EMERGENT COMPOSITIONAL COMMUNICATION IN GENERALIZED SIGNALING GAMES

TOMEK KORBAK
INSTITUTE OF PHILOSOPHY OF SOCIOLOGY,
POLISH ACADEMY OF SCIENCES

PLAN OF THE TALK

What is emergent communication?

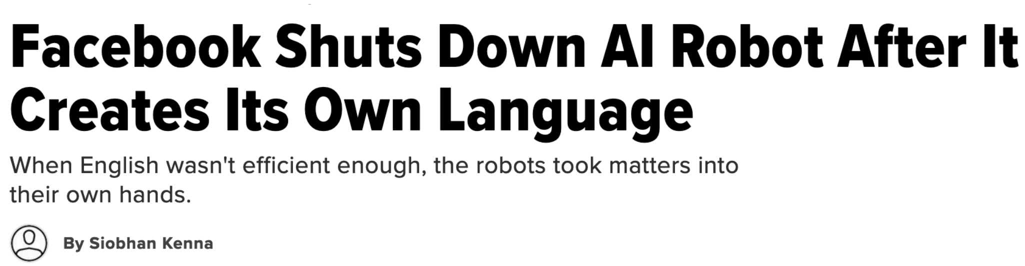
Experimental setup and the problem of inducing compositionality

Template transfer

Results

Discussion

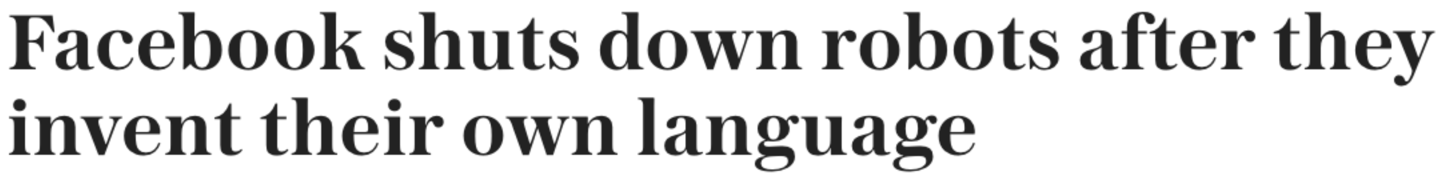
WHAT IS EMERGENT COMMUNICATION?



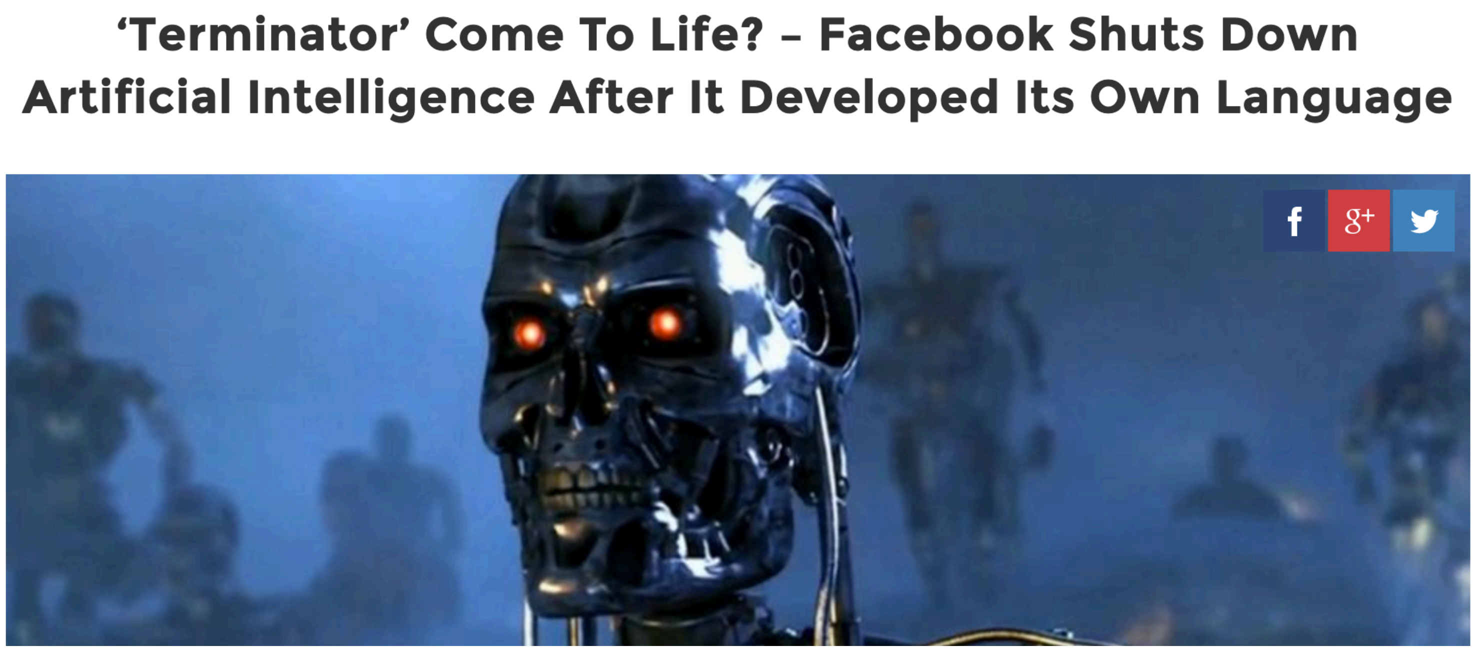
Facebook Shuts Down AI Robot After It Creates Its Own Language

When English wasn't efficient enough, the robots took matters into their own hands.

