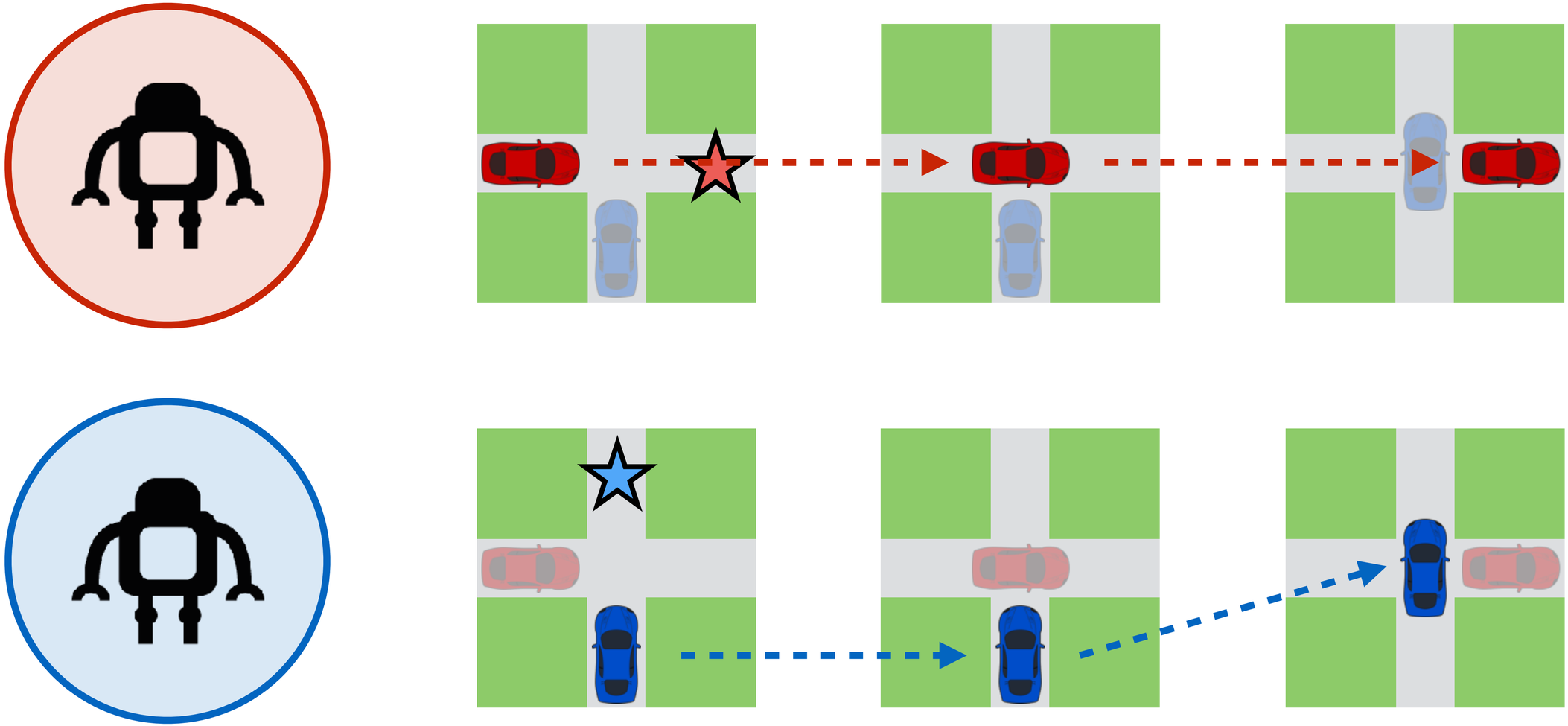
Structure and Interpretation of Neural Codes

Jacob Andreas

Translating Neuralese

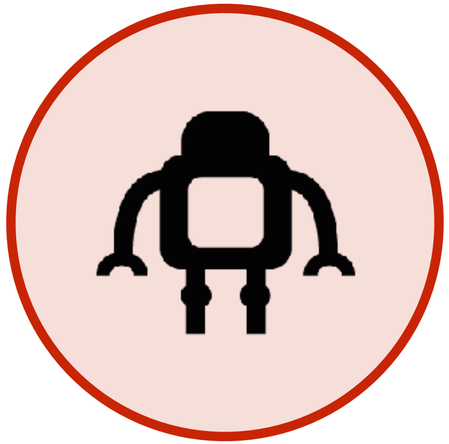
Jacob Andreas, Anca Dragan and Dan Klein

