Scarecrows in Oz: The Use of Large Language Models in HRI

TOM WILLIAMS, Colorado School of Mines, USA
CYNTHIA MATUSZEK, University of Maryland, Baltimore County, USA
ROSS MEAD, Semio, Inc., USA
NICK DEPALMA, Plus One Robotics, USA

The proliferation of Large Language Models (LLMs) presents both a critical design challenge and a remarkable opportunity for the field of Human-Robot Interaction (HRI). While the direct deployment of LLMs on interactive robots may be unsuitable for reasons of ethics, safety, and control, LLMs might nevertheless provide a promising baseline technique for many elements of HRI. Specifically, in this position paper, we argue for the use of LLMs as *Scarecrows*: 'brainless,' straw-man black-box modules integrated into robot architectures for the purpose of quickly enabling full-pipeline solutions, much like the use of "Wizard of Oz" (WoZ) and other human-in-the-loop approaches. We explicitly acknowledge that these Scarecrows, rather than providing a satisfying or scientifically complete solution, incorporate a form of the wisdom of the crowd, and, in at least some cases, will ultimately need to be replaced or supplemented by a robust and theoretically motivated solution. We provide examples of how Scarecrows could be used in language-capable robot architectures as useful placeholders, and suggest initial reporting guidelines for authors, mirroring existing guidelines for the use and reporting of WoZ techniques.

CCS Concepts: • **Computer systems organization → Robotics**; • **Human-centered computing → Natural language interfaces**.

ACM Reference Format:
Tom Williams, Cynthia Matuszek, Ross Mead, and Nick DePalma. 2023. Scarecrows in Oz: The Use of Large Language Models in HRI. *ACM Trans. Hum.-Robot Interact.* 1, 1, Article 1 (January 2023), 11 pages. https://doi.org/10.1145/3606261

1 INTRODUCTION

Human-robot interaction, as a field, focuses on interactive, embodied agents that engage in back-and-forth exchanges with humans. One such form of interaction is language, which is fundamental to human communication, and a growing subfield of human-robot interaction involves language-capable robots. It is in this context that large language models have appeared, and it is natural that researchers in HRI have already begun to try to leverage those models to accomplish a variety of tasks, including modeling humans [3, 66], planning tasks based on language requests [4], commonsense reasoning [35], and parsing to LTL specifications [34, 44]. These uses are not limited to natural language dialog, but extend to a variety of capabilities, especially for models capable of taking in both linguistic and non-linguistic inputs [21].

Large pre-trained language models such as GPT address certain longstanding challenges in artificial intelligence—sentiment analysis, text generation, automatic program writing—in an exciting

Authors' addresses: Tom Williams, twilliams@mines.edu, Colorado School of Mines, Golden, Colorado, USA; Cynthia Matuszek, cmat@umbc.edu, University of Maryland, Baltimore County, Baltimore, Maryland, USA; Ross Mead, ross@semio.ai, Semio, Inc., Los Angeles, California, USA; Nick DePalma, ndepalma@alum.mit.edu, Plus One Robotics, Pittsburgh, Pennsylvania, USA.

2573-9522/2023/1-ART1
https://doi.org/10.1145/3606261

new way, and their fluency demonstrates the power of tackling those problems using nothing but a summarized model of extremely large corpora of text. But at the same time, they expose how many problems remain whose solutions need more than just models of word occurrences. These models are powerful, convincing, and impressive; but they are also flawed, hallucinatory, and incomplete. Given that these technologies are already becoming part of the landscape of problem-solving tools, it is worthwhile to consider both what they can do and where they fall short in the context of HRI.

Here, we consider the use of LLMs as components of a human-interactive systems. Large language models, or LLMs, allow HRI researchers to build novel capabilities into their systems, either as a deliberate choice to take advantage of the new capabilities they offer, or as a quick prototyping mechanism. In this work we highlight the parallels between the uses of LLMs as prototyping mechanisms and the widespread use of Wizard-of-Oz in the design and study of interactive robots. Although LLMs are opaque [10], inconsistent [22], and inappropriate for replication [49], the same criticisms often apply to humans, yet this has not prevented the use of human-driven solutions during HRI design processes, during which humans may serve as oracles, data sources, expert demonstrators, and, most analogously, “wizards” who control robots and provide inputs during experiments as a stand-in for not-yet-implemented or not-yet-perfected capabilities [32, 48].