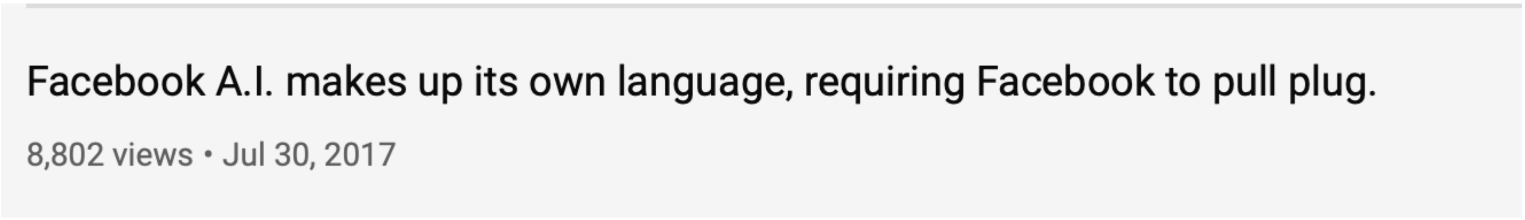
By Siobhan Kenna



Facebook shuts down robots after they invent their own language

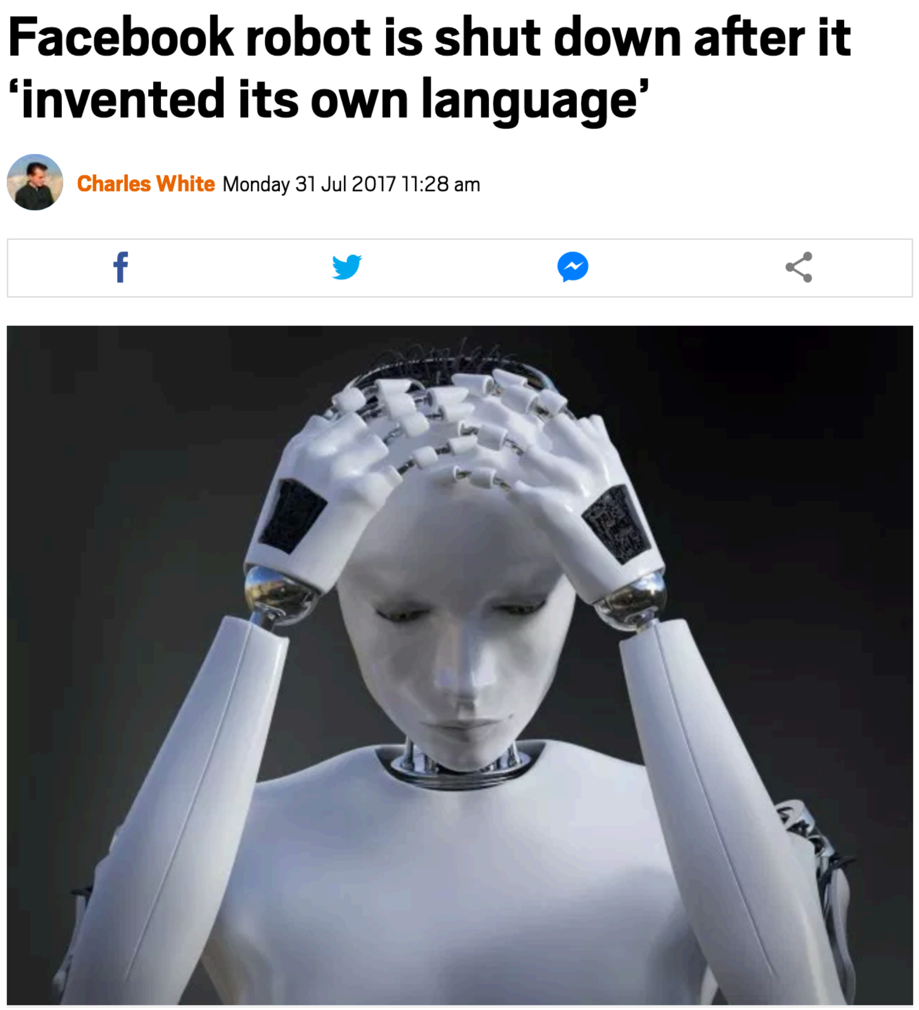


'Terminator' Come To Life? – Facebook Shuts Down Artificial Intelligence After It Developed Its Own Language



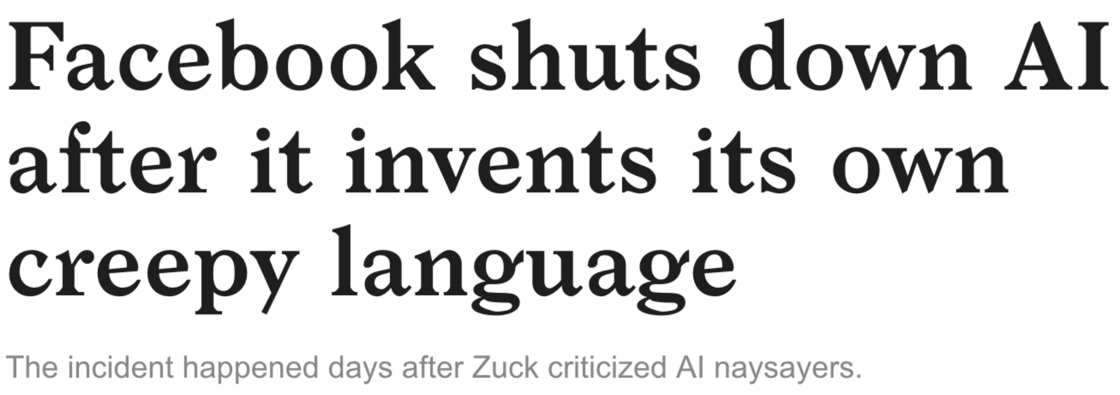
Facebook A.I. makes up its own language, requiring Facebook to pull plug.

8,802 views • Jul 30, 2017



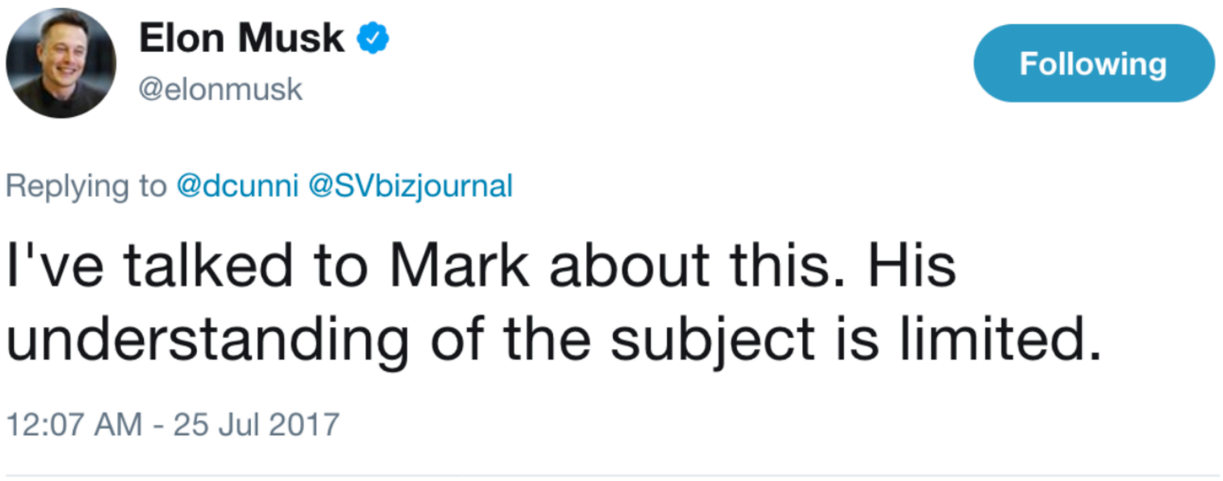
Facebook robot is shut down after it 'invented its own language'

Charles White Monday 31 Jul 2017 11:28 am



Facebook shuts down AI after it invents its own creepy language

The incident happened days after Zuck criticized AI naysayers.