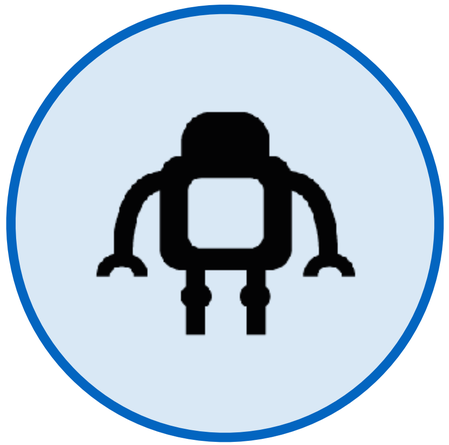
Learning to Communicate

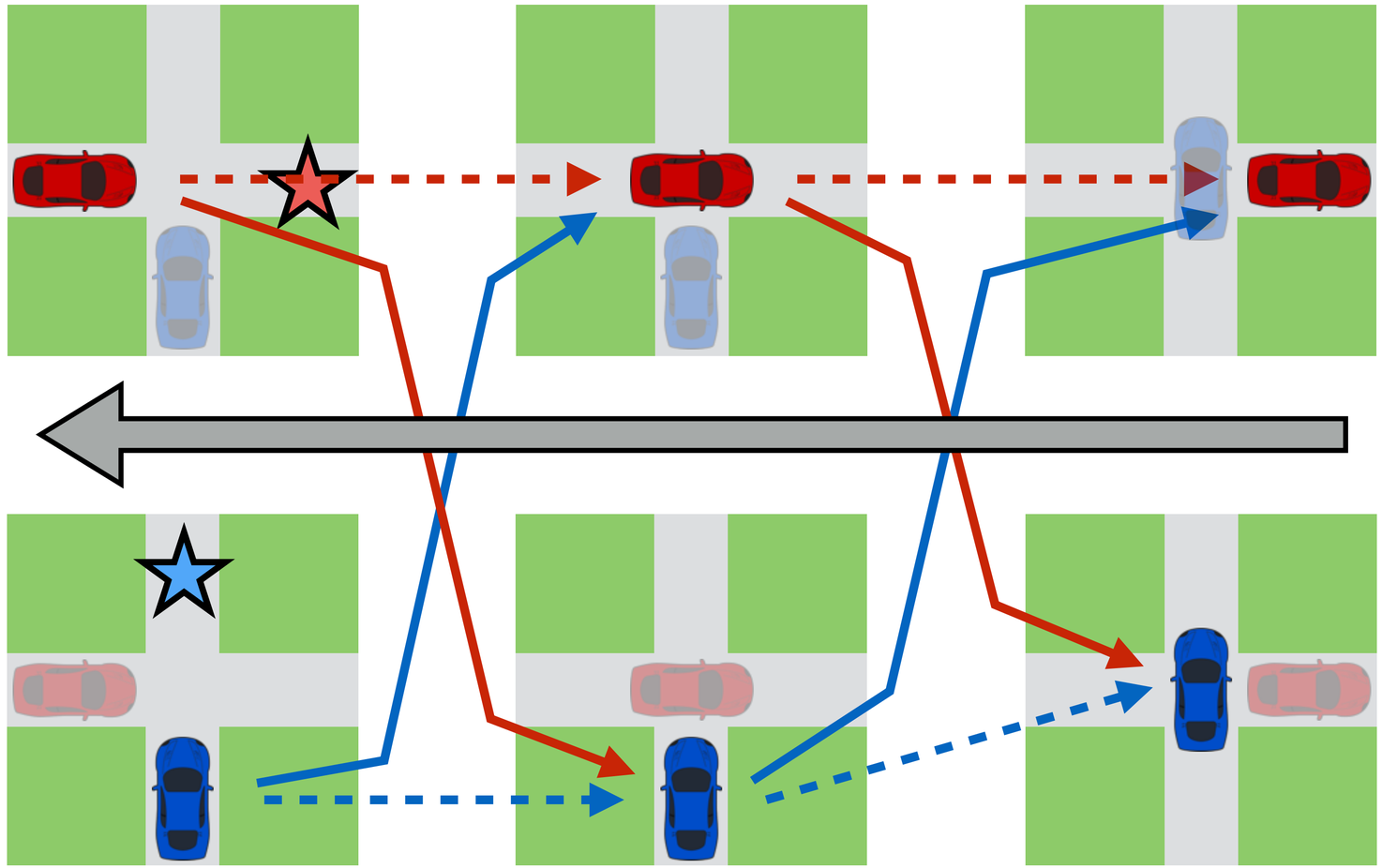


[Wagner et al. 03, Sukhbaatar et al. 16, Foerster et al. 16]

