SocialAI 0.1: Towards a Benchmark to Stimulate Research on Socio-Cognitive Abilities in Deep Reinforcement Learning Agents

Grgur Kovač[*†] Inria (FR)

Rémy Portelas[*†] Inria (FR)

Katja Hofmann Microsoft Research (UK)

Pierre-Yves Oudeyer Inria (FR)

Abstract

Building embodied autonomous agents capable of participating in social interactions with humans is one of the main challenges in AI. This problem motivated many research directions on embodied language use. Current approaches focus on language as a communication tool in very simplified and non diverse social situations: the "naturalness" of language is reduced to the concept of high vocabulary size and variability. In this paper, we argue that aiming towards human-level AI requires a broader set of key social skills: 1) language use in complex and variable social contexts; 2) beyond language, complex embodied communication in multimodal settings within constantly evolving social worlds. In this work we explain how concepts from cognitive sciences could help AI to draw a roadmap towards human-like intelligence, with a focus on its social dimensions. We then study the limits of a recent SOTA Deep RL approach when tested on a first grid-world environment from the upcoming *SocialAI* , a benchmark to assess the social skills of Deep RL agents. Videos and code are available at https://sites.google.com/view/socialai01.

1 Introduction

How do human children manage to reach the social and cognitive complexity of human adults? For Vygotsky, a soviet scholar from the 1920's, a main driver for this path towards "higher-level" cognition are socio-cultural interactions with other human beings (Vygotsky and Cole, 1978). For him, many high-level cognitive functions a child develops first appear at the social level and then at the individual level. This leap from interpersonal processes to intrapersonal processes is referred to as *internalization*. Vygotsky's theories influenced multiple works within cognitive science (Clark, 1996; Hutchins, 1996), primatology (Tomasello, 1999) and the developmental robotics branch of AI (Billard and Dautenhahn, 1998; Brooks et al., 2002; Colas et al., 2020).

A more influential perspective on child development are Jean Piaget's foundational theories of cognitive development (Piaget, 1963). For Piaget, the child is a solitary thinker. While he acknowledged that social context can assist development, for him cognitive maturation happens mainly through the child's solitary exploration of their world. The child is a "little scientist" deciding which experiments to perform to challenge its assumptions and improve its representation of the world.

This Piagetian view on development is well aligned with mainstream Deep Reinforcement Learning (DRL) research, which mainly focuses on sensorimotor development, through navigation and object manipulation problems rather than language based social interactions (Mnih et al., 2015; Lillicrap et al., 2016; Forestier et al., 2017; Andrychowicz et al., 2017). The study of language has been mostly separated from DRL, into the field of Natural Language Processing (NLP), which is mainly focused in learning (disembodied) language models for text comprehension and/or generation (e.g. using large text corpora as in Brown et al. (2020)).

In the last few years however, recent advances in both DRL and NLP made the Machine Learning community reconsider experiments with language based interactions (Luketina et al., 2019; Bender and Koller, 2020). Text-based exploratory games have been leveraged to study the capacities of autonomous agents to properly navigate through language in abstract worlds (Côté et al., 2018; Prabhumoye et al., 2020; Ammanabrolu et al., 2020). While these environments allow meaningful abstractions, they neglect the importance of embodiment for language learning, which has long been identified as an essential component for proper language understanding and grounding (Cangelosi

[*] Equal contribution

[†] Email grgur.kovac@inria.fr & remy.portelas@inria.fr

et al., 2010; Bisk et al., 2020). Following this view, many works attempted to use DRL to train embodied agents to leverage language, often in the form of language-guided RL agents (Chevalier-Boisvert et al., 2018a; Colas et al., 2020; Hill et al., 2020) and embodied visual question answering (EQA) (Das et al., 2017; Gordon et al., 2018), and more recently on interactive question production and answering (Abramson et al., 2020). Multi-agent emergent communication is another subfield which studies how language can emerge from interaction in both embodied and disembodied scenarios (Mordatch and Abbeel, 2018; Jaques et al., 2019; Lowe et al., 2020; Woodward et al., 2020).

One criticism that could be made over previous work in light of Vygotsky's theory is the simplicity of the "social interactions" and language-use situations that are considered: in language-conditioned works, the interaction is merely just the agent receiving its goal as natural language within a simple and rigid interaction protocol (Luketina et al., 2019). In Embodied question answering, language-conditioned agents only need to first navigate and then produce simple one or two words answers. And because of the complexity of multi-agent training, studies on emergent communication mostly consider simplistic language (e.g. communication bits).