Many HRI researchers employ humans as Wizards because doing so enables the science that they care about and abstracts away irrelevant, impossible, or out-of-focus problems—the same high-level goals that LLMs offer. In keeping with the Wizard of Oz metaphor, when LLMs and other large pre-trained models are used as temporary stand-ins for more principled implementations of components, we refer to them as *Scarecrows*[1]—‘brainless’, straw-man, black-box modules integrated into robot architectures for the purpose of enabling full-pipeline solutions. While the use of such Scarecrows is not without perils, the potential uses and rewards they offer have the potential to outweigh their concerns. In this article, we explore some of these high-level opportunities, discuss some of the risks that should be considered by researchers using LLMs, and recommend some reporting guidelines intended to support the principled use of large pre-trained models in HRI.

2 SOME OPPORTUNITIES OF LARGE LANGUAGE MODELS IN HRI

Although attempting to directly put ChatGPT on a robot as a human-interaction system may be a recipe for disaster (see Sec. 3), researchers have recently demonstrated that LLMs may have other uses in robotics that do not require direct dialog with humans. An example of this is recent work at Microsoft Research [58], which demonstrated the use of LLMs not to achieve end-to-end dialogue, but rather to help facilitate narrower tasks within a robot architecture such as the translation of natural language instructions to robot control code in a high-level function library.

Out of the box, pre-trained LLMs offer end-to-end capabilities for general language-based tasks; however, these models can be fine-tuned (e.g., using zero-shot [27] or few-shot learning [38]) to accomplish specific tasks, behaviors, or capabilities within a software architecture. In general, the process of building one of these fine-tuned models reduces the requirements of data, compute power, time, money, and overall workload of researchers in developing new computational capabilities, enabling them to focus instead on the software components that are most relevant to their research and to quickly scale up their existing software infrastructure. Below, we discuss the application and potential value of these models in HRI.

In the context of experimental HRI research, we refer to such fine-tuned LLMs as “Scarecrows”. This metaphor extends that of Wizard of Oz (WoZ) [32, 48], which is readily employed by researchers in HRI experiments. For example, if an experiment requires Automatic Speech Recognition (ASR),

[1] For those unfamiliar, the Scarecrow is a character from the children’s novel, *The Wonderful Wizard of Oz* [9]. In the story, the Scarecrow character is a literal straw-man who lacks a brain and seeks the powerful Wizard of Oz to grant him one.

but ASR is not the focus of the experiment, is not the expertise of the researcher, or is otherwise impractical, then HRI researchers often "Wizard" the ASR component, temporarily replacing it with a human-in-the-loop who provides transcribed speech to the robot software architecture. In other cases, placeholder robot behaviors might be naively implemented or neglected entirely; for example, it is common for robot eye gaze controllers to naively stare at tracked faces and/or avert to random locations, and for robot proxemic behaviors to be relegated to stationary interactions, unreactive to human locations and actions, despite studies showing their impact on teamwork and interactive learning [11, 12, 30, 37, 41, 42]. Neither these Wizard behaviors nor short-term placeholder behaviors are defensible as deployable "engineering" solutions, yet they are reasonably employed in experiments to enable researchers to investigate their phenomena of interest.

In a similar way (and with sufficient forethought), we propose that **Scarecrows can be used as temporary stand-ins** when the implementation of more rigorous models is infeasible given the resources available, enabling HRI science to progress in the face of understandably deprioritized problems. In these use cases, an LLM used as a Scarecrow might quickly offer comparable or better performance than a weaker placeholder or Wizard. Conversely, we also propose non-Scarecrow use cases in which LLMs are employed as sound or rigorous engineering solutions; for example, if an LLM-based parser were known to be the best semantic parser currently available in the research literature, and met all needs of importance to the researcher, it would not be considered a Scarecrow.