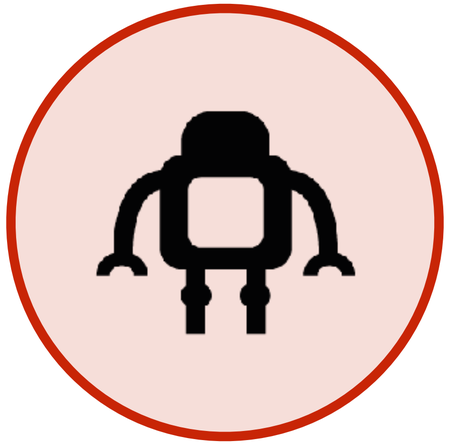
Learning to Communicate

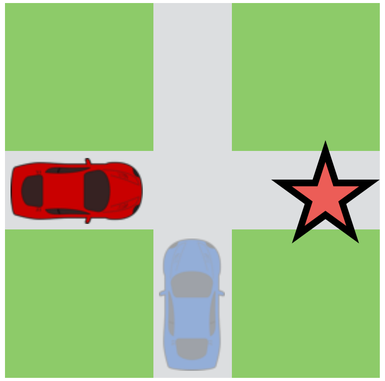


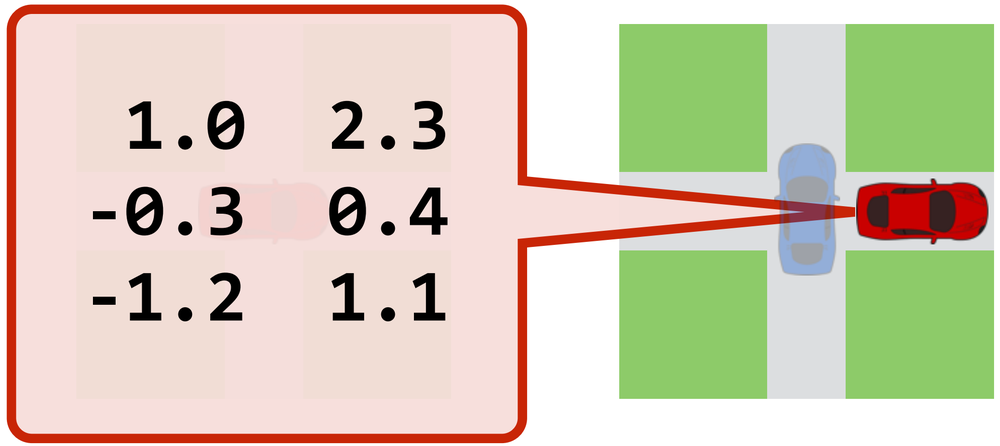


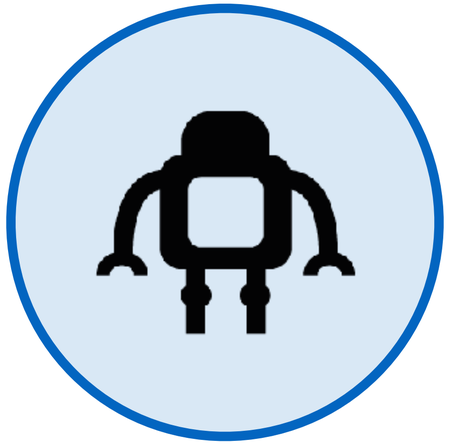


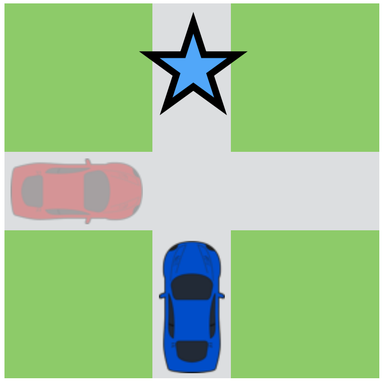
Neuralese

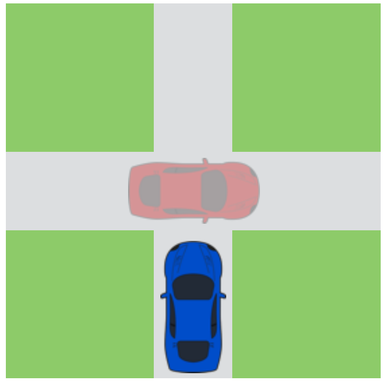


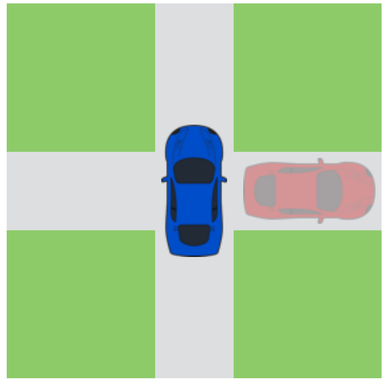




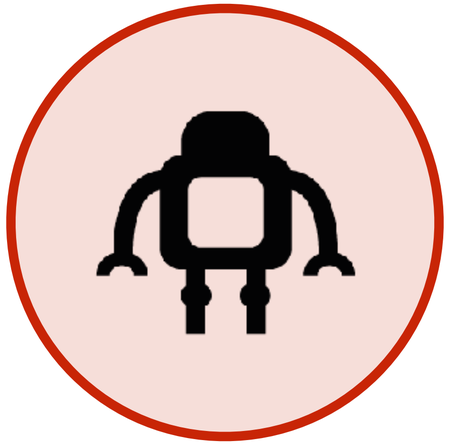


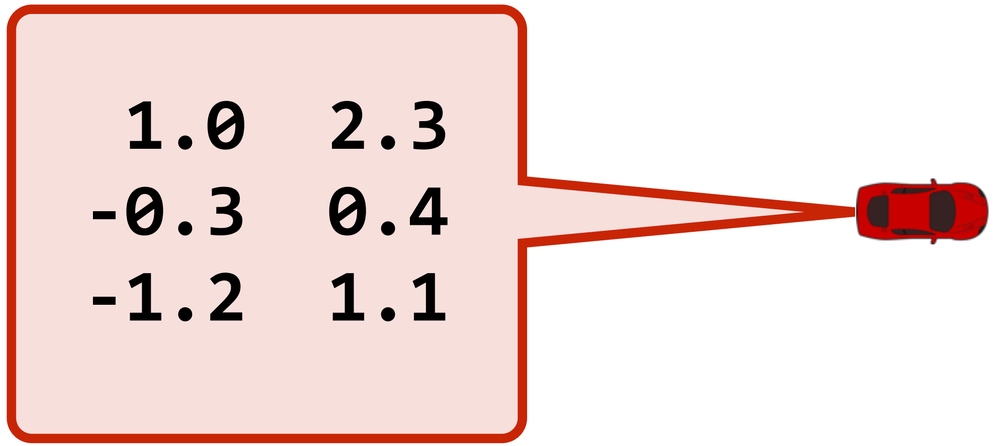






Translating neuralese

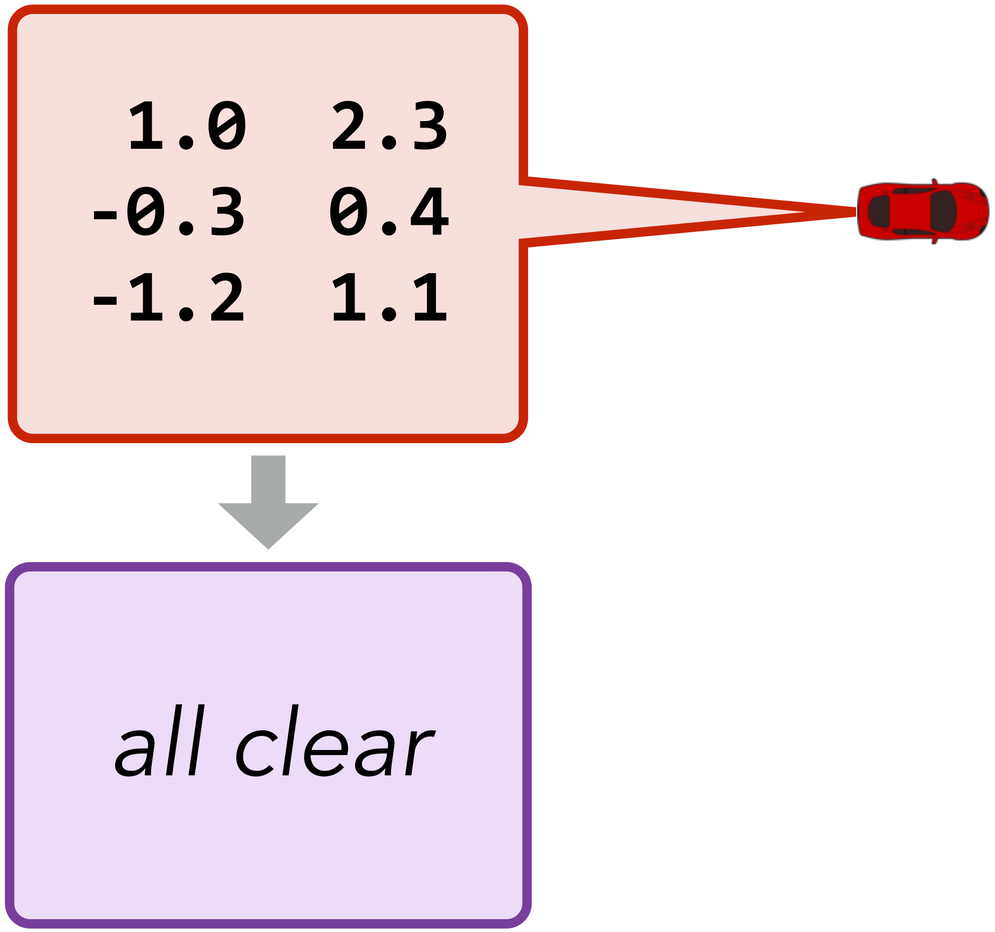




all clear

Translating neuralese

- **Interoperate** with autonomous systems
- **Diagnose** errors
- **Learn** from solutions



[Lazaridou et al. 16]

Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

Outline

Natural language & neuralese

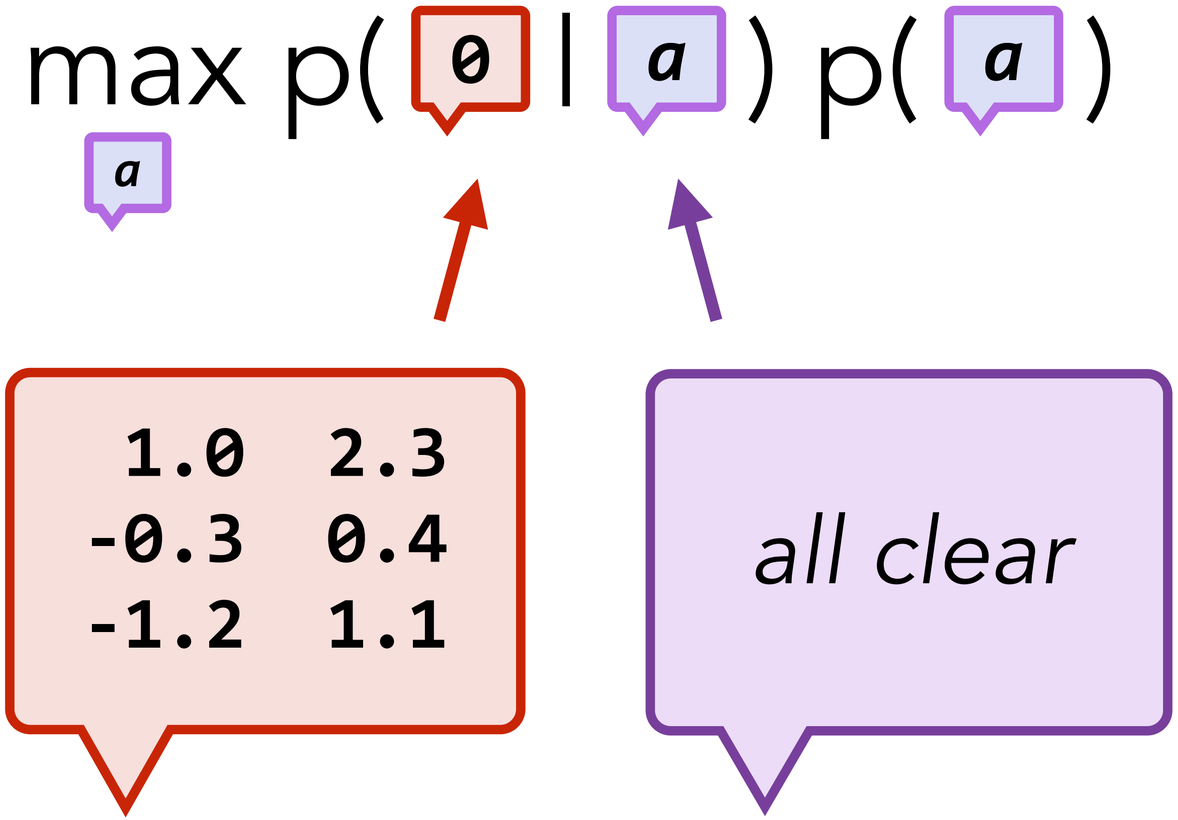
Statistical machine translation

Semantic machine translation

Implementation details

Evaluation

A statistical MT problem



[e.g. Koehn 10]

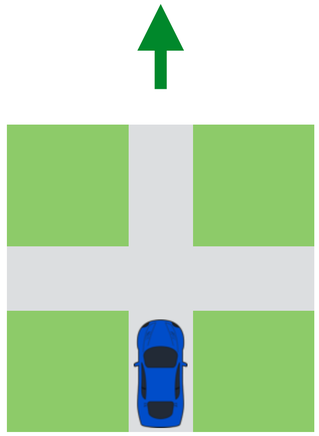
A statistical MT problem

How do we induce a translation model?

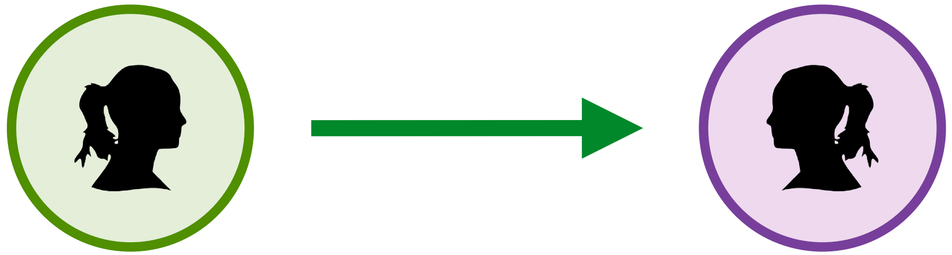
A statistical MT problem

$$\max_{a} p(o \mid a)\, p(a)$$

$$\propto \max_{a} \sum_{\star} p(o \mid \star)\, p(a \mid \star)\, p(\star)$$

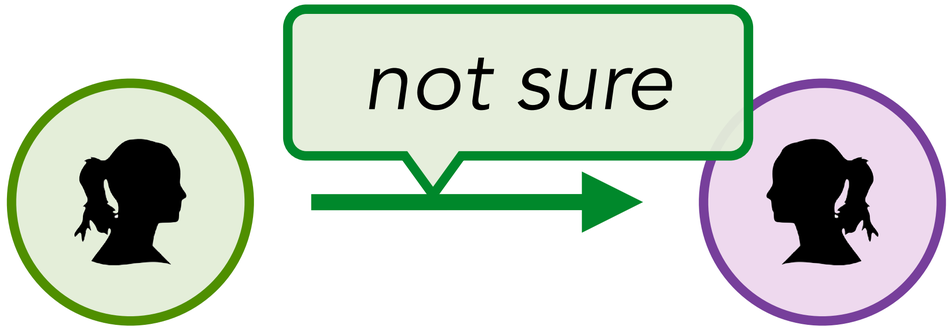


Strategy mismatch



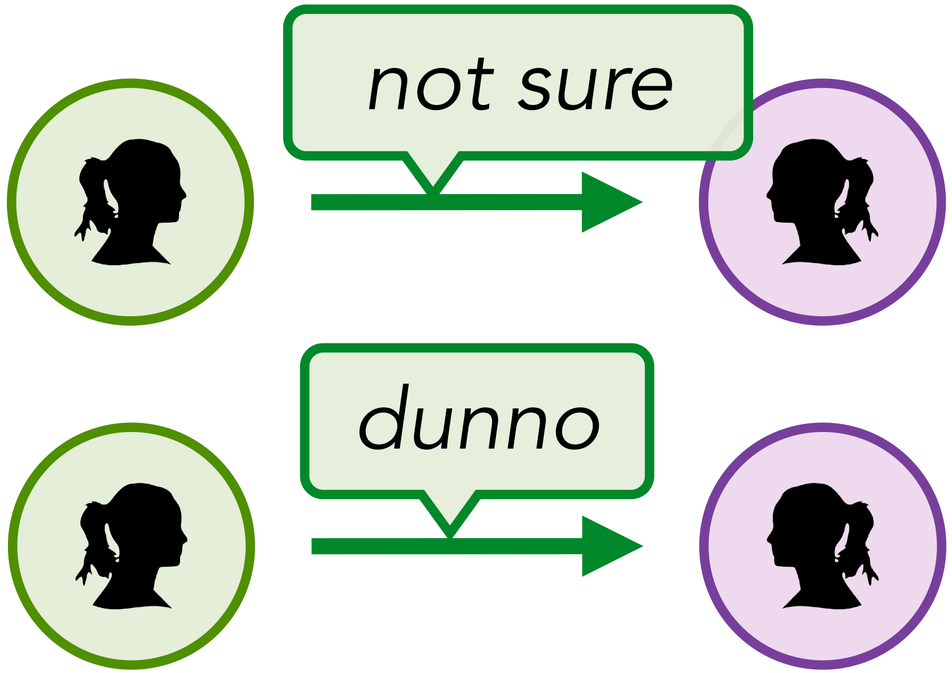
$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{1}{e^x - 1} x^s \frac{\mathrm{d}x}{x}$$

