0: The Developmental Part in Figure 1). Given that this is the first and only framework introduced for the WSC, it does not purport to eliminate any other potentially valuable approach but to guideline the development of systems that focus on human commonsense reasoning abilities while tackling Winograd instances. In this regard, future updates to the framework might include other creative or genuinely innovative solutions that might help us tackle the WSC.

**The Core Model.** The first model relates to the human mind. According to Kinzler and Spelke (2007), it is believed that humans are endowed with four distinguishable systems/cores (objects, agents, numbers, geometry) that are responsible for our values, including the acquisition of language. Ostensibly, research shows that new concepts are built on these cores. The object system refers to principles that help predict when objects move and where/when they stop. The agent system persists over human development and is defined by goal-directed dynamic actions between other agents. Similarly, the number system refers to the representation of numbers throughout human development. The last core system refers to the geometry of space (e.g., distance, angle, sense relations)

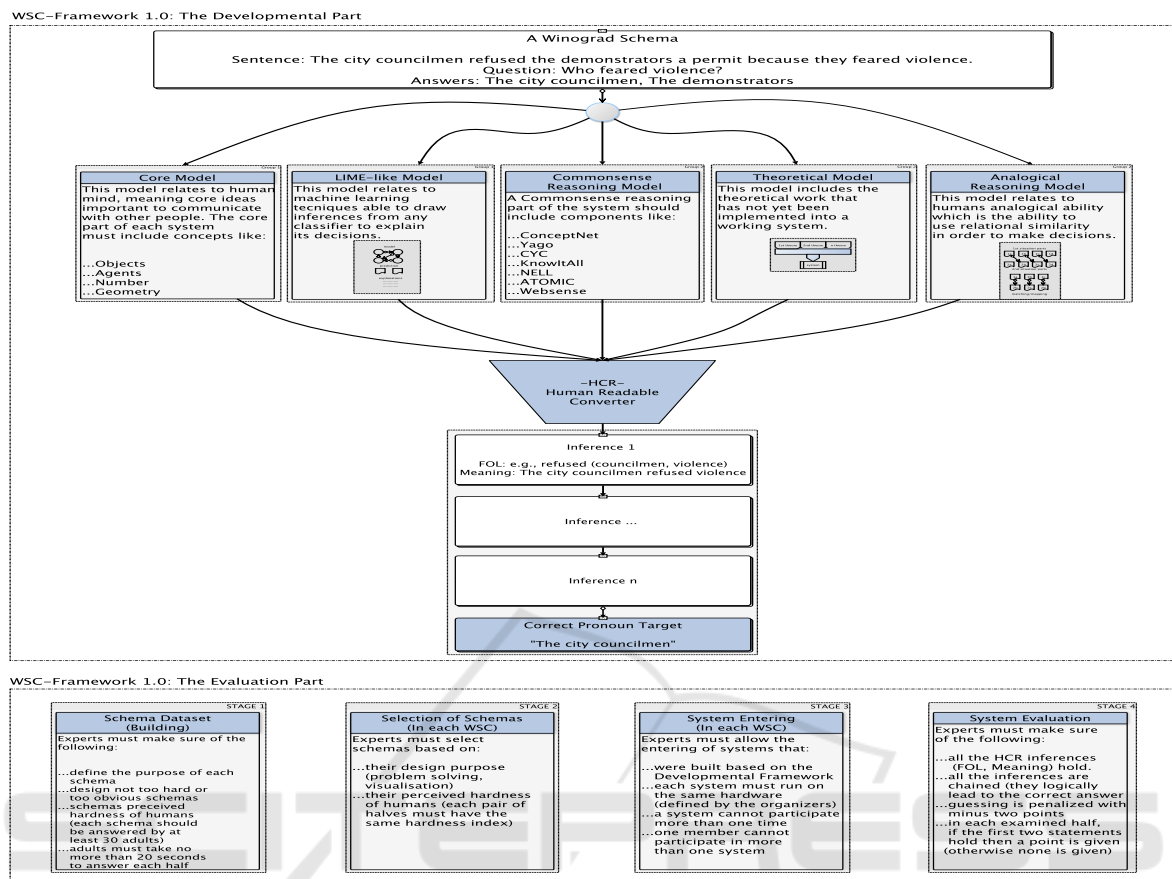


Figure 1: The WSC-Framework 1.0: A Developmental-Evaluation Framework for the WSC.

where human adults use geometric information in pictures. Given that these systems are essential for communicating with other people in our day-to-day lives, researchers need to distinguish what core representations must be utilized while tackling Winograd instances.