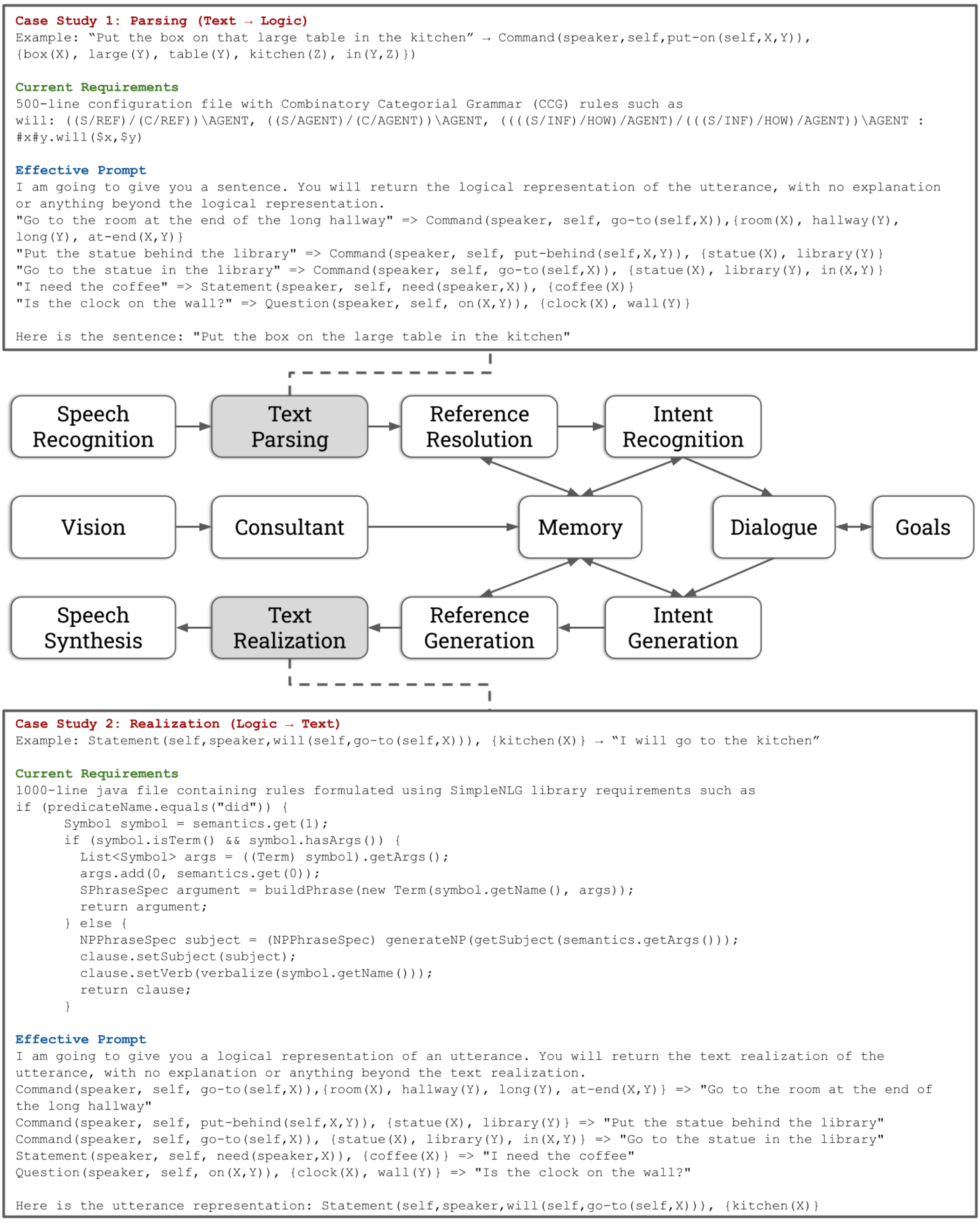
Given the complexity of HRI, we anticipate that many Scarecrows could exist within a single HRI software architecture. This would be similar to an architecture with multiple placeholder robot behaviors and possibly Wizard behaviors. Additionally, the use of Scarecrows affords more complex combinations of robot behaviors, including those that would be challenging within a WoZ methodology; for example, multiple Scarecrows could be employed in software pipelines in which multiple Wizards might be desirable but impractical, and could even be used in conjunction with a Wizard. A multi-Scarecrow architecture would follow recent trends in LLMs (e.g., [26, 47, 55, 63]).

As a practical example, Fig. 1 shows a typical configuration of the Distributed Integrated Affect Reflection Cognition (DIARC) Robot Cognitive Architecture [52, 54], which has been used in a large number of papers published at the HRI conference [13–16, 40, 43, 50, 53, 56, 59–61, 65]. Most research using DIARC focuses on high-level cognitive processes at the language-memory boundary (reference resolution and generation, intent inference and generation, dialogue, and goal management). However, to make practical use of the architecture, a new researcher must configure the system to be able to transform incoming text into logical representations and vice versa: a hurdle that is overly reliant on hard-coded rules, and which presents a substantial challenge to new researchers—especially undergraduate researchers, who are loathe to learn complex formalisms like Combinatory Categorial Grammars [57] when their intended research project focuses on emotion, moral reasoning, or other capabilities far downstream in the architecture. Fig. 1 summarizes these current requirements for Text Parsing and Text Realization, and gives examples of effective prompts that, when delivered to ChatGPT [58], produce the desired response for sample test cases.