Strategy mismatch

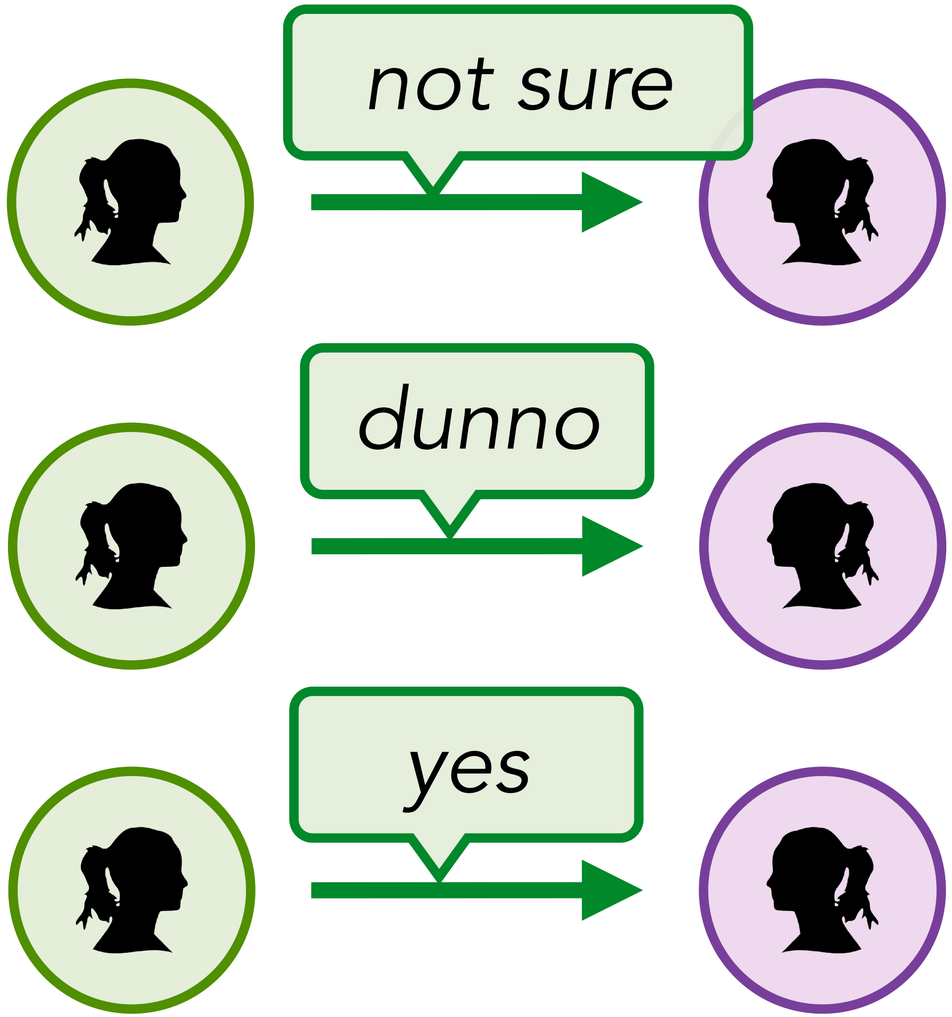


$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{1}{e^x - 1} x^s \frac{\mathrm{d}x}{x}$$