**The LIME-like Model.** Humans can make judgments based on the frequencies of previously experienced situations, also called statistical decisions (Téglás et al., 2011). For instance, we know that techniques like deep learning excel in perceptual classification. It would be of great interest to take advantage of deep learning and combine it with an inference mechanism that can be easily evaluated. According to Marcus and Davis (2019), some parts of the mind might work like deep learning, and some other parts seem to work at a much higher level of abstraction. To combine machine learning techniques with the ability to draw inferences, in our framework, we utilized the LIME model (Ribeiro et al., 2016), which can interpret the results of classifiers. For instance, if a model predicts that in the sentence, “The cat caught the mouse because it is clever”, the definite pronoun *it*

refers to *cat*, then the LIME model could show what led to that prediction by emphasizing the words that led to the conclusion. Drawing inferences from any classifier would arguably help human curators trust the model’s prediction.

**The Commonsense-reasoning Model.** Previous studies have demonstrated that a complete commonsense reasoner would be very important for the AI community (Davis and Marcus, 2015). Knowledge-based systems are systems designed to solve problems by simulating the capabilities of humans. Given that current machines do not purport a wide variety of commonsense reasoning, here, we propose a combination of different techniques from the literature in identifying ways to endow machines with the necessary knowledge. The overall goal is to identify an interpretable model to explain the inferences drawn while tackling Winograd instances. In this regard, any developed WSC system should explicitly define which of the following methods are used:

- 1.) ConceptNet<sup>1</sup>: Projects like the ConceptNet

<sup>1</sup><https://conceptnet5.media.mit.edu>



semantic network provide a broad set of background knowledge that contains facts connected based on different kinds of relations (e.g., relatedTo). Concept-Net knowledge is collected across multiple sources like games with a purpose, expert-created, and crowdsourcing resources. 2.) Yago<sup>2</sup>: Yago is a human-verified project built from Wikipedia, WordNet, and GeoNames. It covers facts for geographical entities, personalities of public life or history, movies, and organizations. Following the RDF model, it builds its knowledge base with facts represented as triples showing relations between subject, predicates, and objects. 3.) Cyc<sup>3</sup>: The Cyc project seems to be one of the largest experiments in symbolic AI. The intention has been to facilitate the creation of AI applications that may perform human-like reasoning, and its facts are mostly taxonomic. Its initial purpose was to make AI programs flexible by providing access to a substantial base of implicit background knowledge. 4.) KnowItAll<sup>4</sup>: KnowItAll automates the process of extracting facts, concepts, and relationships from the WWW, storing at the same time its information in an RDBMS form. Its main component is the *extractor* which outputs/inference categories by reading lists of texts, where every extraction is assessed with a probability. 5.) NELL<sup>5</sup>: NELL has been a project learning to accumulate data from the Web since January 2010. These include taxonomic relations and facts with confidence-weighted beliefs (e.g., served-With (tea, biscuits)). 6.) Atomic<sup>6</sup>: Atomic is a commonsense reasoning atlas that focuses on inferential knowledge for everyday events, causes, and effects (*if then* knowledge). For instance, given a previously unseen event, it can learn to perform *If-Then* commonsense inferences. 7.) Websense: Websense is an engine that can respond to user queries, with logical inferences that are implied from human knowledge found across the Web (Michael, 2013). The engine accumulates its data from Wikipedia autonomously via a crawler. For instance, given a query saying that a “man robbed a bank”, the engine, through semantic scenes between subjects, objects, and verbs, returns inferences in the form of natural language text.

**The Theoretical-Grounding Model.** Many theoretical papers have been published, though few of them will be utilized to develop actual systems (Davis and

Marcus, 2015). A large body of theoretical work does not necessarily lead to convincing potential applications, meaning that the upcoming target is a published paper rather than an implemented program (Davis and Marcus, 2015). Within this model, researchers can define if their approach results from a theoretical paper and how.