The use of Scarecrows has the potential to allow HRI researchers to quickly build capabilities into their architectures that may not be the focus of the immediate research, and which are either "good enough" for their intended purpose or which can be replaced later with more principled solutions. However, using LLMs in HRI is not without risks. Understanding the ways that LLMs and other large pre-trained models might be used as Scarecrows also makes clear some of the potential problems of their use within robotics applications. We discuss these risks in the next section.

3 SOME PERILS OF LARGE LANGUAGE MODELS IN HRI

Robotics researchers consistently underestimate the challenge of working with natural language as an interaction modality, and may face a rude awakening when they realize that they cannot



```
Case Study 1: Parsing (Text → Logic)
Example: “Put the box on that large table in the kitchen” → Command(speaker,self,put-on(self,X,Y)),
{box(X), large(Y), table(Y), kitchen(Z), in(Y,Z)})

Current Requirements
500-line configuration file with Combinatory Categorial Grammar (CCG) rules such as
will: ((S/REF)/(C/REF))\AGENT, ((S/AGENT)/(C/AGENT))\AGENT, ((((S/INF)/HOW)/AGENT)/(((S/INF)/HOW)/AGENT))\AGENT :
#x#y.will($x,$y)

Effective Prompt
I am going to give you a sentence. You will return the logical representation of the utterance, with no explanation
or anything beyond the logical representation.
"Go to the room at the end of the long hallway" => Command(speaker, self, go-to(self,X)),{room(X), hallway(Y),
long(Y), at-end(X,Y)}
"Put the statue behind the library" => Command(speaker, self, put-behind(self,X,Y)), {statue(X), library(Y)}
"Go to the statue in the library" => Command(speaker, self, go-to(self,X)), {statue(X), library(Y), in(X,Y)}
"I need the coffee" => Statement(speaker, self, need(speaker,X)), {coffee(X)}
"Is the clock on the wall?" => Question(speaker, self, on(X,Y)), {clock(X), wall(Y)}

Here is the sentence: "Put the box on the large table in the kitchen"
```

```
Case Study 2: Realization (Logic → Text)
Example: Statement(self,speaker,will(self,go-to(self,X))), {kitchen(X)} → “I will go to the kitchen”

Current Requirements
1000-line java file containing rules formulated using SimpleNLG library requirements such as
if (predicateName.equals("did")) {
      Symbol symbol = semantics.get(1);
      if (symbol.isTerm() && symbol.hasArgs()) {
        List<Symbol> args = ((Term) symbol).getArgs();
        args.add(0, semantics.get(0));
        SPhraseSpec argument = buildPhrase(new Term(symbol.getName(), args));
        return argument;
      } else {
        NPPhraseSpec subject = (NPPhraseSpec) generateNP(getSubject(semantics.getArgs()));
        clause.setSubject(subject);
        clause.setVerb(verbalize(symbol.getName()));
        return clause;
      }

Effective Prompt
I am going to give you a logical representation of an utterance. You will return the text realization of the
utterance, with no explanation or anything beyond the text realization.
Command(speaker, self, go-to(self,X)),{room(X), hallway(Y), long(Y), at-end(X,Y)} => "Go to the room at the end of
the long hallway"
Command(speaker, self, put-behind(self,X,Y)), {statue(X), library(Y)} => "Put the statue behind the library"
Command(speaker, self, go-to(self,X)), {statue(X), library(Y), in(X,Y)} => "Go to the statue in the library"
Statement(speaker, self, need(speaker,X)), {coffee(X)} => "I need the coffee"
Question(speaker, self, on(X,Y)), {clock(X), wall(Y)} => "Is the clock on the wall?"

Here is the utterance representation: Statement(self,speaker,will(self,go-to(self,X))), {kitchen(X)}
```

Fig. 1. Example use of LLMs as Scarecrows for HRI research. This diagram depicts a configuration of the DIARC robot architecture used in previous work [54] (further described in Section 2), and two case studies of how LLM Scarecrows might be used within that architecture [54]. For each *Case Study* (red), the reader should note both (a) the complexity and inscrutability of the *Current Requirements* (green), and (b) the simplicity and legibility of the *Effective Prompt* (blue).

simply add an Automatic Speech Recognition system to their robot and quickly enable natural dialogue-based interactions. The introduction of Large Language Models may present the opposite problem. By integrating a Large Language Model into a robot architecture, a researcher could indeed quickly enable natural-sounding dialogue-based interactions. But the simplicity of such integration obscures the ways that the resulting solution will fail to meet the practical and ethical standards of human-robot interaction, and obscures how far such an integration might be from an actually deployable solution. In this section, we summarize some of the concerns that HRI researchers need to carefully consider before pursuing LLM-based solutions. In particular, we argue that while the use of LLMs as Scarecrows presents a promising opportunity for HRI researchers, such use *may* elicit some of the well-known risks regarding LLMs—depending on both the manner and scope of the linguistic capabilities replaced in this way.