In our work, we propose to identify a richer set of socio-cognitive skills than those currently considered in most of the DRL and NLP literature. We organise this set along 3 dimensions: *intertwined multimodality* (coordinating multimodal actions based on multimodal observations), *theory of mind* (inferring other's mental state, e.g. beliefs, desires, emotions, etc) and *social games* (taking part in time-extended structured social interactions). We then study the failure case of a current SOTA DRL approach on a grid-world social environment. To enable the design and study of complex social scenarios in reasonable computational time, we consider single-agent learning among scripted agents (a.k.a. Non-Player-Characters or NPCs) and use low-dimensional observation and action spaces. We use templated-language, enabling to emphasize the under-studied challenges of dealing with more natural social and pragmatic situations.

Social affordances of NPCs. Although NPCs can be seen as merely complex interactive objects, we argue they are in essence quite different. NPCs, as humans, can have very complex and changing internal states, including intents, moods, knowledge states, preferences, emotions, etc. The resulting set of possible interactions with NPCs (social affordances) is essentially different than those with objects (classical affordances). In cognitive science, an affordance refers to what things or events in the environment afford to an organism (de Carvalho, 2020). A flat surface can afford "walking-on" to an agent, while a NPC can afford "obtaining directions from". The latter is a social affordance, which may require a social system and conventions (e.g. politeness), implying that the NPC must have complex internal states and the ability to reciprocate. Successful interaction might also be conditioned on the NPC's mood, requiring communication adjustments.

Training an agent for such social interactions most likely requires drastically different methods – e.g. different architectural biases – than classical object-manipulation training. We argue that studying isolated social scenarios featuring NPCs in tractable environments is a promising direction towards designing proficient social agents.

Grounding language in social interactions. In AI, *natural language* often refers to the ability of an agent to use a large vocabulary and complex grammar. We argue that this is but one dimension of the *naturalness* of language. Another, often overlooked, dimension of this *naturalness* refers to language grounding, i.e. the ability of an agent to map specific meaning from some domain to language (Steels, 2007). Command following (Chevalier-Boisvert et al., 2018a; Colas et al., 2020) is an example of language grounding in the environment. To understand the meaning of "grow green plant", an agent must relate both the plant in the environment to the word "plant", and the word "grow" to the action of watering the plant. We aim to go a step further by grounding language in social interactions, i.e. requiring social context to be understood in order to make sense of a given utterance. For example, the meaning of a NPC's utterance can change if one knows this NPC is a liar.

Social skills for socially competent agents Social skills have been extensively studied in cognitive science (Riggio, 1986; Beauchamp and Anderson, 2010) and social and developmental robotics (Cangelosi et al., 2010). Here we outline some of those skills for the purpose of studying them in the context of training *social* artificial agents.

1 - Intertwinded multimodality refers to the ability to interact using multiple modalities (verbal and non-verbal) in a coordinated manner. A proficient agent should be able to act using both primitive actions (moving) and language actions (speaking), and to process both visual and language observations (spoken by other NPCs). Importantly, this agent must be able to learn and adapt its multimodal interaction sequence, rather than following a pre-established interaction protocol, e.g. as in EQA. (Das et al., 2017), where 1) a question is given to the agent at the beginning of the episode, 2) the agent moves through the environment to gather information, and 3) upon finding an answer it responds (in language) and the episode ends. By the term *intertwined* multimodality we aim to emphasize that the modalities often interchange and the question of "when to use which modality" is non-trivial, e.g. sometimes the relevant information can be obtained by *asking* for it and sometimes by *looking* for it.

2 - Theory of Mind(ToM) refers to the ability of an agent to attribute to others and itself mental states, including beliefs, intents, desires, emotions and knowledge (Wellman, 1992; Flavell, 1999).

An agent that has ToM perceives other participants as *minds* like itself. This enables the agent to theorise about other's intents, knowledge, lack of knowledge etc. Here we outline some, of many, different perspectives of ToM to better demonstrate how ToM is essential for human social interactions.

- **inferring intents:** the agent is able to infer, based on verbal or non-verbal cues, what others will do or want to do, e.g. that some social peers are liars/trustworthy.
- **false belief:** the agent understands that someone's belief (including its own) can be faulty (Baillargeon et al., 2010).
- **self-awareness:** the capacity to take oneself as the object of thought (Wicklund, 1975).
- **imitating or emulating social peer's behaviour:** agent can imitate a behaviour seen in a social peer, or emulate its goal, e.g. upon observing a peer cut onions the agent is able to cut the onions himself, either with the same movement or with its own strategy.