**The Analogical-Reasoning Model.** Adults can make rational judgments based on frequencies of previously experienced events. According to Mitchell (2019), how humans connect concepts between similar ideas is crucial towards achieving a human-like AI, albeit not much attention has been paid to it. In this sense, once we pay attention to analogies’ crucial role in cognition, we might unlock human-like cognition in machines. We can think of analogies as the process of drawing inferences between similar situations where structured representations are connected with first-order-like statements. In the case of similar situations, at first, we find out the possible matches, then the matches are combined into consistent clusters, and finally, we conclude the overall result/mapping (Gentner and Colhoun, 2010). In this regard, any developed system should answer if it uses analogical reasoning to resolve Winograd instances.

**Human Readable Converter (HRC).** It is essential to combine the strengths of the various approaches presented above to help researchers in their designing process but also human curators in their evaluation process. Our proposed framework aims to bring together all the above to a single inference mechanism for the tackle of Winograd instances. Given that most of the approaches use unique knowledge representation mechanisms, below we propose an inference converter (called Human Readable Converter) that would act as a unified proxy to show all of the implied mechanisms needed to answer Winograd instances. Pieces of knowledge might not be explicitly stated but be implicitly encoded across the above approaches. Hence, the HRC component gathers all the knowledge and outputs inferences depicted in a logical human-readable form. Specifically, the HRC component outputs the inferences in two forms, i) first-order semantic scenes (logical *formulae*), similar to Prolog rules, and ii) simple English sentences, with no more than ten words. To illustrate, for the Winograd instance, *Sentence: The city councilmen refused the demonstrators a permit because they feared violence. Question: Who feared violence?, Answers: The city councilmen, The demonstrators* the results could be something like, 1.) *refused (city-councilmen, violence) → The city councilmen refused violence.* 2.) *fear (people, violence) → People fear violence.*

<sup>2</sup><https://yago-knowledge.org>

<sup>3</sup><https://cyc.com/platform/>

<sup>4</sup><http://projectsweb.cs.washington.edu/research/knowitall/>

<sup>5</sup><http://rtw.ml.cmu.edu/rtw/>

<sup>6</sup><https://homes.cs.washington.edu/~msap/atomic/>

3.) *refuse (people, violence) → People refuse violence.* 4.) *fear (city-councilmen, violence) → The city councilmen fear violence.* Answer: *The city councilmen.*

## 4.2 The Evaluation Part

The automatic evaluation of generated language is still an open research question, meaning that human evaluation is still considered the gold standard (Isaak and Michael, 2019, 2021a). Furthermore, given the problems of the automatic evaluation of systems like GLUE and SuperGLUE, it seems that the empirical evaluation by human curators cannot be avoided. In this regard, the Evaluation part of the WSC-Framework provides guidelines to human curators for organizing and evaluating upcoming Winograd challenges. It consists of four parts, relating to the dataset building of schemas, the schema selection process, the entering of systems, and finally, the evaluation of the results (see WSC-Framework 1.0: The Evaluation Part in Figure 1).

**Dataset Building.** Because of human curators, we believe that our proposed framework can offer guarantees on the soundness of its results. For that purpose, human evaluators should test every developed system with several new Winograd schemas designed by human experts in the field.

For each schema, experts should previously make sure of the following: 1.) Define the purpose of their design. As Levesque (2014) has argued, each schema should tell us something about human behavior, meaning it should serve a specific purpose. 2.) That their questions are not hard for people, nor their answers are obvious enough (Levesque et al., 2012). Especially their answers should not be resolvable with selectional restrictions or syntactically resolved just by parsing the sentence. 3.) Estimate their perceived hardness for humans. Work in the literature (Bender, 2015; Isaak and Michael, 2021b) has shown that not all humans tackle schemas with the same ease, meaning that there are schemas that are harder for humans to resolve. On another, it is only when they go wrong that machines remind us how powerful they are, meaning that human results should be comparable to machine results. Therefore, for human curators to access the schemas' hardness indexes (values in the range of 0 - 1), all of the testing schemas should be previously tested with human adults. Like with previous studies (Bender, 2015; Isaak and Michael, 2021b), each schema should be answered by at least 30 adult English native speakers —additionally, human adults should not need more than 20 seconds to answer each Winograd instance.

