My research vision is to **ground AI systems** with an understanding the world – which is needed for humans to communicate with them safely and reliably. I perform **interdisciplinary** research: combining methodology from **natural language processing** and **computer vision** with insights from **developmental psychology**. I am inspired by how young children learn a commonsense mental model of 'how the world works,' and then learn language on top of this model [8]. Over time, they master **physical** reasoning over objects and actions, along with higher-order **event** reasoning about complex situations. Today's machines struggle with both. My research seeks to bridge this gap from three angles:

a. **Grounding physical dynamics**, through *symbolic* and *graph-structured* representations, that transfer to the world of unstructured language (Section 1; [1, 11, 14, 18]).

b. **Grounding events** – complex situations that evolve over time – through *multimodal neural script knowledge* learned through large-scale self-supervision (Section 2; [9, 10, 12, 13, 17]).

c. **Adversarial evaluation of grounding** through new methodology to test machines **at the limits and at scale**, guiding field-wide progress (Section 3; [1–4, 6, 9, 10, 12, 13]).

Figure 1 shows an example that 'grounds' these directions. The image comes from Visual Commonsense Reasoning (VCR; [9]), one of several 'AI grand challenges' created through my work. Three of these (VCR, HellaSwag [12], and PIQA [1]) were chosen as official benchmarks for the DARPA Machine Commonsense Program, out of eight total selected. The challenges have seen over 100 model submissions so far, and they have been featured by popular press outlets like The New York Times.[1]

Though my work on grounded models has established state-of-the-art results on challenges like VCR [16, 17], the challenge of grounding AI is far from solved. I am excited to continue pursuing it as a professor, together with students and collaborators.

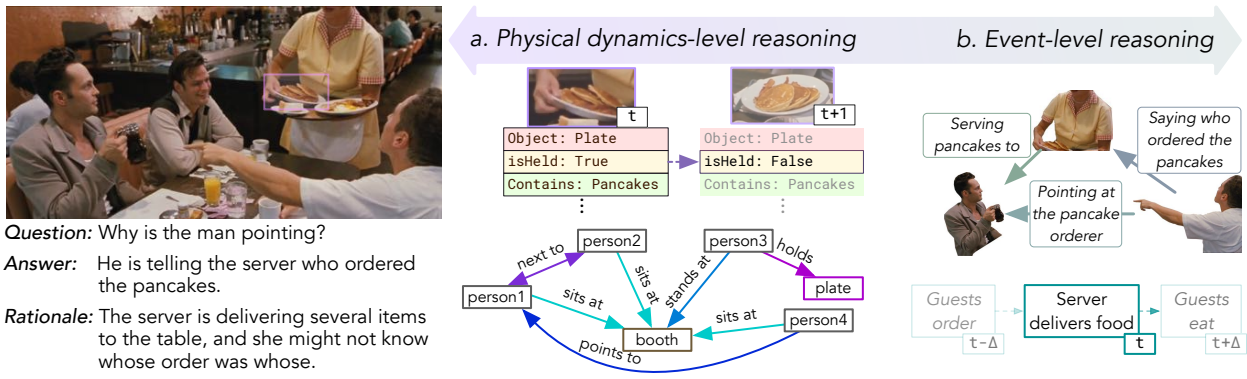


Figure 1: Synthesis of my work on commonsense grounding, through an example from VCR [9]. I have worked on physical reasoning through embodied interaction in a 3D world [14] and over graph-structured representations [18], along with event-level reasoning [10, 12] over time [16, 17].

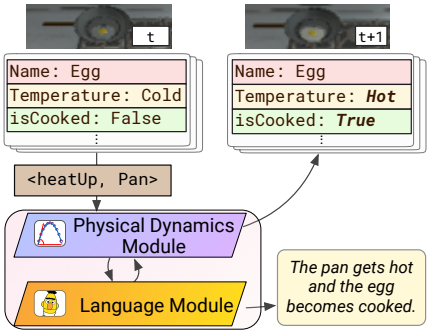
1 Grounding physical dynamics through symbolic structure

When using language in real-world situations, we are guided by physical knowledge about objects and actions. As humans, we intuitively think of many objects in a **structured** and **symbolic** way, along with actions as *transforming* these objects. I have worked towards modeling such physical understanding, and using it to inform how machines understand and generate language.

[1] 'Finally, a Machine That Can Finish Your Sentence' (nyti.ms/2DycutY), covering SWAG [10, 12].