3 - Social games is a concept closely related to pragmatic frames (Bruner, 1985; Vollmer et al., 2016) and language games (Wittgenstein, 1953). A social game refers to the pattern characterizing the unfolding of possible interactions (equivalent to a "grammar" for social interactions or an interaction protocol). For example, by playing turn taking games a child extracts the rule of each participant having his "turn". It then generalizes this role to a conversation where it understands that it shouldn't speak while someone else is speaking.

Closely related is the concept of roles which was proposed to be one of the key differences between human and ape socio-cognitive abilities (Tomasello, 2020). A human understands that a shared goal is completed by various participants playing different roles, which are often equally important. Crucially, we learn about others' roles by playing our own. For example, in the game of catch where one participant throws the ball and another one catches it. By playing the *catcher* role we understand what the *thrower* role consists of. This makes it easy for us to switch roles and play the *thrower* role.

Furthermore, humans have the ability to quickly detect when the *social game* changes and adapt to that change (ex. while playing football we are able to participate in *small talk* with another player).

Main contributions:

- An outline of the core socio-cognitive skills necessary to enable artificial agents to efficiently act and learn in a social world.
- A case-study of a SOTA Deep RL approach on a grid-world environment[1] featuring scripted NPCs to easily assess social skills.

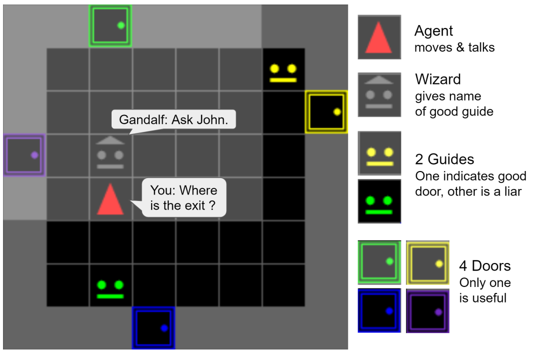


Figure 1: *TalkItOut*, a simple environment to study social skills of DRL agents. Solving it requires to master *intertwined multimodality*, basic *Theory of Mind* (detecting trustworthy agents), and a basic form of *Social Game* (standing near NPCs to interact with them). See app. A.1.4 for a discussion of required social skills.

[1] Based on Minigrid (Chevalier-Boisvert et al., 2018b)

2 Experiments and Results

Prior to a broader study of social skills in DRL, this work's experiments focus on a simple environment requiring a limited subset of social skills.

TalkItOut is a one-room grid-world environment. The agent is rewarded upon exiting the room, i.e. saying the right passphrase ("Open sesame") in front of the correct door (out of four, randomly chosen for each new episode). It can both navigate (turn left/right, go forward) and use natural language (template based, 64 possibilities), and observe a partial agent-centric symbolic pixel grid along with the history of observed language outputs from nearby NPCs. To locate the target door, the agent can question three randomly placed NPCs (by asking "Where is the exit" while standing near them): two guides – one trustworthy and one lying – and a Wizard that indicates which guide is trustworthy (e.g. "Ask Jack"). NPC names are added to their utterances to allow identification (e.g. "Jack: Go to red door"). See app. A.1 for details, fig. 1 for a visualization.

Implemented Baselines. Our main baseline is a PPO-trained (Schulman et al., 2017) Deep RL architecture proposed in (Hui et al., 2020). We chose this model as it was designed for language-conditioned navigation in grid worlds, which is similar to our setup (although in our case language input is not fixed but varies along interactions). We modify the original architecture to be Multi-Headed (*MH-BabyAI*), since our agent has to both navigate and talk. We also consider an ablated version that does not receive language inputs (*Deaf-MH-BabyAI*) and a randomly acting agent (*Random*). See appendix A.2 for details.

Results. Table 1 shows post-training success rates on a fixed random test-set of 1000 environments for all conditions. See appendix B for additional results.

The MH-BabyAI agent doesn't solve TalkItOut: its average success rate of 26% is not statistically significant from Deaf-MH-BabyAI (p>0.05, using Welch's student t-test). MH-BabyAI does not leverage language inputs: both approaches learn the suboptimal policy of going to a random door and saying the passphrase (NPCs are ignored). Similar results are observed on an ablated version of the environment that does not feature the lying guide.