**Schema Selection.** The whole idea behind the challenge is to have pairs of halves (schemas) with slight differences. However, on the flip side of the coin, having pairs of halves with small differences might lead systems to guess the answers, which is not a demonstration of intelligent behavior —e.g., if you know the answer of the first half, you do not need to bother for the answer of the second half. To avoid this kind of behavior, we could use some randomly displayed instances with slightly different words in each schema. Regarding the final selection of schemas, this should be done based on their perceived hardness for humans. For example, the first and only WSC consisted of 60 schemas from which 38 of them were correctly resolved by nine people, one schema had both halves correctly solved by eight people, and 21 of them had the one half correctly solved by nine and the other half by eight people (Davis et al., 2016).

**Participants - Entering Systems.** Based on our framework, in every upcoming WSC, organizers should allow only entries developed based on the Developmental part. In this regard, the organizers should strictly define what models of the Developmental part are used.

In the first WSC (Morgenstern et al., 2016; Davis et al., 2017), six systems participated, though representing four different teams, meaning that one participant was allowed to enter three times with the same system. To avoid such problems, each team should be allowed to participate with one system, and each member should be allowed to participate in no more than one team. Furthermore, the challenge should occur in a physical place, where access to the Internet would not be allowed for safety reasons. Additionally, all systems should be built to run on a laptop or desktop computer provided by the testing committee. Therefore, all systems should be tested on the same hardware, clearly stated prior to the competition. Given that human participants need at least 18 seconds to answer a Winograd instance (Bender, 2015; Isaak and Michael, 2021b), and that in the first WSC, the longest it took for a system to answer was around 3.5 minutes/half, 3 minutes for each half should be fair enough.

**The Scoring Function.** Given that AI researchers are a competitive bunch, we proceed to the development of a scoring function that human curators will administer for evaluation purposes. In this regard, we have built the scoring function based on Levesque work (Levesque et al., 2012; Levesque, 2014), the first and only Winograd challenge (Davis et al., 2017), and the schema selection process mentioned above. The whole idea is based on building a scoring function

Table 1: Our Proposed Framework’s Scoring Function.

Half id: 01A			
The city councilmen refused the demonstrators a permit because they feared violence. Who feared violence? The city councilmen, The demonstrators			
city councilmen == x, demonstrators == y, violence, refuse, fear, people			
	FOL	Meaning	Curators
1/4	refused (x, violence)	The city councilmen refused violence	<input checked="" type="radio"/>
2/4	fear (people, violence)	People fear violence	<input checked="" type="radio"/>
3/4	refuse (people, violence)	People refuse violence	<input checked="" type="radio"/>
4/4	fear (x, violence)	The city councilmen fear violence.	<input checked="" type="radio"/>
Answer	The city councilmen		Score: 1.0

to motivate researchers to participate in the upcoming challenges. The goal is to validate the nuggets of knowledge used to draw inferences to answer Winograd instances. To that end, several human curators should examine every given answer ( $n = 3$ ). For compatibility reasons, each system should output a text file for every single Winograd instance, based on the Human Readable Converter (HRC) component. Then, each Winograd instance should be displayed on a computer screen, where the curators would evaluate the results (see Table 1).

We know that an inference is valid if it would be typically recognized as such by humans (Michael, 2013). Hence, for every answer of each examined system, curators should consider the drawn inferences, meaning the slightly modified output of the HRC component (see the *FOL* and *Meaning* columns in Table 1). In short, these are the premises that must hold for the correct resolving of a Winograd instance. For every valid transformation of *FOL* to *Meaning*, curators must check the relevant option button (see the left sub-column under the curator’s column in Table 1). Finally, if all of the option buttons are checked, the *chaining* option button is enabled (see the right sub-column under the curator’s column in Table 1). Following, if all the inferences are chained, meaning they logically lead to the correct answer of the Winograd instance at hand, the *chaining* option button is manually checked. If this is the case for most curators, the examined system takes one point. If the drawn inferences cannot lead to the correct answer, the score equals 0. Similarly, if a *FOL* formula does not make sense or a *Meaning* cannot be concluded, the examined system is penalized with minus two points. As stated by Levesque (2014), a WSC test involves asking a number of questions with a strong penalty for wrong answers to preclude guessing.