Structured models for objects and actions. I have studied grounding real-world actions through English verb frames [11] – e.g., a verb like 'jump' implies a short duration. I also proposed a method for building **graph models** of real-world scenes, while integrating common motifs – e.g. if somebody is riding a bicycle, it probably has wheels [18]. Collaborators and I have used such symbolic knowledge to constrain natural language generators [5].

Learning physical dynamics through interaction. I have investigated new approaches for grounding language, inspired by how children learn *grounded* language. I introduced PIGLeT, a framework and model that learns physical dynamics **through interaction in a 3D environment**, and then uses this to ground language [14]. We train our model by having it predict "what happens next" explicitly and symbolically: for example, when an `Egg` is placed on a `Pan` that is `Hot`, it changes state to `Cooked`. We link this learned dynamics model with a language model. The combined PIGLeT model can 1) reliably generate natural language summaries of physical state changes, and 2) reason about physical dynamics written through language. It outperforms expensive 'text-only' models with 100x the parameters.



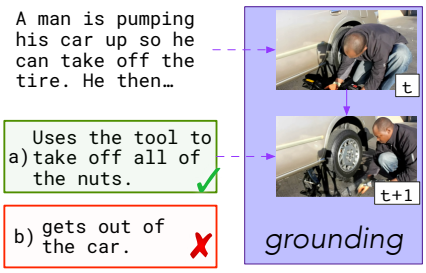
2 Grounding events through multimodal script knowledge

Human-level language understanding is further grounded in everyday events and situations. Suppose we are familiar with cars, and encounter the sentence:

(🚙) A man is pumping his car up so he can take off the tire.

As humans, we can ground the meaning of this new *event*, to a lived experience of similar situations. For example, we might imagine what it could 'look like,' or predict what might happen next. In my work, I have defined this understanding as **multimodal script knowledge**: how events happen causally in the world, in what order, and perceived through which perceptual modalities [16, 17].

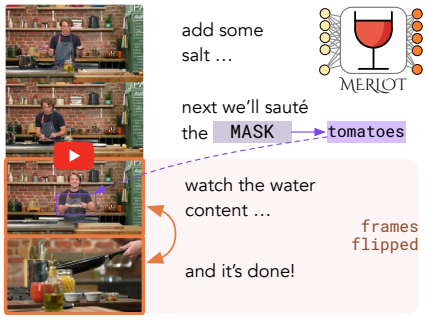
Formalizing commonsense event-level grounding. Over the course of my Ph.D., I have worked towards defining *how* machines can demonstrate event-level understanding of descriptions like (🚙). My work on the Situations With Adversarial Generations (SWAG) benchmark defined this task concretely [10]. Given a sentence like (🚙), a machine must choose the most likely follow-up. The sentences describe real-world events, so the 'correct' answer is the one that *actually happens next*, ensuring grounding. Our task remains challenging [12].



Visual commonsense reasoning over script knowledge. When looking at the restaurant photo in Figure 1, we can infer *beyond the frame* about what might be going on in the scene and why: for instance, that a man is likely pointing at his companion *to tell the server who ordered the pancakes*. Through my work on the Visual Commonsense Reasoning (VCR) benchmark, I proposed evaluating this capacity through multiple-choice question answering – *why is the man pointing?* Beyond answering correctly, models must also *provide a rationale* justifying their answer: connecting event-level commonsense reasoning to explanations from physics-level object, activity, and scene recognition. This work has seen great interest from the community: over 80 models and counting have been submitted to VCR's leaderboard.

Learning multimodal (neural) script knowledge. Over the last few years, many large-scale NLP and computer vision models have been trained on a combination of text, images, and manual annotations – yet, this approach has not been sufficient to 'solve' tasks like VCR.

My work introduces a new approach, where we train a model on *multimodal* and *temporal* data from YouTube. We use new self-supervised objectives to learn multimodal script knowledge [17]. We dub our model MERLOT, short for Multimodal Event Representation Learning over Time. Our model sets new **state-of-the-art results on twelve video reasoning tasks**, as well as on VCR. In doing so, it outperforms larger, industry-submitted models that that learn from *static* data: images annotated with object detections, and literal descriptions.



I recently built on this work through a new model named MERLOT Reserve [16]. The idea is to learn through developmental psychology-inspired *reentry* – learning connections between all modalities **including sound** to understand videos [8]. For example, predicting the sound in the above figure requires understanding if the person is sautéing or frying the vegetables; sound thus supervises visual and world understanding. Perhaps surprisingly, our work shows that integrating **sound improves vision-and-text representations**. We set a new state-of-the-art on VCR, even though it doesn't include any sound for models.