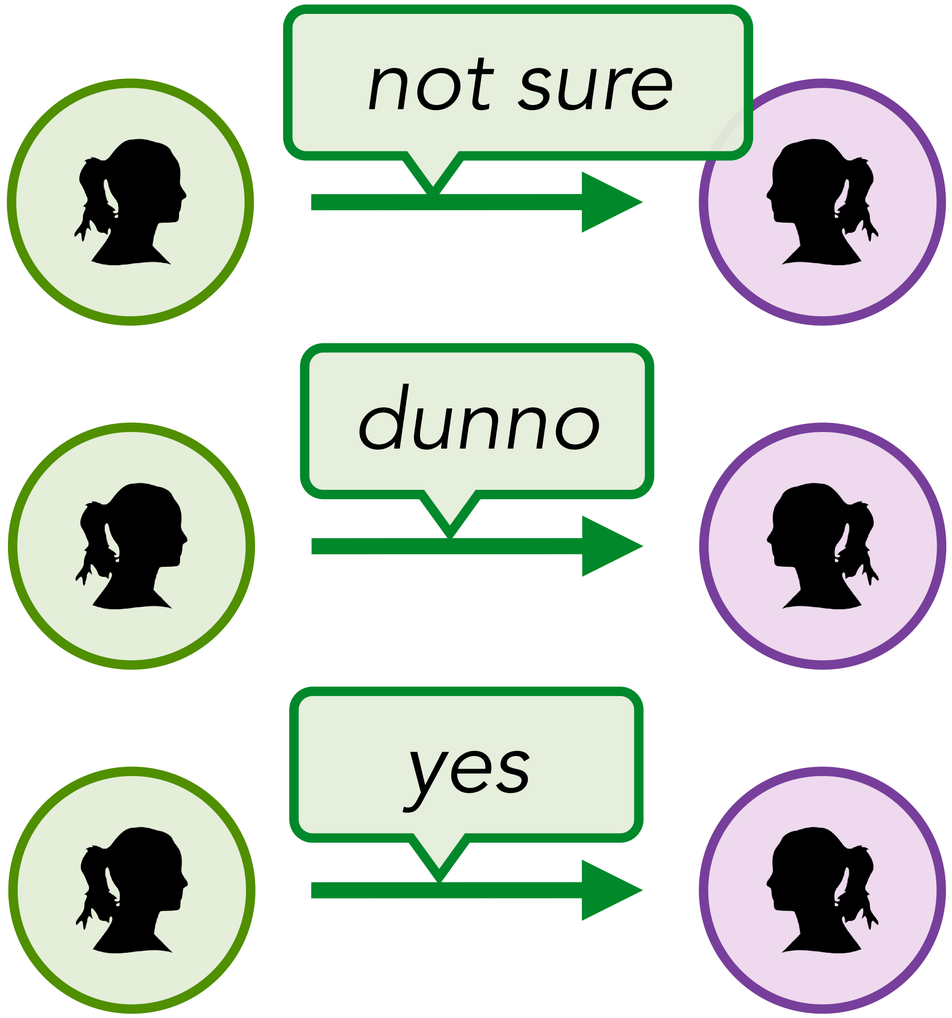
Strategy mismatch

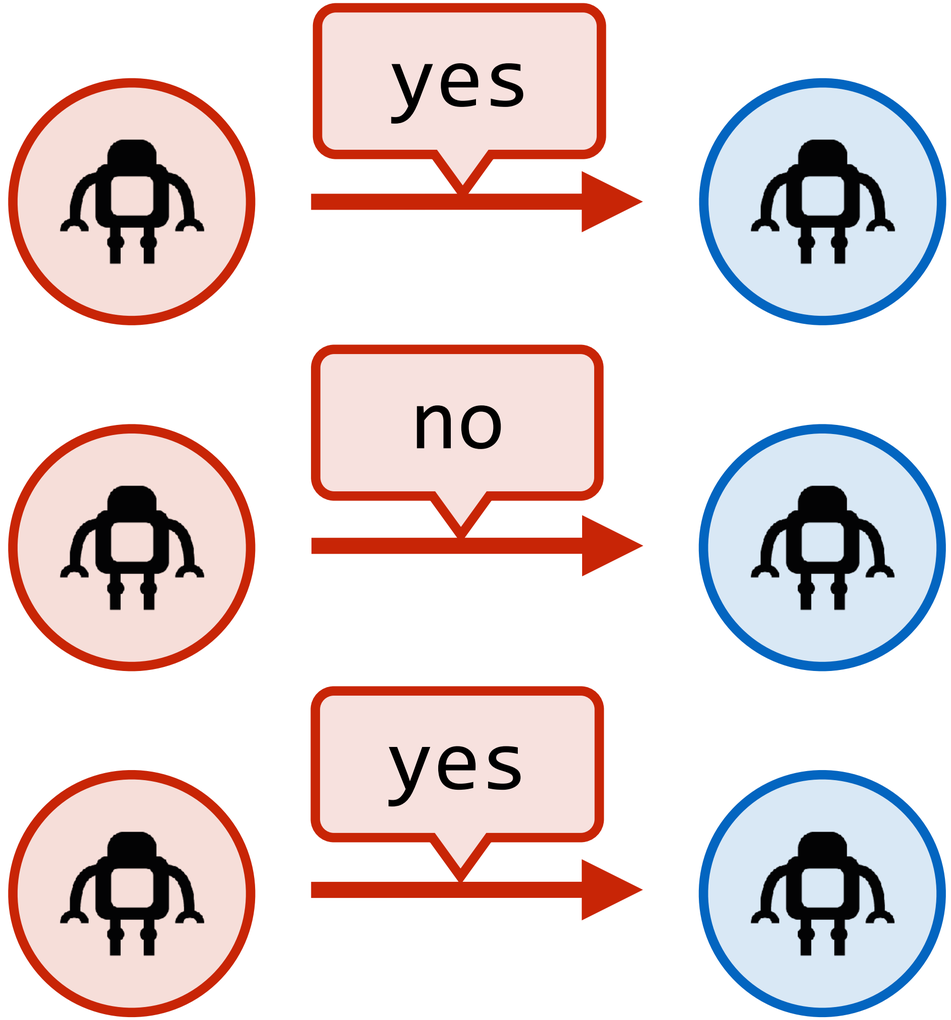


Strategy mismatch

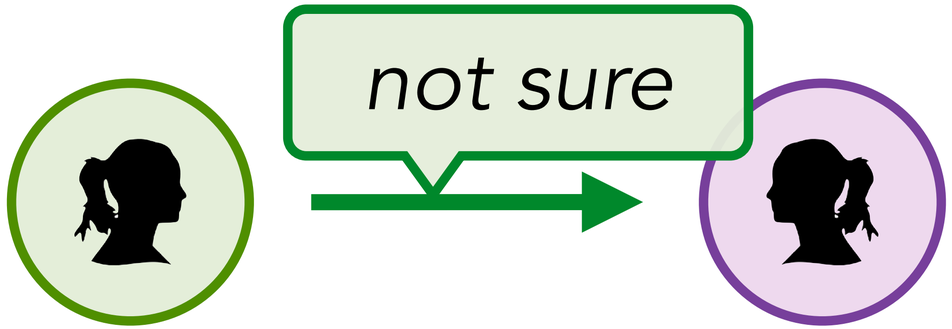


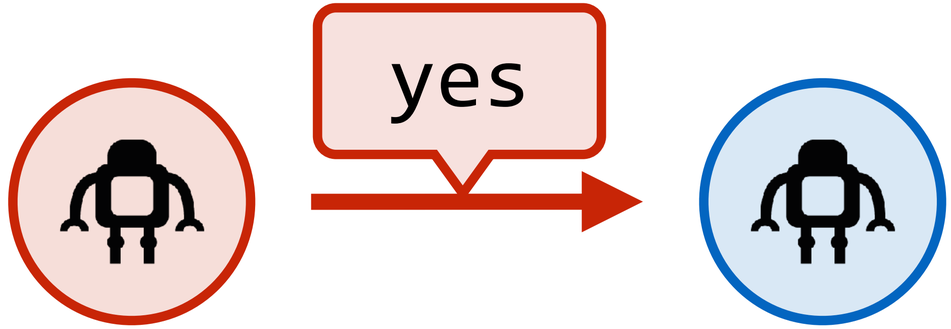
Strategy mismatch





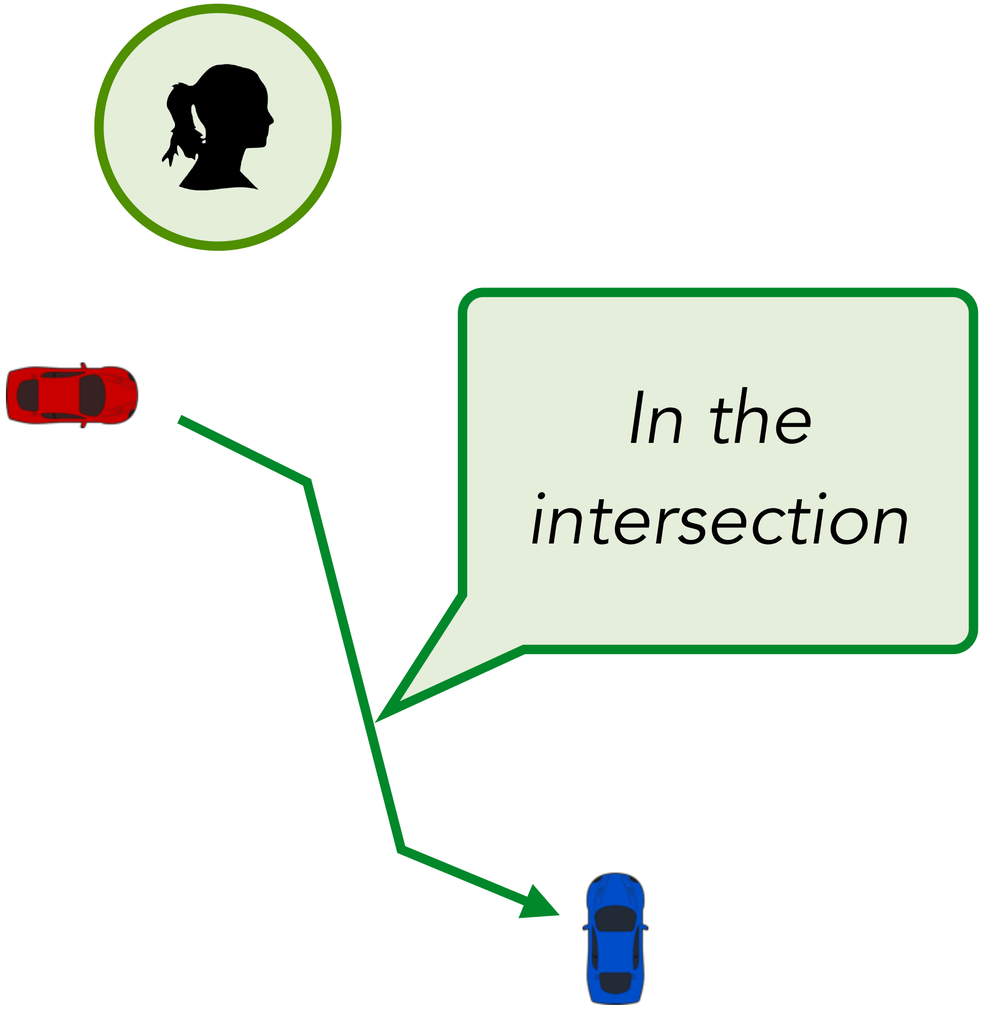
Strategy mismatch

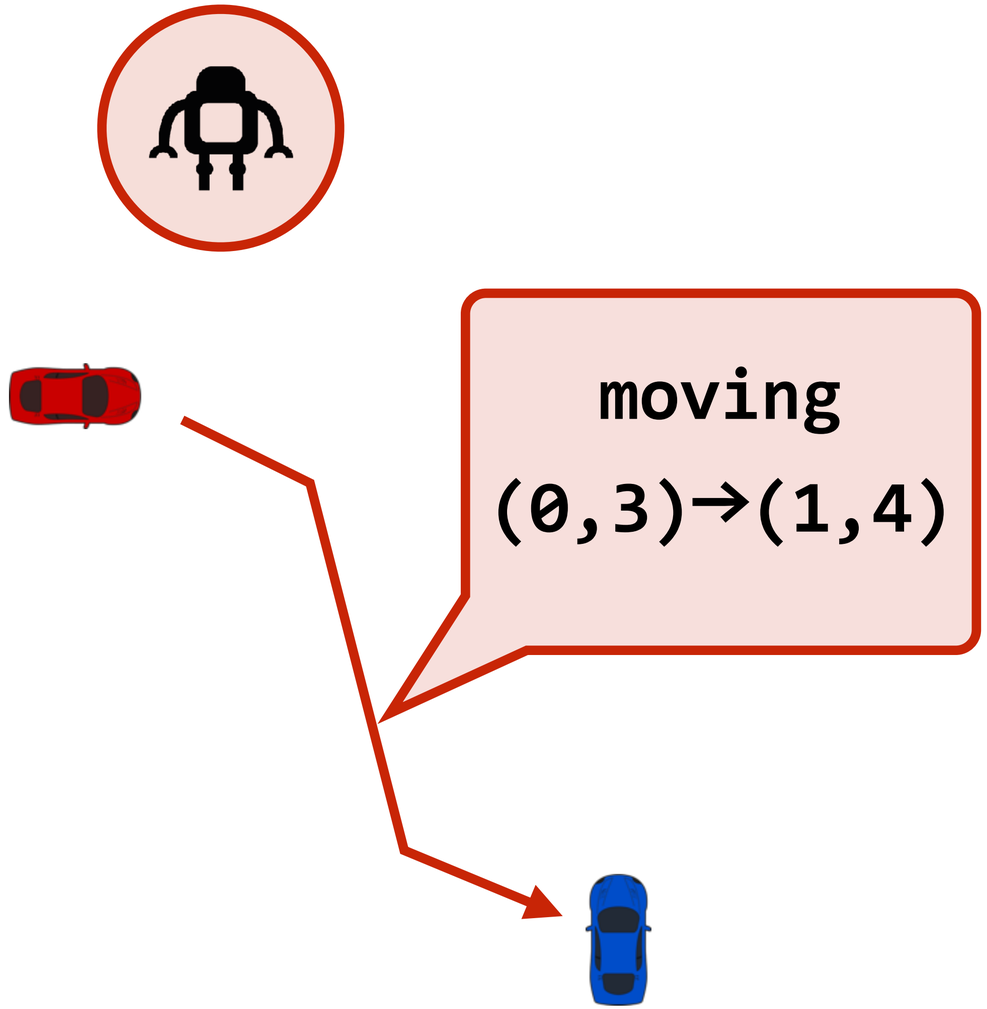




$$\Sigma\ p(\texttt{0}, \blacksquare \mid \textit{not sure})\ p(\textit{not sure})$$