Elon Musk
@elonmusk

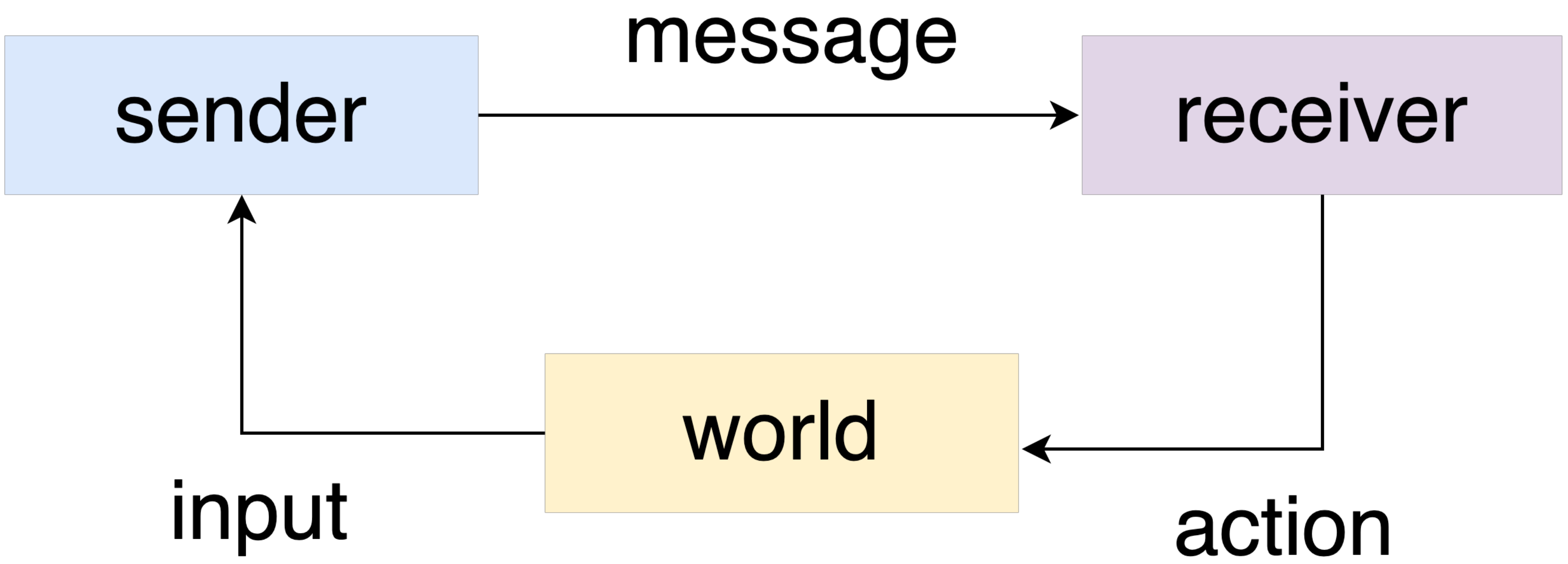
Replying to @dcunni @SVbizjournal

I've talked to Mark about this. His understanding of the subject is limited.

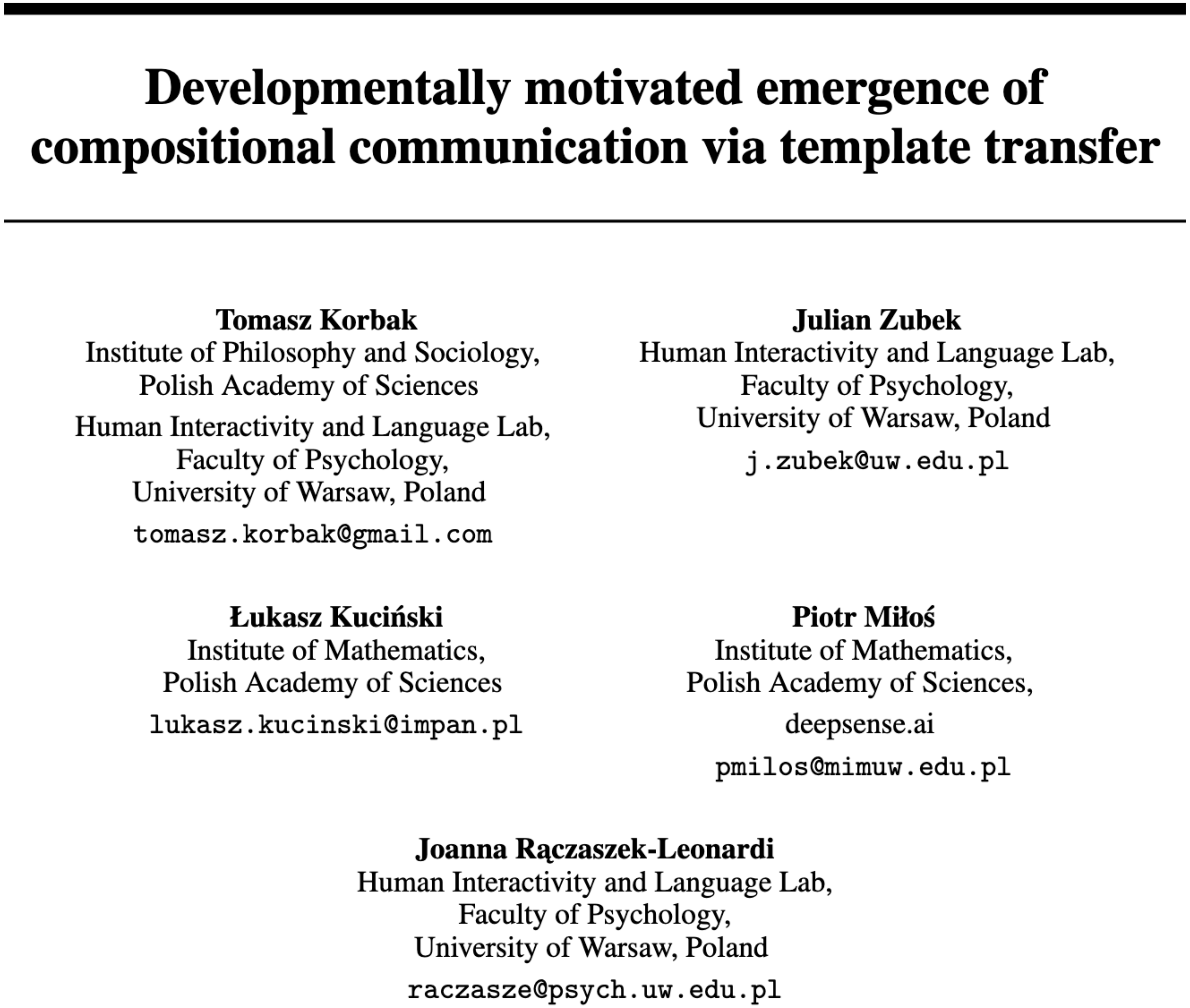
12:07 AM - 25 Jul 2017

LEWIS SIGNALING GAMES

A Lewis signaling game demands a **sender** and a **receiver** to invent a **communication protocol** so that the **receiver can act** based on information only available to the sender and maximize reward for both of them.