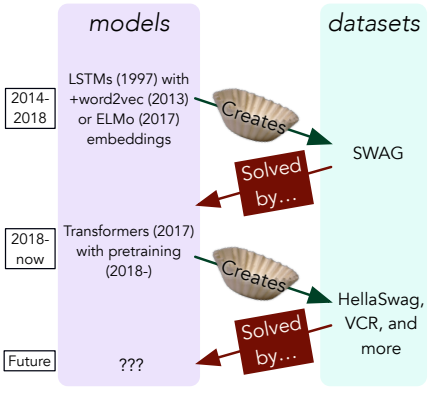
3 Adversarial evaluation of grounding

Building grounded language understanding systems is a hard challenge: one that will require research across many disciplines, both in AI and beyond. To this end, I have helped **direct field-wide efforts** through my work on evaluating AI grounding. My benchmarks, along with the new algorithms and methodology I have proposed for making benchmarks, have been widely adopted by the community and showcase issues with models, even at the extreme (GPT-3) scale.

Adversarial Algorithms for Benchmark Dataset Creation. The key challenge in benchmarking grounding is that neural models excel at "gaming" benchmarks. When presented with thousands of *human-written* exam questions (and answers) as training data, models can learn to identify spurious patterns. This enables them to *outperform* humans within such a closed setting, yet while struggling to generalize to new or out-of-distribution examples that might be found in the external world.

In our work on SWAG, we introduced a new approach for dataset creation, called **Adversarial Filtering (AF)** [10]. The idea is to use machines themselves in the exam-creation process: both to generate 'wrong' answers, and to *filter out* easily-spotted wrong answers – in turn replacing them with others. The result is a dataset that is adversarial for that class of models (and weaker ones).

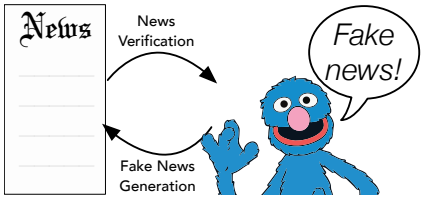
Adversarial filtering enables **co-evolution between benchmarks and models**: as models improve, we can use those models to make updated benchmarks. The past few years have seen several rounds of this co-evolution. SWAG was difficult for 2018-era models, yet 300-million parameter models reached close to human performance. In 2019, we used the latest round of modeling advancements to create both HellaSWAG, as well as VCR. Both remain challenging for machines [9, 12].



In follow-up work, collaborators and I showed that AF can be used to make standard benchmarks like ImageNet more robust for both model evaluation and training [4]. The community has adapted AF to create new challenges as well, for tasks like question answering, entailment, and beyond. My collaborators' paper, of applying AF to Winograd schemas, won a best paper award at AAAI 2020 [7].

Evaluating open-ended language generation. Grounding language to world understanding

requires being able to *use it* in open-ended situations. I introduced a benchmark for this named TuringAdvice, where a model must *generate* language that would help a human resolve a real-world situation [13]. TuringAdvice reveals key weaknesses of ungrounded language models, even at extreme scale: text generated by GPT3 reveals deep misunderstandings of the situations being discussed. My recent work with collaborators, proposing divergence frontiers to compare machine generations to human text, recently won an outstanding paper award at NeurIPS 2021 [6].



Adversarial threat modeling. My research also tests what models can do at scale, from a computer security lens. A concern with language generators is 'neural fake news': the risk of machine-generated propaganda that reads like real news [15]. I introduced Grover, the first *threat model* for neural fake news, in turn leveraging our adversarial filtering methodology to test the limits of *when machine text can be detected*. Through our work, we discovered a **defense against neural fake news** as well – it turns out to be another fake news *generator* itself, with over 97% accuracy at telling apart real- from machine-generated news across a variety of domains. Our work on Grover was covered by The Washington Post[2] and The New York Times.[3]

4 Future plans

My short term plans include:

Unifying simulation and videos. My work proposes two paradigms for learning grounding: through interacting in a 3D world, and through watching (and listening to) YouTube. I believe these are complementary. 3D environments can teach agents the connection between their actions and the world, while YouTube enables learning about many more actions and objects, through all their modalities. I am excited to study grounding AI through a synthesis of both learning paradigms.