Condition \ Env.	Original	No liar NPC
MH-BabyAI-EB	0.236 ± 0.01	$\mathbf{0.996 \pm 0.002}$
MH-BabyAI	0.259 ± 0.01	0.252 ± 0.010
Deaf-MH-BabyAI	0.260 ± 0.02	0.246 ± 0.014
Random	$0.002 \pm 1e^{-3}$	0.005 ± 0.002

Table 1: Success rate of studied baselines on TalkItOut and a variant without the lying guide (16 seeds, mean ± stddev, 28M steps). SOTA fails to solve TalkItOut.

One potential explanation for this failure could be that the language space is too large for the agent. To diagnose whether this is the case, we augment the MH-BabyAI baseline with intrinsic episodic exploration bonuses on observed language (i.e. a curiosity bias). The resulting MH-BabyAI with Exploration Bonuses (MH-BabyAI-EB) manages to solve the ablated no-liar NPC environment, reaching over 0.99 success rate. However, MH-BabyAI-EB still does not solve the original environment.

Additional architectural changes were tested as an attempt to improve performances (e.g. language processing, see app. B), without success.

These results showcase that the original TalkItOut environment seems to be a complex challenge DRL learners, especially for the social skills it requires, i.e. handling multi-NPCs multimodal interactions (asking the wizard then the guide) and inferring ill intentions of the false guide.

3 Conclusion And Discussion

In this work we classified and described the main socio-cognitive skills needed to build socially competent autonomous agents. As a first step towards building *SocialAI* – a test-bed to assess the social skills of DRL learners – we performed a preliminary case study on a simple social environment, and showed that a current SOTA DRL approach was unable to learn the required social skills to solve it, making it a relevant test-bed. In future work we plan to release the full *SocialAI* benchmark, which will include more SOTA baselines and multiple complex social environments to encompass a broader range of social skills.

This work suggests that architectural improvements are needed for DRL agents to learn to behave appropriately in multimodal social environments. One avenue towards this is to endow agents with mechanisms enabling to learn models of others' minds, which has been identified in cognitive neuroscience works as a key ingredient of human social proficiency (Vélez and Gweon, 2020).

References

Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Stephen Clark, Andrew Dudzik, Petko Georgiev, Aurelia Guy, Tim Harley, Felix Hill, Alden Hung, Zachary Kenton, Jessica Landon, Timothy P. Lillicrap, Kory Mathewson, Alistair Muldal, Adam Santoro, Nikolay Savinov, Vikrant Varma, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu. 2020. Imitating interactive intelligence. *ArXiv*, abs/2012.05672.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2020. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In *NeurIPS*.

Renée Baillargeon, Rose M. Scott, and Zijing He. 2010. False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3):110–118.

Miriam Beauchamp and Vicki Anderson. 2010. Social: An integrative framework for the development of social skills. *Psychological bulletin*, 136:39–64.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Aude Billard and Kerstin Dautenhahn. 1998. Grounding communication in autonomous robots: An experimental study. *Robotics and Autonomous Systems*, 24(1):71 – 79. Scientific Methods in Mobile Robotics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, and et al. 2020. Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rodney Brooks, Cynthia Breazeal, Matthew Marjanovic, Brian Scassellati, and Matthew Williamson. 2002. The cog project: Building a humanoid robot. *Lecture Notes in Artificial Intelligence*, 1562.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NeurIPS 2020*.

Jerome Bruner. 1985. Child's talk: Learning to use language. *Child Language Teaching and Therapy*, 1(1):111–114.

Angelo Cangelosi, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Chrystopher Nehaniv, Kerstin Fischer, Jun Tani, Tony Belpaeme, Giulio Sandini, Francesco Nori, et al. 2010. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2018a. Babyai: A platform to study the sample efficiency of grounded language learning.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. 2018b. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`.

Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2067–2075. JMLR.org.

Andy Clark. 1996. *Being There: Putting Brain, Body, and World Together Again*, 1st edition. MIT Press, Cambridge, MA, USA.

Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter F. Dominey, and Pierre-Yves Oudeyer. 2020. Language as a cognitive tool to imagine goals in curiosity driven exploration. In *NeurIPS 2020*.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. *ArXiv*, abs/1806.11532.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017. Embodied question answering. *ArXiv*, abs/1711.11543.

Eros Moreira de Carvalho. 2020. Social affordance. *Encyclopedia of Animal Cognition and Behavior*.

John H. Flavell. 1999. Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, 50(1):21–45.

Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. 2017. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: visual question answering in interactive environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4089–4098. IEEE Computer Society.