## 5 CONCLUSION

This paper is all about the WSC and the science of AI. We have shown that we still do not have a system that can read any Winograd instance and tell you about the unfolding human mechanisms behind it, about who did what to whom, when, where, and why. Following Levesque (2014) line of research, that even a single Winograd instance should tell us something important about how people behave, we proposed a novel but straightforward framework that covers both the building of systems and their evaluation by human experts. Additionally, we have provided the necessary insights into organizing future Winograd challenges. We have also provided guidelines for building and selecting the testing schemas according to their perceived hardness for humans, organizing participants in teams, and finally, how human curators can evaluate their results.

Regarding future work, and given that there are no silver bullets, a lot more remains to be done. About the framework itself, given that AI is a dynamic field, future updates might include other creative solutions that might bring us closer to understanding the unfolding human mechanisms while tackling Winograd instances. Considering the schema design difficulties even for experts, one could argue that the design of future challenges, even yearly, might be a difficult task. Regarding the schema design and selection part, of interest would be discovering automated ways to amplify human and machine intelligence without taking human experts out of the loop, which would help reduce the volume of experts’ work.

## REFERENCES

- Bailey, D., Harrison, A., Lierler, Y., Lifschitz, V., and Michael, J. (2015). The Winograd Schema Challenge and Reasoning about Correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

- Bender, D. (2015). Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*, pages 39–45.
- Davis, E. and Marcus, G. (2015). Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM*, 58(9):92–103.
- Davis, E., Morgenstern, L., and Ortiz, C. (2016). Human Tests of Materials for the Winograd Schema Challenge 2016.
- Davis, E., Morgenstern, L., and Ortiz, C. L. (2017). The First Winograd Schema Challenge at Ijcai-16. *AI Magazine*, 38(3):97–98.
- Emami, A., De La Cruz, N., Trischler, A., Suleman, K., and Cheung, J. C. K. (2018). A Knowledge Hunting Framework for Common Sense Reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1958, Brussels, Belgium. Association for Computational Linguistics.
- Gentner, D. and Colhoun, J. (2010). Analogical Processes in Human Thinking and Learning. In *Towards a theory of thinking*, pages 35–48. Springer.
- Isaak, N. and Michael, L. (2016). Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In Pearce, D. and Pinto, H. S., editors, *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press.
- Isaak, N. and Michael, L. (2019). WinoFlexi: A Crowdsourcing Platform for the Development of Winograd Schemas. In Liu, J. and Bailey, J., editors, *AI 2019: Advances in Artificial Intelligence*, pages 289–302, Cham. Springer International Publishing.
- Isaak, N. and Michael, L. (2021a). Blending NLP and Machine Learning for the Development of Winograd Schemas. In Rocha, A. P., Steels, L., and van den Herik, J., editors, *Agents and Artificial Intelligence*, pages 188–214, Cham. Springer International Publishing.
- Isaak, N. and Michael, L. (2021b). Experience and Prediction: A Metric of Hardness for a Novel Litmus Test. *Journal of Logic and Computation*. exab005.
- Kinzler, K. D. and Spelke, E. S. (2007). Core Systems in Human Cognition. *Progress in brain research*, 164:257–264.
- Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., and Lukasiewicz, T. (2019). A Surprisingly Robust Trick for Winograd Schema Challenge. *arXiv preprint arXiv:1905.06290*.
- Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G., and Morgenstern, L. (2020). A Review of Winograd Schema Challenge Datasets and Approaches.
- Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge. In *Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning*.
- Levesque, H. J. (2014). On Our Best behaviour. *Artificial Intelligence*, 212:27–35.
- Marcus, G. and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.
- Michael, L. (2013). Machines with Websense. In *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning*.
- Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.
- Morgenstern, L., Davis, E., and Ortiz, C. L. (2016). Planning, Executing, and Evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1):50–54.
- Peng, H., Khashabi, D., and Roth, D. (2015). Solving Hard Coreference Problems. *Urbana*, 51:61801.
- Rahman, A. and Ng, V. (2012). Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.
- Sharma, A., Vo, N. H., Aditya, S., and Baral, C. (2015). Towards Addressing the Winograd Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 25–31.
- Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J., and Bonatti, L. (2011). Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference. *Science (New York, N.Y.)*, 332:1054–9.
- Wang, S., Zhang, S., Shen, Y., Liu, X., Liu, J., Gao, J., and Jiang, J. (2019). Unsupervised Deep Structured Semantic Models for Commonsense Reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 882–891.