The most well-known problems with end-to-end LLMs result from their lack of trustworthiness. LLMs regularly generate text that is plainly false, and “hallucinate” declarative knowledge that is phrased in a way that is confident and assertive yet bears no grounding to reality. Moreover, notwithstanding the “guard rails” set in place in some such models, they regularly generate text that is either overtly norm-violating or that exhibits encoding of racist or sexist stereotypes [25] and biases. In some recent studies, LLMs have been found to inadvertently produce toxic utterances depending on the prompt [20]. The generation of counterfactual or immoral text by these models is not clearly solvable, as by their design these models are simply not designed for accuracy or normative adherence. As we have discussed in previous work [62], these models operate by “bullshitting” (in the sense described by Frankfurt [24]): speaking without regards for the truth of what is said, to convince the listener of some attribute (deservingness of reward due to predictive accuracy). This stands in stark contrast to much of the work on human-robot dialogue, which is intentional and goal-directed.

The sacrifice of accuracy, morality, and intentionality in service of fluency stands in stark contrast to the needs of virtually all applications requiring Natural Language Generation (NLG). As NLG pioneer Ehud Reiter argues, “authors [of Deep Learning-based NLG papers] are fixated on... fluency, but regard factual accuracy as unimportant. This is a bizarre perspective because one thing that 30 years of NLG research has taught me is that readers of NLG texts care hugely about accuracy, and indeed prefer accurate-but-poorly-written texts over inaccurate-but-fluent texts” [46]. For robotics, however, this lack of accuracy, morality, and intentionality is particularly fraught. As a wealth of HRI research has demonstrated, robots wield substantial persuasive power [5, 64]. Social robots in general pose risks of manipulating people [33, 51], but language-capable robots in particular have unique potential for moral influence [28], perhaps due to robots’ perceived mental states [31] and perceived competence [36], the social and moral judgements this evokes [6], and the resulting downstream attributions of moral and social agency to robots [23, 29].

This persuasive power is also particularly fraught for language-capable robots because of the human tendency to ascribe identity characteristics to anthropomorphic robots, whether that anthropomorphism stems from human-like morphology [2, 8, 45] or from language capability [1, 7]. If robot speech is grounded in models like Generative Pre-trained Transformers (GPTs), which are trained to represent the language patterns found on the internet, the ultimate result will be the white male default voice of the internet becoming the voice of robotics. This would not only exacerbate the “whiteness of AI” [17], but moreover when combined with robots’ persuasive power, would turn robots into a tool for reinforcing white male hegemonic power.

The white male monoculture of Large Language Models also contrasts with positive trends in the HRI literature surrounding increased attention to identity and culture during robot design, and surrounding the use of techniques like Participatory Design. The HRI community is increasingly embracing an approach to research, engineering, and design grounded in attending to a community’s needs and working to address those needs through practices like co-design and by leveraging the rich collection of design guidelines presented in the HRI literature. The sacrifice of the intentionality of not only robots but also their designers, through an abdication of the patterns of human-robot dialogue to “however text looks on the internet,” would seem to stand in complete opposition to these community trends.

Finally, LLMs present significant challenges from a scientific perspective. Given the black box nature of LLMs (especially for those behind the digital walls of companies like OpenAI), there is no way of telling exactly what serves as the foundation for HRI research building on such models. The black box nature of LLMs is especially concerning when considering the inconsistency of these models, both *within* models (as multiple model runs can lead to wildly different outcomes) and *between* models (as model performance can differ in opaque ways between model versions).

These critiques help us to understand *where* and *when* the use of LLMs in robotics might be most worrisome: in contexts where biases and stereotypes risk bleeding into robot speech, where problematic inaccuracies may be introduced, and where explainability is needed. These risks are obviously not equally distributed throughout a robot architecture. Instead they are differently likely to manifest during the process of translation from speech to utterance representation; during dialogue modeling; during norm violation response; and during sentence realization.

This uneven distribution of risks highlights the need for HRI researchers using LLMs as Scarecrows to transparently document where, how, and why they are using them. To consider *how* HRI researchers might best report on their use of Scarecrows, we suggest inspiration be taken from reporting guidelines that have been developed in our community for experimental researchers’ use of Wizard of Oz techniques during disparate parts of their architectures. Specifically, the community has developed standards and practices for understanding when and how to appropriately employ WoZ and how to report the use of WoZ in publications [48]. Drawing parallels to that work, we argue that the field is in need of a similar set of criteria for considering when it is appropriate to use Scarecrows, and guidelines for explaining this rationale, the means of deployment, and the anticipated risks of their use, even if that use is intended to be short term. We discuss these guidelines in the next section.