PRESENTED APPROACH



Developmentally motivated emergence of compositional communication via template transfer

Tomasz Korbak
Institute of Philosophy and Sociology,
Polish Academy of Sciences
Human Interactivity and Language Lab,
Faculty of Psychology,
University of Warsaw, Poland
tomasz.korbak@gmail.com

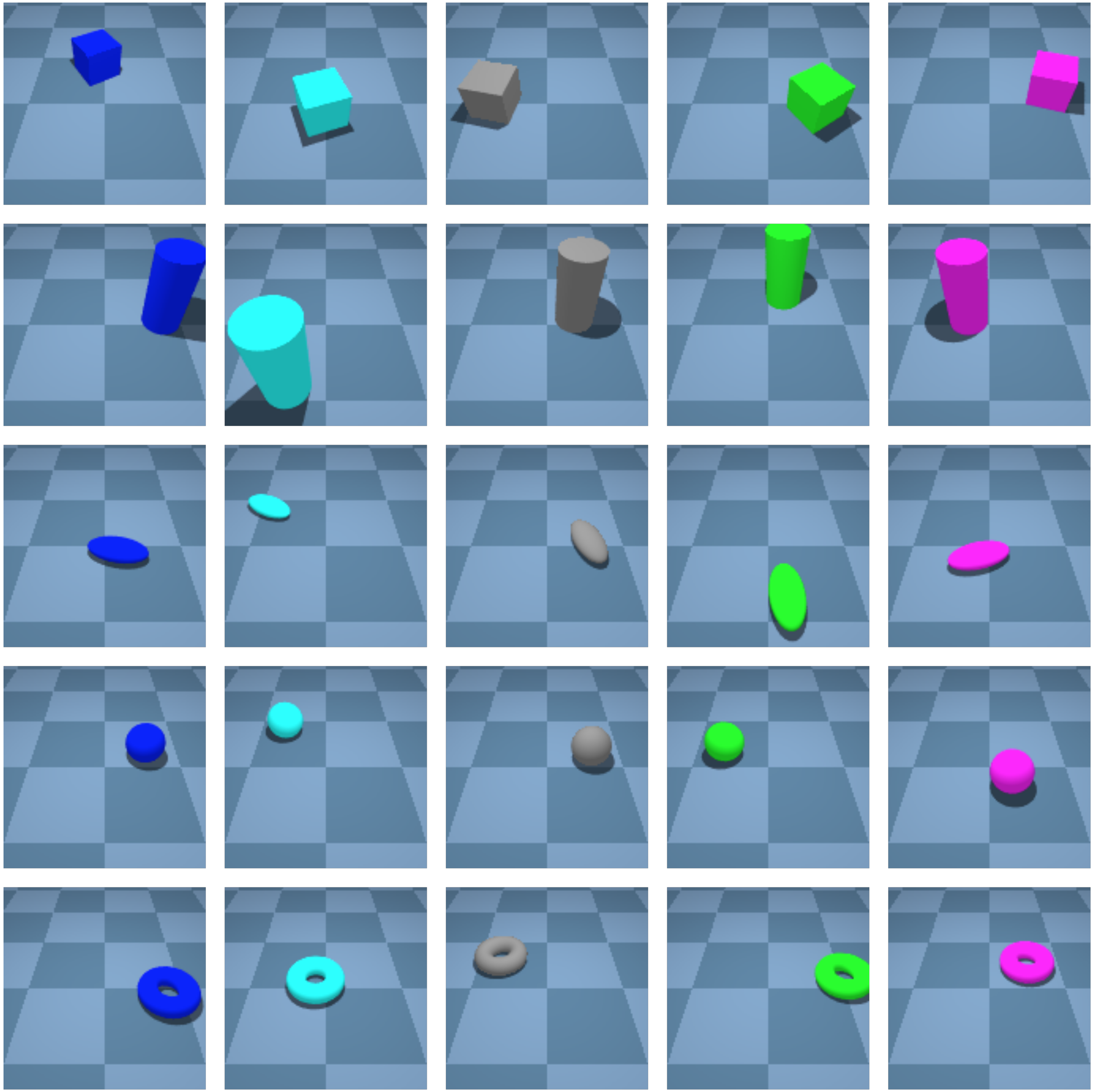
Julian Zubek
Human Interactivity and Language Lab,
Faculty of Psychology,
University of Warsaw, Poland
j.zubek@uw.edu.pl

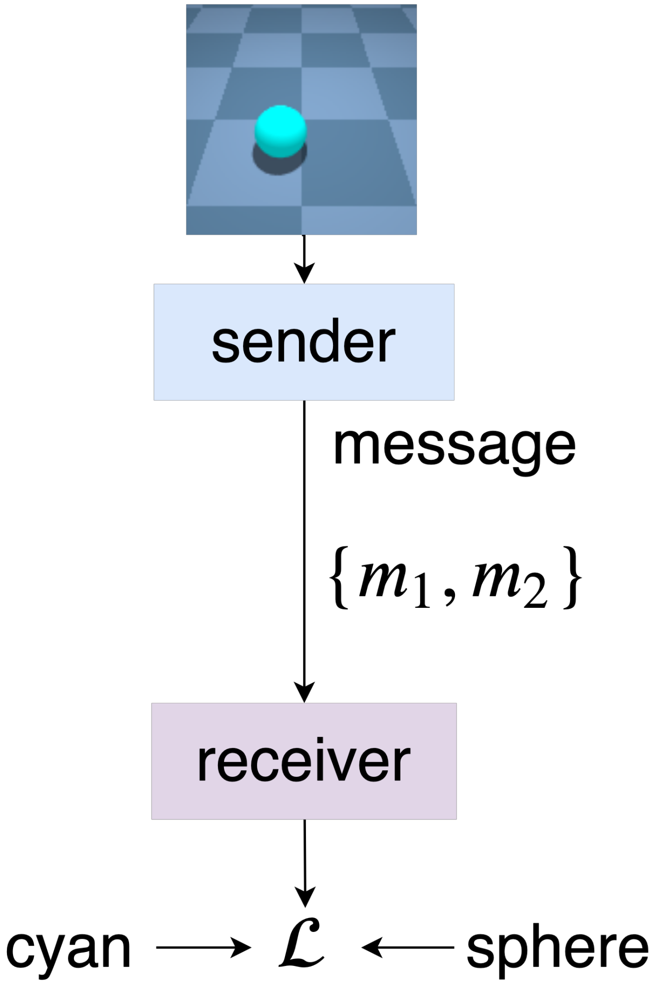
Łukasz Kuciński
Institute of Mathematics,
Polish Academy of Sciences
lukasz.kucinski@impan.pl