Felix Hill, Sona Mokra, N. Wong, and Tim Harley. 2020. Human instruction-following with deep reinforcement learning via transfer-learning from text. *ArXiv*, abs/2005.09382.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

David Yu-Tung Hui, Maxime Chevalier-Boisvert, Dzmitry Bahdanau, and Yoshua Bengio. 2020. Babyai 1.1.

Edwin Hutchins. 1996. *Cognition in the Wild (Bradford Books)*. The MIT Press.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97, pages 3040–3049. PMLR.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *ICLR*.

Ryan Lowe, Abhinav Gupta, Jakob N. Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. In *8th International Conference on Learning Representations, ICLR 2020*.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6309–6317.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1495–1502. AAAI Press.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *ICML*.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2017. Film: Visual reasoning with a general conditioning layer. *CoRR*, abs/1709.07871.

Jean Piaget. 1963. The origins of intelligence in children. *W W Norton & Co.*

Shrimai Prabhumoye, Margaret Li, Jack Urbanek, Emily Dinan, Douwe Kiela, Jason Weston, and Arthur Szlam. 2020. I love your chain mail! making knights smile in a fantasy game world: Open-domain goal-oriented dialogue agents.

Ronald Riggio. 1986. Assessment of basic social skills. *Journal of Personality and Social Psychology*, 51:649–660.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy P. Lillicrap, and Sylvain Gelly. 2018. Episodic curiosity through reachability. *ArXiv*, abs/1810.02274.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Luc Steels. 2007. The symbol grounding problem has been solved. so what's next? *Symbols, Embodiment and Meaning. Oxford University Press, Oxford, UK*.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2017. Exploration: A study of count-based exploration for deep reinforcement learning.

Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.

Michael Tomasello. 2020. The role of roles in uniquely human cognition and sociality. *Wiley: Journal for the Theory of Social Behaviour*.

Anna-Lisa Vollmer, Britta Wrede, Katharina J. Rohlfing, and Pierre-Yves Oudeyer. 2016. Pragmatic frames for teaching and learning in human–robot interaction: Review and challenges. *Frontiers in Neurorobotics*, 10:10.

L. S. Vygotsky and Michael Cole. 1978. *Mind in society : the development of higher psychological processes*. Harvard University Press Cambridge.

Natalia Vélez and Hyowon Gweon. 2020. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *PsyArXiv*.

H. M. Wellman. 1992. *The child's theory of mind*. The MIT Press.

Robert A. Wicklund. 1975. Objective self-awareness. *Advances in Experimental Social Psychology*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Mark Woodward, Chelsea Finn, and Karol Hausman. 2020. Learning to interactively learn and assist. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 2535–2543. AAAI Press.

Templates	
Action	Template
0	Where is <noun>.
1	Open <noun>.
2	Close <noun>.
3	What is <noun>.
Nouns	
Action	Noun
0	sesame
1	the exit
2	the wall
3	the floor
4	the ceiling
5	the window
6	the entrance
7	the closet
8	the drawer
9	the fridge
10	oven
11	the lamp
12	the trash can
13	the chair
14	the bed
15	the sofa

Table 2: Template based grammar

A Experimental details

A.1 Environment

A.1.1 Action space

The action space of the environment consists of two modalities (*primitive actions* and *language*) which results in a 3D discrete action vector.

The first dimension corresponds to the primitive actions modality. It consists of 7 actions (turn left, turn right, move forward, pickup, drop, toggle, done). In the TalkItOut task *pickup* and *drop* actions do not do anything and *toggle* and *done* terminate the episode with 0 reward. The reason for this is that we intend to use those actions in the full benchmark.

The second and third dimensions regard the language modality. The second dimension selects a template (four possibilities) and the third a noun (8 possibilities). The full grammar is shown in table 2

Both modalities can also be undefined, in which case no action is taken in the undefined modality. Examples of such actions are shown in table 3.

Action	description
(1, -, -)	moves left without speaking
(1, 1, 5)	moves left and utters "Open the window"
(-, 1, 5)	doesn't move but utters "Open the window"
(-, -, -)	nothing happens

Table 3: Examples of various actions in the environment. Second and third dimension must both either be underfined or not.

A.1.2 State space

The multimodal state space consists of the *vision* modality and the *language* modality.

The *vision* modality is manifested as a *7x7* grid displaying the space in front of the agent (shown as highlighted grids in figure 1). Each location of this grid is encoded as three integers depicting the object type, color and additional information (ex. NPC type: wizard or guide). For example, a blue wizard will be encoded as $(11, 2, 0)$ and a blue guide as $(11, 2, 1)$.