4 SOME USAGE AND REPORTING GUIDELINES FOR LANGUAGE MODELS IN HRI

Careful reporting guidelines are critical when using LLMs for several reasons. First, as described in Section 3, the advantages of LLMs come at the cost of a number of ethical risks which, if unmitigated, could have serious impacts on the humans with whom robots are interacting, or even on society at large. Our community must be careful to characterize, discuss, and mitigate these effects. Second, the inconsistent and stochastic performance of LLMs may conflict with the HRI community’s values regarding replicability and reproducibility. Reporting precisely which models were used and how, and the ways this may have impacted robot behavior within and between experiments, may help to head off these concerns. Finally, some researchers have raised questions as to whether the use of LLMs in scientific practice constitutes academic misconduct, due to the view that these models may operate via plagiarism; while these concerns have largely been raised in the context of LLM-driven generation of papers and code [18, 19], HRI researchers should nevertheless be maximally transparent about their use of LLMs in order to head off similar concerns. In this section, we explore some of the considerations that researchers may wish to report on when using Scarecrows in their HRI research. Here, we suggest an initial sample of possible guidelines to seed a discussion in the HRI community of more rigorous and comprehensive reporting goals. These recommendations are deliberately high level because this is a rapidly evolving research space.

Consideration 1: Model specification. The text generated by LLMs differs extensively based not only on the LLM used, but also based on the specific version of the LLM and the way it was used. As such, HRI researchers using LLMs should consider reporting data necessary to most effectively reproduce their results. This would include not only which LLM was used, but also what version of the model and, if available, the snapshot of the weights. More broadly, it also requires including information about how the model was initialized and what kinds of prompts were used, as well as whether it was fine-tuned and how. To be clear, this level of reporting is not standard practice in the machine learning community, yet it may be necessary for scientific replicability. At the design level, it is also worth reporting on why these specific decisions were made and whether there is a model card (e.g., [39]) or similar reporting mechanism that can be used.

Consideration 2: Motivation for using LLMs. When using LLMs as Scarecrows, it should be obvious how and why LLMs were used (regardless of the details of the specific LLM used). As such, HRI researchers using LLMs should consider reporting their answers to the following questions: what capabilities were replaced by Scarecrows, and why were LLMs used rather than fully implementing the capability? The use of LLMs, as opposed to other (possibly less complex) baselines should be justified, and the way in which the use of LLMs changes robot behavior (positively or negatively) as compared to the use of a baseline policy or full implementation should be discussed.

Consideration 3: Ethical concerns and their mitigation. The use of LLMs has not only practical implications for effective robot deployment, but also potential ethical risks, as described in Section 3. For each use of LLMs within their robot architecture, researchers should be specific about the extent to which the types of risks we have described do or do not manifest for that particular use. For example, the potential to generate hate-speech may be substantial when LLMs are used to generate free-form text based directly on user utterances, but minimal or non-existent when LLMs are used to parse task descriptions into Linear Temporal Logic. For risks that do present themselves, researchers should make reasonable effort to explain how those risks were mitigated, and whether the use of LLMs and their known risks were reported to IRBs when obtaining ethical approval for any associated human-participant experiments.

Consideration 4: Path to deployment. Finally, it needs to be clear whether the use of LLMs is intended as a short-term stand-in or a long-term solution. Depending on the answer to this question, different information might be necessary for other researchers to weigh the merits of the proposed approach. If LLMs are intended as a short-term stand-in (that is, if they are truly being used as Scarecrows), researchers should do their best to explain what would be needed to replace the LLM with an alternate long-term solution in their architecture, and the engineering and socio-technical challenges that would need to be addressed to do so successfully. Alternatively, if an LLM is being used not as a Scarecrow, but as a genuine long-term solution, researchers should be clear how their use of the LLM would need to be validated in other work. Moreover, researchers should be clear about the license and cost of using their selected LLMs, for example, in terms of money and energy, including the estimated impact on the environment, if possible.

5 CONCLUSION

LLMs allow AI agents to demonstrate previously unseen capabilities, and researchers are only beginning to explore the possibilities they offer. LLM can be added to human-interactive architectures either as a mechanism for efficient prototyping—that is, as Scarecrows—or to add state-of-the-art capabilities. In either case, researchers who make use of these models should bear in mind that they produce factually untrustworthy language, lack any inbuilt concept of morality or intentionality, have the potential to be dangerously persuasive when combined with robots, and are fundamentally neither transparent nor consistent.