Concept transfer for text understanding. As humans, we represent entities in the world as unified concepts, and we can easily transfer those concepts from the visual world to language and back. I hypothesize that this is part of why humans are such efficient learners, compared to today's large ungrounded language models. I envision a machine that synthesizes knowledge from a variety of modalities to perform tasks. For instance, it might use script knowledge learned from videos to answer (🚙): 'what's likely to happen next if someone is pumping up his tire?'

My long-term vision for grounded AI includes:

Lifelong multimodal learning. Unlike today's machines, humans learn a robust grounded model of both language and the world, from a variety of complementary modes – by watching others, by interacting, or by receiving instruction through language. Supervision from these modes is often based on what we need in the moment, like when we ask a question about a concept we don't fully understand. Taking inspiration from humans, I am excited about exploring these new paradigms for model training, that will enable us to build robust grounded learners.

Interdisciplinary impact beyond Computing. The next generation of grounded machines have great potential to impact society. I believe many of these impacts are positive, but there is a risk of amplifying many of the issues that these models already have, including bias and dual-use. Some of these issues might be addressable with technical solutions, but for others, technical solutions might not be enough. Building off my work on Grover [15], I will continue collaborating with people across disciplines, including computer security, public policy, and beyond, to further study these impacts and solutions – with a goal towards directing grounding research towards just outcomes.

[2] 'Top AI researchers race to detect 'deepfake' videos: 'We are outgunned' (WaPo; archived at archive.ph/FAIwv)
[3] 'How A.I. Could Be Weaponized to Spread Disinformation' (nyti.ms/2VeaIGY)

References

[1] Yonatan Bisk, **Rowan Zellers**, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. "PIQA: Reasoning about Physical Commonsense in Natural Language". *AAAI*. Oral (top 3%). 2020.

[2] Jeff Da, Maxwell Forbes, **Rowan Zellers**, Anthony Zheng, Jena D Hwang, Antoine Bosselut, and Yejin Choi. "Edited Media Understanding Frames: Reasoning About the Intent and Implications of Visual Misinformation". *ACL*. 2021.

[3] Gabriel Ilharco, **Rowan Zellers**, Ali Farhadi, and Hannaneh Hajishirzi. "Probing Contextual Language Models for Common Ground with Visual Representations". *NAACL*. 2021.

[4] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, **Rowan Zellers**, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. "Adversarial Filters of Dataset Biases". *ICML*. 2020.

[5] Ximing Lu, Peter West, **Rowan Zellers**, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. "NeuroLogic Decoding: Unsupervised Neural Text Generation with Predicate Logic Constraints". *NAACL*. 2021.

[6] Krishna Pillutla, Swabha Swayamdipta, **Rowan Zellers**, John Thickstun, Yejin Choi, and Zaid Harchaoui. "MAUVE: Human-Machine Divergence Curves for Evaluating Open-Ended Text Generation". *NeurIPS*. Outstanding paper (top 0.1%). 2021.

[7] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. "Winogrande: An Adversarial Winograd Schema Challenge at Scale". *AAAI*. 2020.

[8] Linda Smith and Michael Gasser. "The development of embodied cognition: Six lessons from babies". *Artificial life* 11.1-2 (2005).

[9] **Rowan Zellers**, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "From Recognition to Cognition: Visual Commonsense Reasoning". *CVPR*. Oral (top 5%). 2019.

[10] **Rowan Zellers**, Yonatan Bisk, Roy Schwartz, and Yejin Choi. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference". 2018.

[11] **Rowan Zellers** and Yejin Choi. "Zero-Shot Activity Recognition with Verb Attribute Induction". *EMNLP*. 2017.

[12] **Rowan Zellers**, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "HellaSwag: Can a Machine Really Finish Your Sentence?" *ACL*. 2019.

[13] **Rowan Zellers**, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. "TuringAdvice: A Generative and Dynamic Evaluation of Language Use". *NAACL*. 2021.

[14] **Rowan Zellers**, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. "PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World". *ACL*. 2021.

[15] **Rowan Zellers**, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. "Defending Against Neural Fake News". *NeurIPS*. 2019.

[16] **Rowan Zellers**, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. "Merlot Reserve: Neural Script Knowledge through Language, Vision, and Sound". *arXiv*. 2021.

[17] **Rowan Zellers**, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. "MERLOT: Multimodal Neural Script Knowledge Models". *NeurIPS*. Oral (top 1%). 2021.

[18] **Rowan Zellers**, Mark Yatskar, Sam Thomson, and Yejin Choi. "Neural Motifs: Scene Graph Parsing with Global Context". *Conference on Computer Vision and Pattern Recognition*. 2018.