The *language* modality is represented as a string containing the currently heard utterances, i.e. utterances uttered by NPCs next to the agent, and their names (ex. "John: go to the green door"). In case of silence an "empty indicator" symbol is used.

As it is often more convenient to concatenate all the utterances heard, to simplify the implementation of the agent, the implementation of the environment also supports giving the full history of heard utterances with the "empty indicator" symbols removed as additional information.

A.1.3 The task

As discussed in the main text the task consists of three NPCs and four doors. The agent has to find out which door is the correct one by asking the true guide. To find out which guide is the correct one the agent has to ask the wizard. Upon finding out which door is the correct one the agent has to stand in front of it and utter "Open sesame". Then the episode ends and the reward is calculated by the following equation:

$$r_{extr} = 1.0 - 0.9 * \frac{t}{t_{max}} \quad (1)$$

, where t is the number of steps agent made in the environment and $t_{max} = 40$ is the maximum allowed number of steps. If the agent executes *done*, *toggle* or utters "Open sesame" in front of the wrong door the episode ends with no reward.

True guide: John Correct door color: blue
agent goes to the wizard
Agent: Where is the exit?
Wizard: Ask John.
agent goes to one guide
Agent: Where is the exit?
Jack: Go to the red door.
agent goes to the other guide
Agent: Where is the exit?
John: Go to the blue door.
agent goes to the blue door
Agent: Open sesame

Table 4: An example of a successful episode

An example of a dialogue that might appear in a successful episode is shown in table 4

For each episode the colors of doors and NPCs are selected randomly from a set of six and the names of the two guides are selected randomly from a set of two (Jack, John). Furthermore, the grid width and height are randomized from the minimal size of 5 up to 8 and the NPCs and the agent are placed randomly inside (omitting locations in front of doors).

A.1.4 Required social skills

In this section we will discuss the TalkItOut environment in the context of social skills required of the agent. The upcoming *SocialAI* benchmark will contain various environments each specialized for testing different social skills i.e. some will be specialized for multimodality and others for ToM or Social games.

Intertwined multimodality

To solve this task the agent must use both modalities both in the action and in the observation space. Furthermore, this multimodality is intertwined because the progression in which the modalities are used is non-trivial. To discuss this notion further let's imagine an example of command following. The progression of modalities here is trivial because the agent always *listens* for the command first and then *looks* and *moves/acts* to complete the task. Another good example is embodied question answering. Here the agent again always first *listens* to the question, then *looks* and *moves* in the environment to finally, at the end, *speak* the answer.

In our environment, however, the agent must choose which modality to use based on the current state. And it will often be required to switch between modalities many times. For example, to talk to an NPC the agent first *looks* to find the NPC, then it *moves* to the NPC, finally the agent *speaks* to it and *listens* to the response. This progression is then used, if needed, for other NPCs, and finally a similar one used to go to the correct door and open it. Furthermore, depending on the current configuration of environment, the progression can also be different. Usually, after finding out the correct door the agent needs to *look* for it and *move* to it to *speak* the password, but if the true guide is already next to the correct door only *looking* for the door and *speaking* the password is required.

Theory of Mind

Since the agent must be able to infer good or bad intentions of other NPCs, a basic form of ToM is needed. Primarily, the agent needs to infer that the wizard is well-intended, wants to help, and is therefore trustworthy. Using the inferred trust in the wizard it is possible to infer the good intentions of the true guide, and likewise the bad intentions of the false guide.

On the other hand, as the false guide chooses which false direction to give each time asked, it is also possible to infer its ill-intentions by asking him many times in the same episode and observing this inconsistency. If an NPC gives different answers for the same question in the same episode then it is evident its intentions are bad.

Social games

Since social games were not the focus of this environment, and will be studied in more detail in the upcoming environments, they are present in this environment only in a simple form. To talk with an NPC the agent needs to stand in next to it, to get an answer the agent needs to ask "where is the exit". These simple rules (a.k.a. social conventions) are social games i.e. grammars describing the possible and impossible interactions. It is impossible to communicate if you are far and get directions if you ask "Where is the floor". The agent needs to be able to extract these rules and use them in relation to all the NPCs.