Despite these risk factors, the idea of using LLMs as Scarecrows in larger HRI architectures is both powerful and inevitable. Scarecrows have the potential to serve as stand-ins for people or for complex system elements, allowing researchers to sidestep engineering questions and focus on the aspects of a system they are actively studying, with the understanding that those components are to be replaced with a more principled solution at a later time. They may also represent state-of-the-art solutions for certain problems or sub-problems, despite their opacity and lack of replicable behavior. LLMs are unexpectedly capable, but for that very reason, we need to deliberately consider when and whether they are appropriate before plugging them into human-interactive architectures.

It would be dangerously reactionary to refuse to use tools that work well for their intended purpose; however, there is also danger in using powerful tools thoughtlessly. The use of large pre-trained models in HRI is in a relatively early stage, meaning that now is the time to think about constraints and guidelines: how can we meaningfully take advantage of the capabilities enabled by these models without losing replicability, transparency, and ethical design? Now—*before* the use of these models becomes widespread—is the right time to consider these questions, balance benefits and risks, and develop guidelines and standards as a community.

ACKNOWLEDGMENTS

We thank attendees of the 2023 NSF workshop on *Dialogue with Robots*[2], whose insightful discussions on the future of robotics and language led to this paper. Tom Williams' work on this paper was funded in part by grant N00014-21-1-2418 from the US Office of Naval Research and Young Investigator Award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research. Cynthia Matuszek's work on this paper was supported in part by NSF grants IIS-2024878 and IIS-2145642.

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. 24–33.

[2] Arifah Addison, Christoph Bartneck, and Kumar Yogeeswaran. 2019. Robots can be more than black and white: examining racial bias towards robots. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 493–498.

[3] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[4] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can and Not As I Say: Grounding Language in Robotic Affordances. In *arXiv preprint arXiv:2204.01691*.

[5] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3 (2011), 41–52.

[6] Jaime Banks. 2021. Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics* 13, 8 (2021), 2021–2038.

[7] Jessica Barfield. 2023. Designing Social Robots to Accommodate Diversity, Equity, and Inclusion in Human-Robot Interaction. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 463–466.

[2] https://www.malihealikhani.com/dialogue-with-robots

[8] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots and racism. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 196–204.

[9] L. Frank Baum. 1900. *The Wonderful Wizard of Oz*. George M. Hill Company.

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[11] Matt Berlin, Cynthia Breazeal, and Crystal Chao. 2008. Spatial scaffolding cues for interactive robot learning. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1229–1235.

[12] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[13] Timothy Brick and Matthias Scheutz. 2007. Incremental natural language processing for HRI. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 263–270.

[14] Gordon Briggs, Meia Chita-Tegmark, Evan Krause, Will Bridewell, Paul Bello, and Matthias Scheutz. 2022. A Novel Architectural Method for Producing Dynamic Gaze Behavior in Human-Robot Interactions. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 383–392.

[15] Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. 2010. Robust spoken instruction understanding for HRI. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 275–282.

[16] Rehj Cantrell, Kartik Talamadupula, Paul Schermerhorn, J Benton, Subbarao Kambhampati, and Matthias Scheutz. [n. d.]. Tell me when and why to do it!. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*. 471.

[17] Stephen Cave and Kanta Dihal. 2020. The whiteness of AI. *Philosophy & Technology* 33, 4 (2020), 685–703.

[18] ACL 2023 Program Chairs. 2023. ACL 2023 Policy on AI Writing Assistance. https://2023.aclweb.org/blog/ACL-2023-policy.

[19] ICML 2023 Program Chairs. 2023. Clarification on Large Language Model Policy LLM. https://icml.cc/Conferences/2023/llm-policy.

[20] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. arXiv:2304.05335 [cs.CL]

[21] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).

[22] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics* 9 (2021), 1012–1031.

[23] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14 (2004), 349–379.

[24] Harry G Frankfurt. 2009. On bullshit. In *On Bullshit*. Princeton University Press.

[25] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.findings-emnlp.301

[26] Edd Gent. 2023. The Stickle-Brick Approach to Big AI. *IEEE Spectrum* (April 2023). https://spectrum.ieee.org/large-language-models

[27] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*. PMLR, 9118–9147.

[28] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 401–410.

[29] Ryan Blake Jackson and Tom Williams. 2021. A theory of social agency for human-robot interaction. *Frontiers in Robotics and AI* 8 (2021), 687726.