Stat MT criterion doesn't capture meaning





Outline

Natural language & neuralese

✗ **Statistical** machine translation

Semantic machine translation

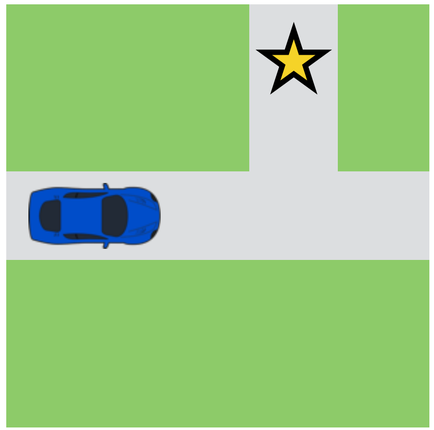
Implementation details

Evaluation

A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

I’m going north

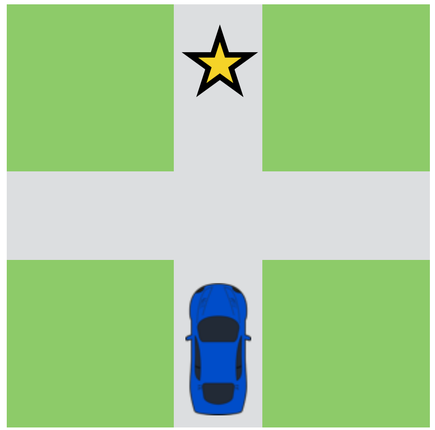


[Davidson 67]