A.2 Baselines details

BabyAI baseline In this work we use a PPO-trained (Schulman et al., 2017) DRL architecture initially designed for the BabyAI benchmark (Chevalier-Boisvert et al., 2018a). The policy design was improved in a follow-up paper by Hui

Hyperparameter	value
learning rate	$1e^4$
GAE λ	0.99
clip ϵ	$1e^5$
batch size	1280
γ	0.99
recurrence	10
epochs	4
expl. bon. C	0.125
expl. bon. M	50

Table 5: Training hyperparametres

et al. (2020) (more precisely, we use their *original_endpool_res* model). See figure 2 for a visualization of the complete architecture. First, symbolic pixel grid observations are fed into two convolutional layers (LeCun et al., 1989; Krizhevsky et al., 2012) (3x3 filter, stride and padding set to 1), while dialogue inputs are processed using a Gated Recurrent Unit layer (Chung et al., 2015). The resulting image and language embeddings are combined using two FiLM attention layers (Perez et al., 2017). Max pooling is performed on the resulting combined embedding before being fed into an LSTM (Hochreiter and Schmidhuber, 1997) with a $128D$ memory vector. The LSTM embedding is then used as input for the navigation action head, which is a two-layered fully-connected network with tanh activations and has an 8D output (i.e. 7 navigation actions and no_op action).

In order for our agent to be able to both move and talk, we add to this architecture a talking action head, which is composed of three networks. All of them are two-layered, fully-connected networks with tanh activations, and take the LSTM's embedding as input. The first one is used as a switch: it has a one-dimensional output to choose whether the agent talks (output > 0.5) or not (output < 0.5). If the agent talks, the two other networks are used to respectively sample the template (4D output) and the word (16D output).

Note that the textual input given to the agent consists of the full dialogue history (without the "empty string" indicator) as we found it works better than giving only current utterances (see figure 3b).

Exploration bonus The exploration bonus we use is inspired by recent works in intrinsically motivated exploration (Pathak et al., 2017; Savinov

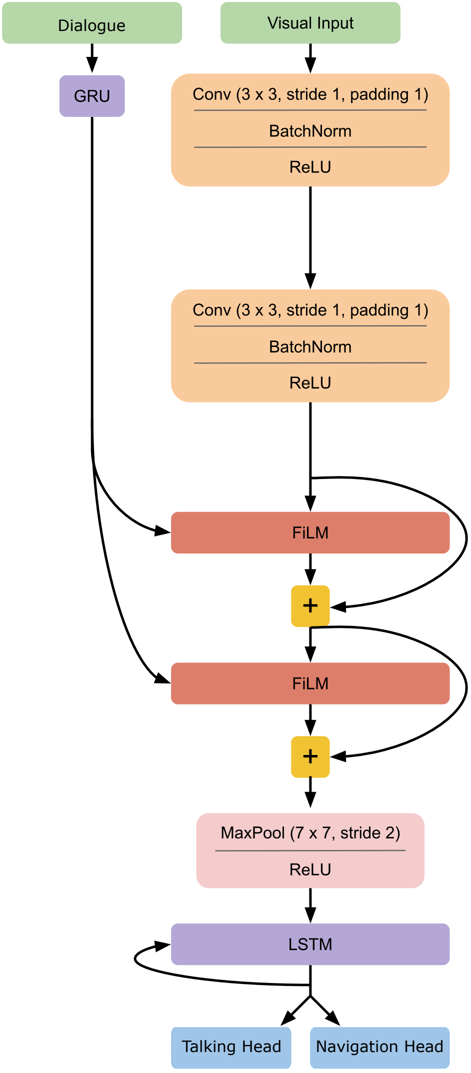


Figure 2: Our Multi-Headed BabyAI baseline DRL agent. Architecture visualization is a modified version of the one made by Hui et al. (2020). We perform two modifications: 1) Instead of fixed instruction inputs our model is fed with NPC's language outputs (if the agent is near an NPC), and 2) We add a language action head, as our agent can both navigate and talk.

et al., 2018; Tang et al., 2017). These intrinsic rewards estimate novelty of the currently observed state and add the novelty based bonus to the extrinsic reward.

In this work we study a multi modal state space and we calculate the exploration bonus only on the language modality. We count how many times was each utterance observed and compute an additional bonus based on the following equation:

$$r_{intr} = \frac{C}{(N(s_{lang}) + 1)^M} \tag{2}$$

, where M and C are hyperparameters and $N(s_{lang})$ is the number of times the utterance s_{lang} was observed during this episode.

We make our reward episodic by resetting the counts at the end of each episode. In the current version of the environment the agent cannot hear his own utterances and the NPCs speak only when spoken to. Therefore, this exploration bonus can be seen as analogous to social influence (Jaques et al., 2019) in the language modality, as the reward is given upon *making the NPC respond*.