[30] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1555–1566.

[31] Merel Keijsers, Christoph Bartneck, and Friederike Eyssel. 2022. Pay them no mind: the influence of implicit and explicit robot mind perception on the right to be protected. *International Journal of Social Robotics* (2022), 1–16.

[32] John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2, 1 (1984), 26–41.

[33] Cherie Lacey and Catherine Caudwell. 2019. Cuteness as a 'dark pattern'in home robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 374–381.

[34] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as Policies: Language Model Programs for Embodied Control. In *arXiv preprint arXiv:2209.07753*.

[35] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language Models of Code are Few-Shot Commonsense Learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[36] Bertram F Malle, K Fischer, J Young, A Moon, and E Collins. 2020. Trust and the discrepancy between expectations and actual capabilities. *Human-robot interaction: Control, analysis, and design* (2020), 1–23.

[37] Ross Mead and Maja J Matarić. 2016. Robots Have Needs Too: How and Why People Adapt Their Proxemic Behavior to Improve Robot Social Signal Understanding. *Journal of Human-Robot Interaction (JHRI)* 5, 2 (2016), 48–68. https://doi.org/10.1007/s10514-016-9572-2

[38] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. 2023. Grounding Language with Visual Affordances over Unstructured Data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. London, UK.

[39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[40] Terran Mott and Tom Williams. 2023. Rube-Goldberg Machines, Transparent Technology, and the Morally Competent Robot. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 634–638.

[41] J. Mumm and B. Mutlu. 2011. Human-Robot Proxemics: Physical and Psychological Distancing in Human-Robot Interaction. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI-2011)*. Lausanne, Switzerland, 331–338. https://doi.org/10.1145/1957656.1957786

[42] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 1–33.

[43] Bradley Oosterveld, Luca Brusatin, and Matthias Scheutz. 2017. Two bots, one brain: Component sharing in cognitive robotic architectures. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 415–415.

[44] Jiayi Pan, Glen Chou, and Dmitry Berenson. 2023. Data-Efficient Learning of Natural Language to Linear Temporal Logic Translators for Robot Task Specification. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.

[45] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 110–119.

[46] Ehud Reiter. 2019. Generated Texts Must Be Accurate! https://ehudreiter.com/2019/09/26/generated-texts-must-be-accurate/ Accessed: 2023-02-02.

[47] Toran Bruce Richards. 2023. Auto-GPT. https://github.com/Significant-Gravitas/Auto-GPT.

[48] Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[49] Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed AI Models Make Bad Baselines. https://hackingsemantics.xyz/2023/closed-baselines/

[50] Paul Schermerhorn, Matthias Scheutz, and Charles R Crowell. 2008. Robot social presence and gender: Do females view robots differently than males?. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 263–270.

[51] Matthias Scheutz. 2011. The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. In *Robot ethics: The ethical and social implications of robotics*. MIT Press.

[52] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. 2013. Novel Mechanisms for Natural Human-Robot Interactions in the DIARC Architecture. In *Proceedings of the 2013 AAAI Workshop on Intelligent Robotic Systems*.

[53] Matthias Scheutz, Paul Schermerhorn, and James Kramer. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 226–233.

[54] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An Overview of the Distributed Integrated Cognition Affect and Reflection DIARC Architecture. In *Cognitive Architectures*, Maria Isabel Aldinhas Ferreira, João S.Sequeira, and Rodrigo Ventura (Eds.). Springer.

[55] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

[56] Rafael Sousa Silva and Tom Williams. 2023. Enabling Human-like Language-Capable Robots Through Working Memory Modeling. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 730–732.

[57] Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell* (2011), 181–224.

[58] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. (2023).

[59] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 353–362.

[60] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated open world reference resolution for human-robot dialogue. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 311–318.

[61] Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. " Thank You for Sharing that Interesting Fact!" Effects of Capability and Context on Indirect Speech Act Use in Task-Based Human-Robot Dialogue. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*. 298–306.

[62] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The confucian matador: three defenses against the mechanical bull. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 25–33.

[63] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–10.

[64] James Young. 2021. Danger! This robot may be trying to manipulate you. *Science Robotics* 6, 58 (2021).

[65] Chen Yu, Matthias Scheutz, and Paul Schermerhorn. 2010. Investigating multimodal real-time patterns of joint attention in an hri word learning task. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 309–316.

[66] Bowen Zhang and Harold Soh. 2023. Large Language Models as Zero-Shot Human Models for Human-Robot Interaction. arXiv:2303.03548 [cs.RO]