Piotr Miłoś
Institute of Mathematics,
Polish Academy of Sciences,
deepsense.ai
pmilos@mimuw.edu.pl

Joanna Rączaszek-Leonardi
Human Interactivity and Language Lab,
Faculty of Psychology,
University of Warsaw, Poland
raczasze@psych.uw.edu.pl

EXPERIMENTAL SETUP





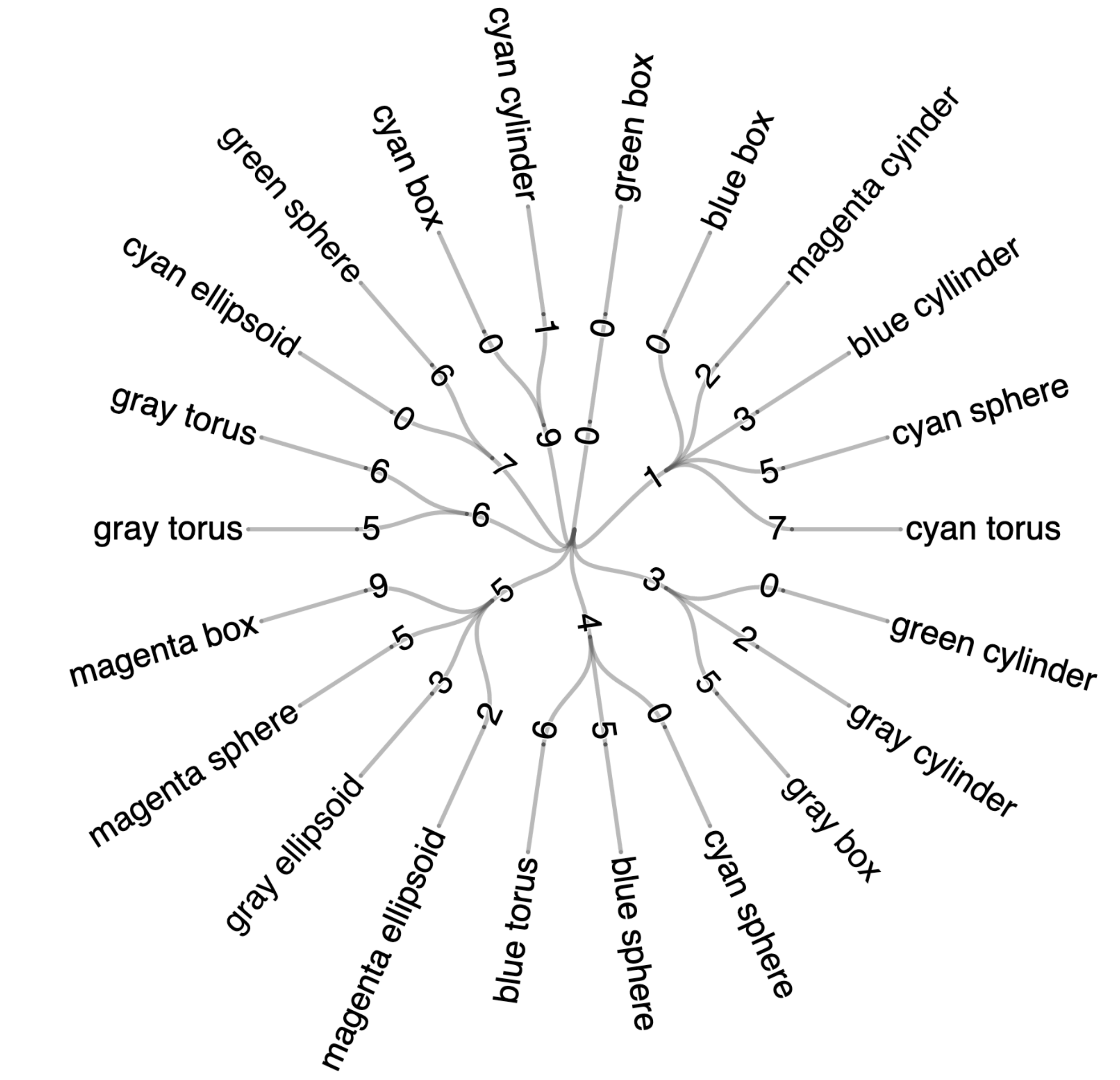
WHEN IS A COMMUNICATION PROTOCOL COMPOSITIONAL?

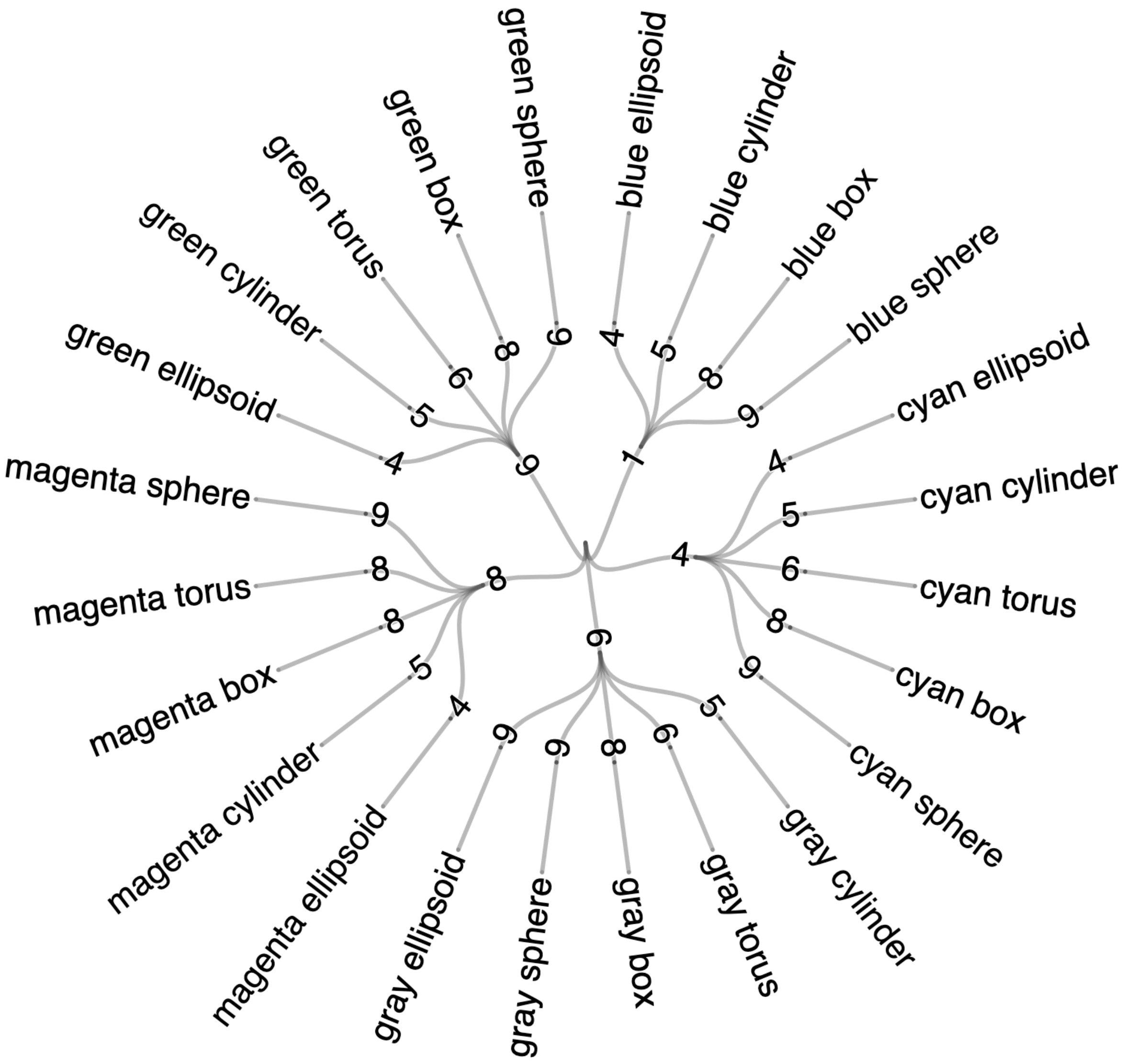
(a) A non-compositional communication protocol