Our verbal episodic intrinsic reward, which uses only the language modality, is a good example of a bias that had to be discovered for training social agents.

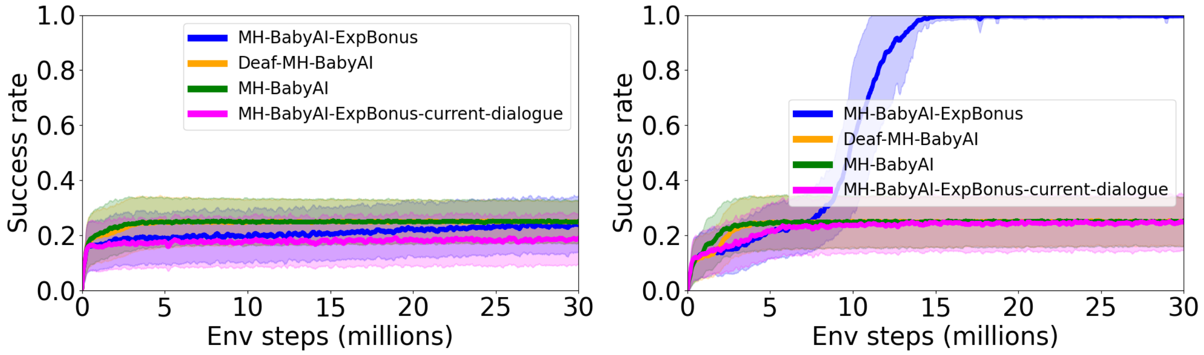
B Additional experiments

In this section we will discuss some additional experiments we ran on the two environments.

Figure 3 shows the success rates for the configurations discussed in the main text and displayed in table 1. One additional configuration shown in this figure is the one denoted "MH-BabyAI-ExpBonus-current-dialogue". In this configuration instead of giving the agent the full dialogue history only the dialogue observed in the current timestep or, if no dialogue is observed, an "empty" indicator string ("NA") is given. It is clearly visible that this configuration is inferior to the one providing the agent with the full dialogue history.

Furthermore, we ran some experiments varying the architecture of the network. These experiments are visible in figure 4. In this figure the "no-mem" refers to the network lacking the final LSTM layer. However, this network still has some form of memory as the full dialogue history is fed into the GRU unit. We can see that, without the LSTM, the agent is not able to solve the ablation environment. We likewise ran experiments where we replaced the GRU unit with a bidirectional-GRU unit ("bigru"), and where we used attention on top of that GRU unit ("attgru")[2]. We also experimented with a different approach of representing the vision modality: a BOW based embedding as used in (Hui et al., 2020). In this approach each grid is represented as a BOW and this representaiton used to retrieve the embedding from a trainable lookup table. We can see that these architectures, like the basic one with the GRU, are able to solve the environment without the liar NPC, but not the full environment.

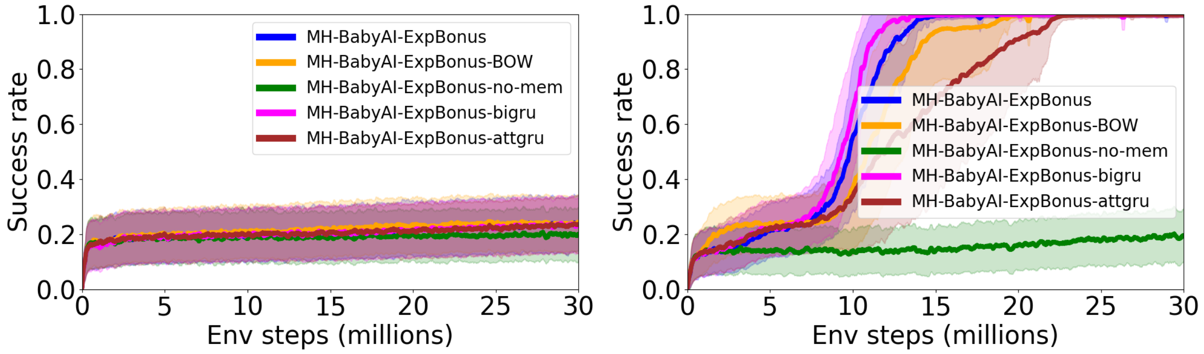
[2] The attention vector was computed using a linear layer on the LSTM's hidden state from the previous step.



(a) Full environment

(b) "NO liar NPC" environment

Figure 3: Training configuration experiments



(a) Full environment

(b) "NO liar NPC" environment

Figure 4: Architectural experiments