	box	sphere	cylinder	torus	ellipsoid
blue	1 0	4 5	1 3	4 6	5 0
cyan	9 0	9 1	3 0	1 7	7 0
gray	3 5	6 5	3 2	6 6	5 3
green	0 0	7 6	3 0	6 0	7 6
magenta	5 9	5 5	1 2	1 5	5 2

(b) A highly compositional communication protocol

	box	sphere	cylinder	torus	ellipsoid
blue	1 8	1 9	1 5	1 6	1 4
cyan	4 8	4 9	4 5	4 6	4 4
gray	6 8	6 9	6 5	6 6	6 9
green	9 8	9 9	9 5	9 6	9 4
magenta	8 8	8 9	8 5	8 8	8 4



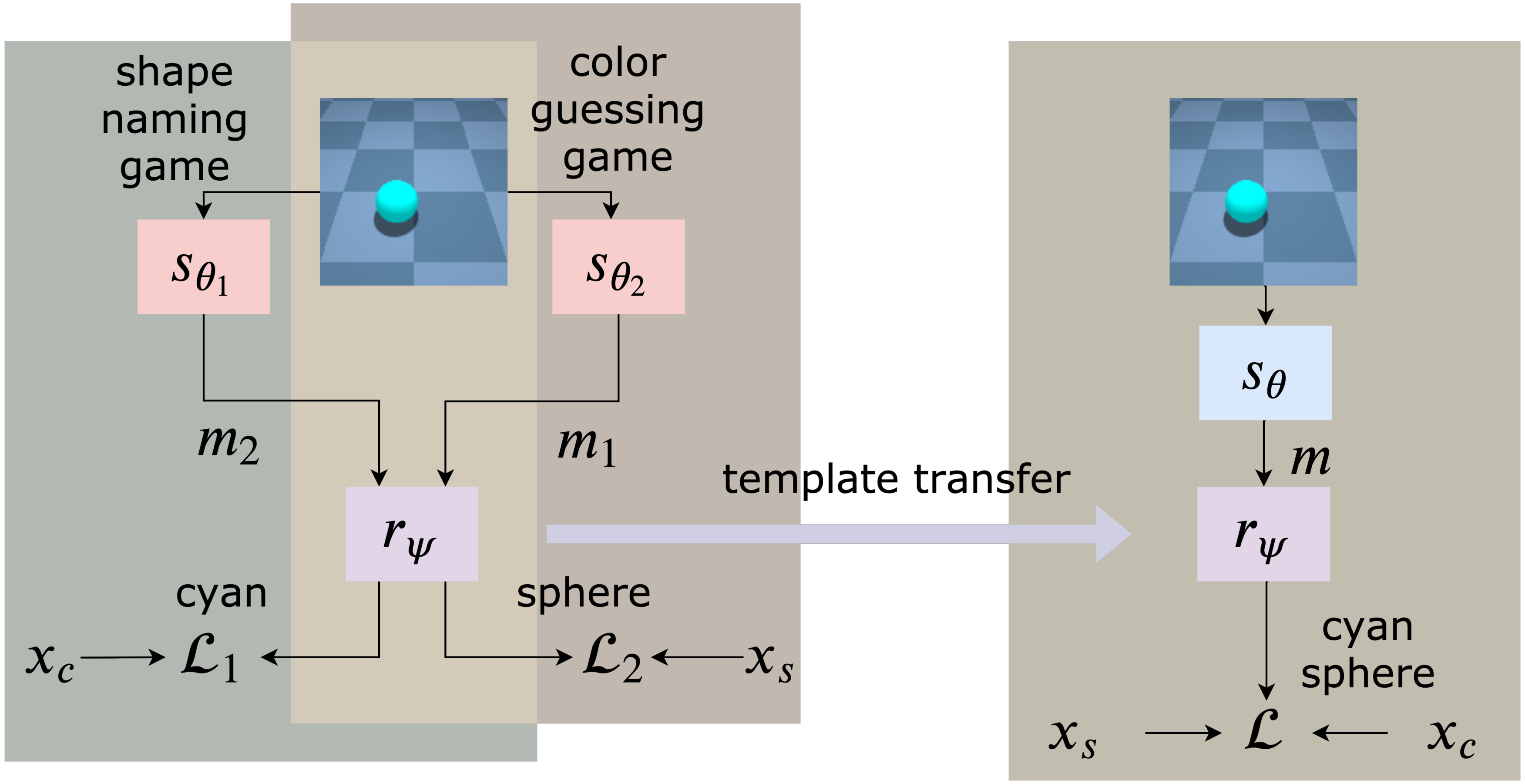


TEMPLATE TRANSFER

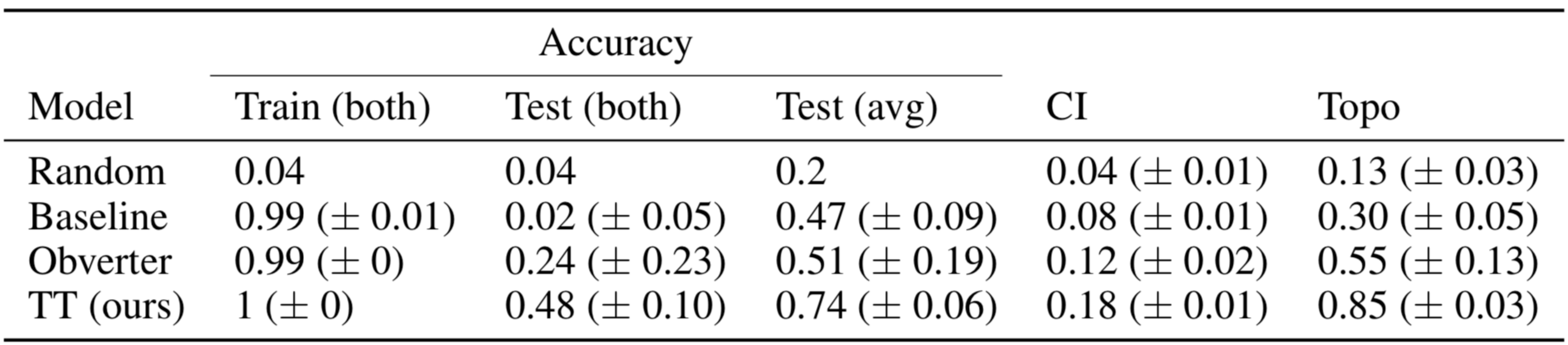
Template transfer means **appropriating** a communication protocol learned in one game **to a new game** (Skyrms & Barrett, 2017).

In our approach, the receiver is pre-trained on **two simpler games**: (i) predicting only the color and (ii) only the shape. The receiver is then **passed** to play the original game with a new sender.

The communication protocol acquired in the first phase serves as a **training bias** in the second phase. The new sender learns to emulate messages sent by the two (color/shape)-specialized senders of the previous phase.



RESULTS

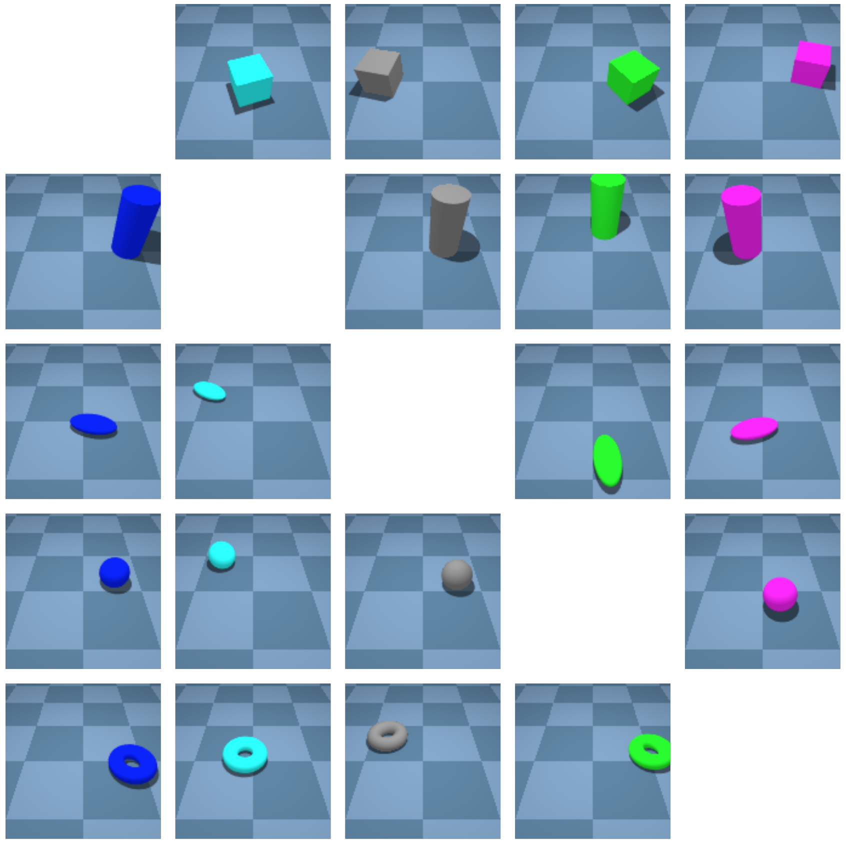


Model	Accuracy Train (both)	Test (both)	Test (avg)	CI	Topo
Random	0.04	0.04	0.2	0.04 (± 0.01)	0.13 (± 0.03)
Baseline	0.99 (± 0.01)	0.02 (± 0.05)	0.47 (± 0.09)	0.08 (± 0.01)	0.30 (± 0.05)
Obverter	0.99 (± 0)	0.24 (± 0.23)	0.51 (± 0.19)	0.12 (± 0.02)	0.55 (± 0.13)
TT (ours)	1 (± 0)	0.48 (± 0.10)	0.74 (± 0.06)	0.18 (± 0.01)	0.85 (± 0.03)

Test accuracy measures the ability to **generalize** to unseen combinations of seen colors and shapes

Context independence (CI) measures **consistency** associating symbols with shapes irrespective of color (and vice versa).

Topographical similarity (Topo) measures the **correlation** between distances over messages and distances over targets.



DISCUSSION

Our procedure:

learning a visual classifier
learning non-compositional protocols
learning a compositional protocol

Peirce’s hierarchy of forms of reference:

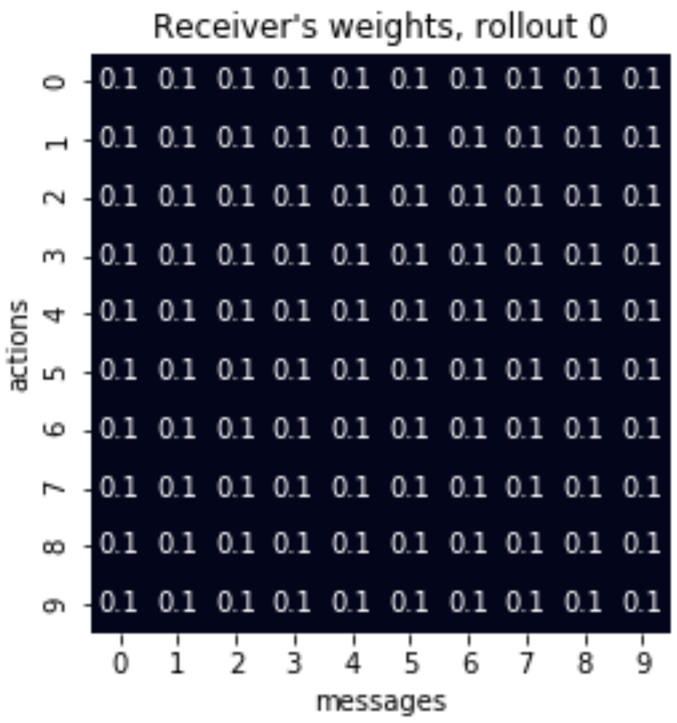
Iconic reference
Indexical reference
Symbolic reference?

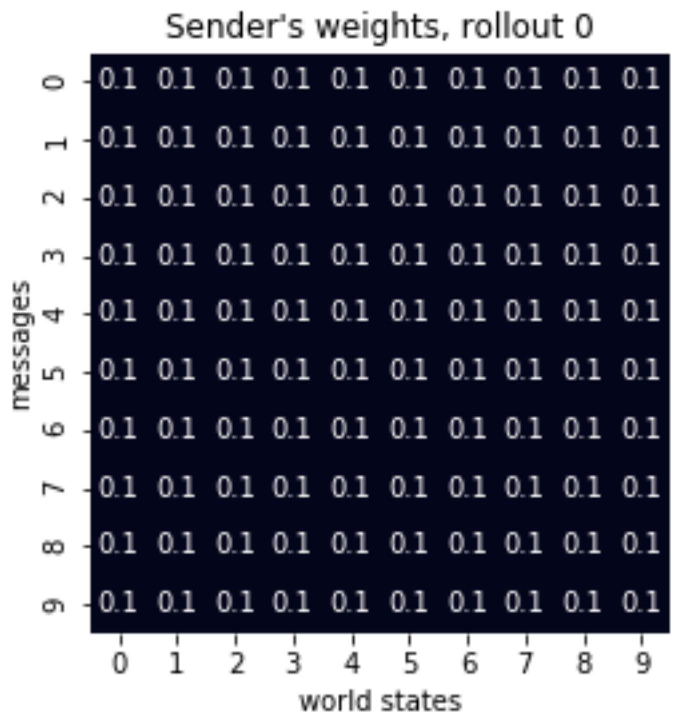
Compositional communication protocols are easier to learn when **bootstrapped** on pre-existing simpler protocols.

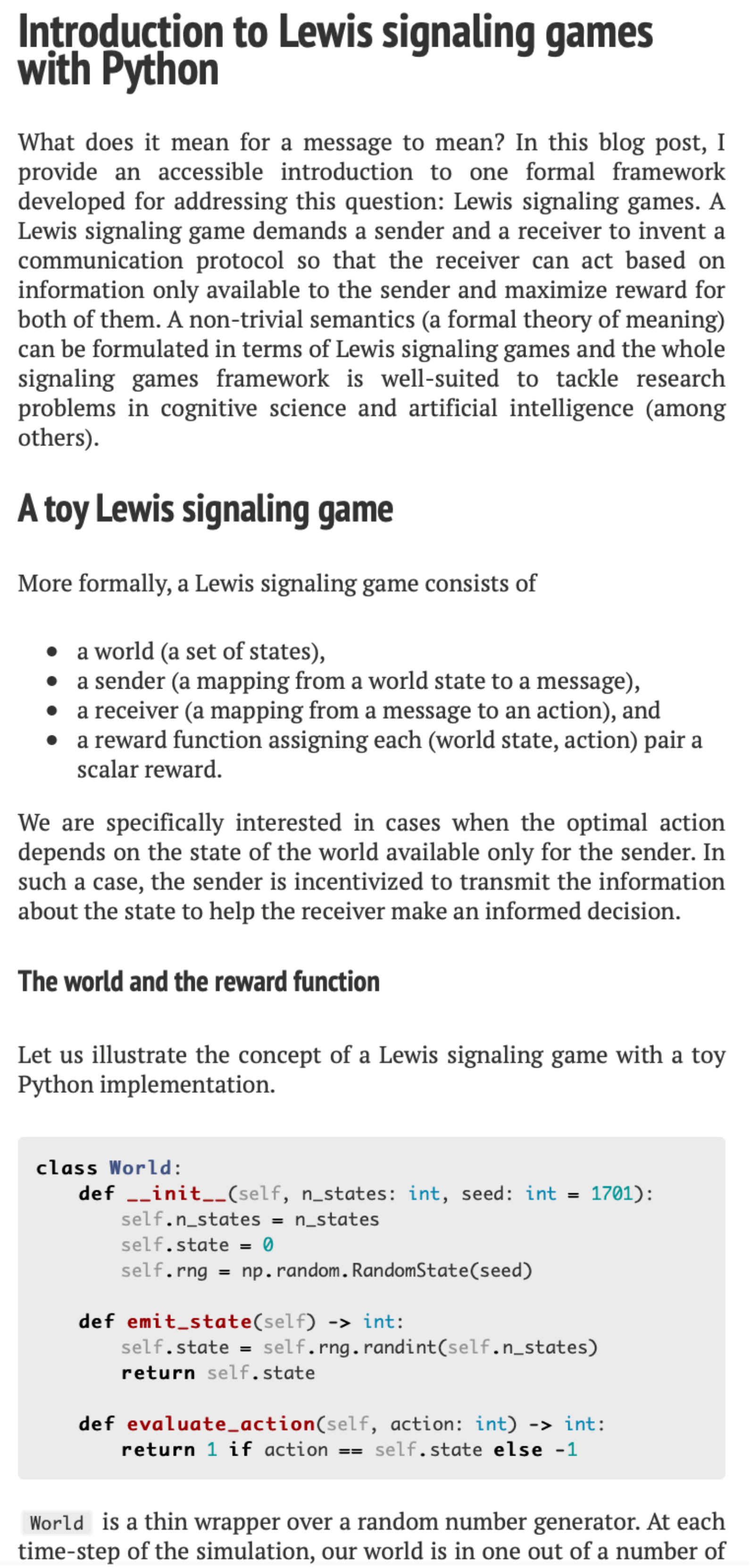
Children does not learn to speak compositionally from **scratch**, but through a series of language games in a rich and highly **structured social environment**.

The ability to communicate compositionality can emerge in a **model-free**, **cognitively undemanding** setting

An introduction to Lewis signaling games with Python examples: https://tomekkorbak.com/2019/10/08/lewis-signaling-games/







Introduction to Lewis signaling games with Python

What does it mean for a message to mean? In this blog post, I provide an accessible introduction to one formal framework developed for addressing this question: Lewis signaling games. A Lewis signaling game demands a sender and a receiver to invent a communication protocol so that the receiver can act based on information only available to the sender and maximize reward for both of them. A non-trivial semantics (a formal theory of meaning) can be formulated in terms of Lewis signaling games and the whole signaling games framework is well-suited to tackle research problems in cognitive science and artificial intelligence (among others).

A toy Lewis signaling game

More formally, a Lewis signaling game consists of

- a world (a set of states),
- a sender (a mapping from a world state to a message),
- a receiver (a mapping from a message to an action), and
- a reward function assigning each (world state, action) pair a scalar reward.

We are specifically interested in cases when the optimal action depends on the state of the world available only for the sender. In such a case, the sender is incentivized to transmit the information about the state to help the receiver make an informed decision.

The world and the reward function

Let us illustrate the concept of a Lewis signaling game with a toy Python implementation.

```
class World:
    def __init__(self, n_states: int, seed: int = 1701):
        self.n_states = n_states
        self.state = 0
        self.rng = np.random.RandomState(seed)

    def emit_state(self) -> int:
        self.state = self.rng.randint(self.n_states)
        return self.state

    def evaluate_action(self, action: int) -> int:
        return 1 if action == self.state else -1
```

`World` is a thin wrapper over a random number generator. At each time-step of the simulation, our world is in one out of a number of

THANK YOU