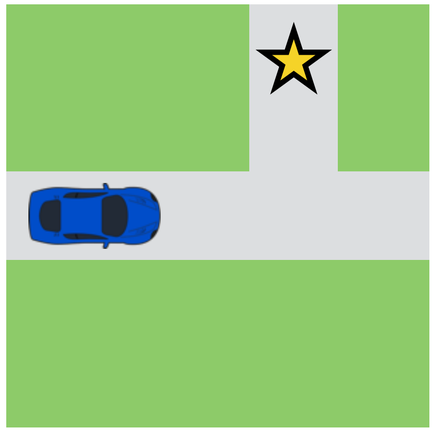
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

I’m going north



✔



✔

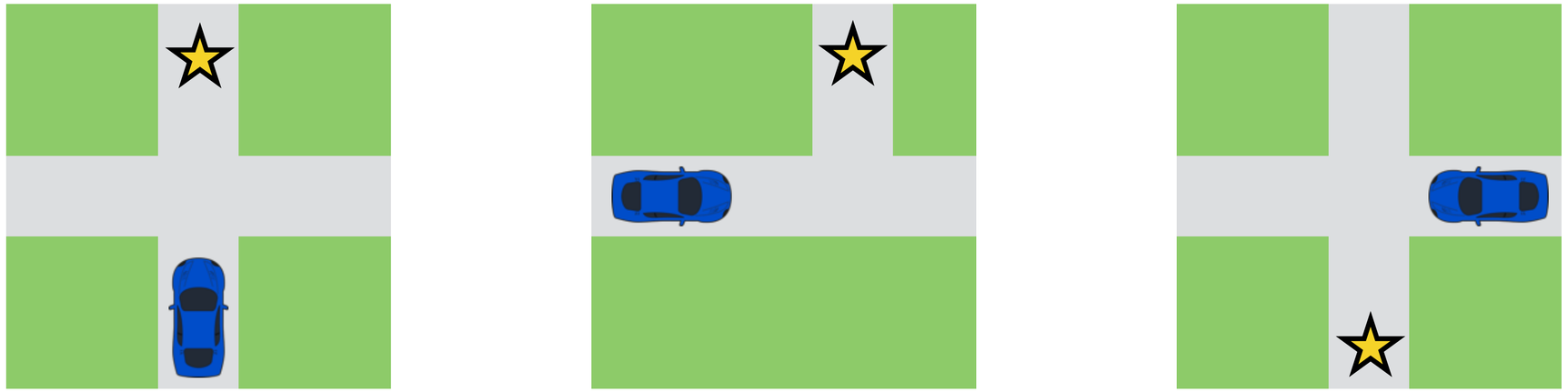
✘

[Davidson 67]

A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

I’m going north



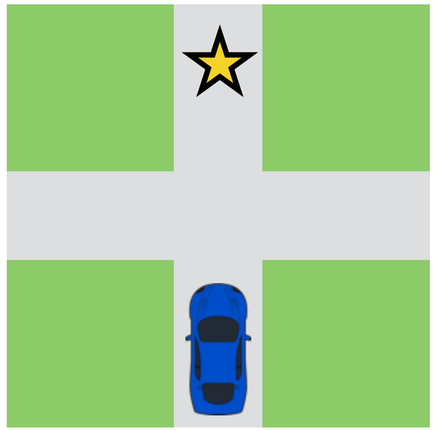
`(loc (goal blue) north)`

A “semantic MT” problem

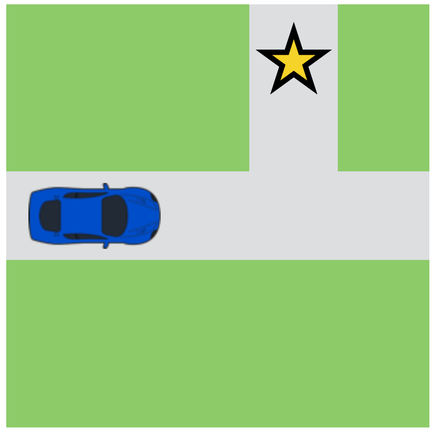
The meaning of an utterance is given by its truth conditions

the **distribution over states** in which it is uttered

I’m going north



0.4



0.2

0.001

[Beltagy et al. 14]

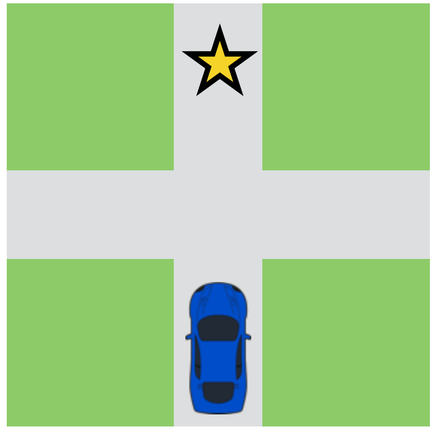
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

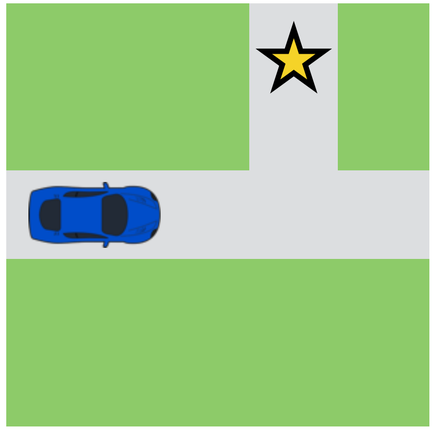
the **distribution over states** in which it is uttered

the **belief** it induces in listeners

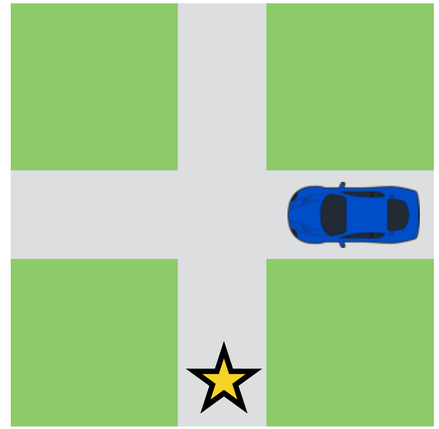
I’m going north



0.4



0.2



0.001

[Frank et al. 09, A & Klein 16]

Representing meaning

The meaning of an utterance is given by

the **distribution over states** in which it is uttered

or equivalently, the **belief** it induces in listeners

Representing meaning

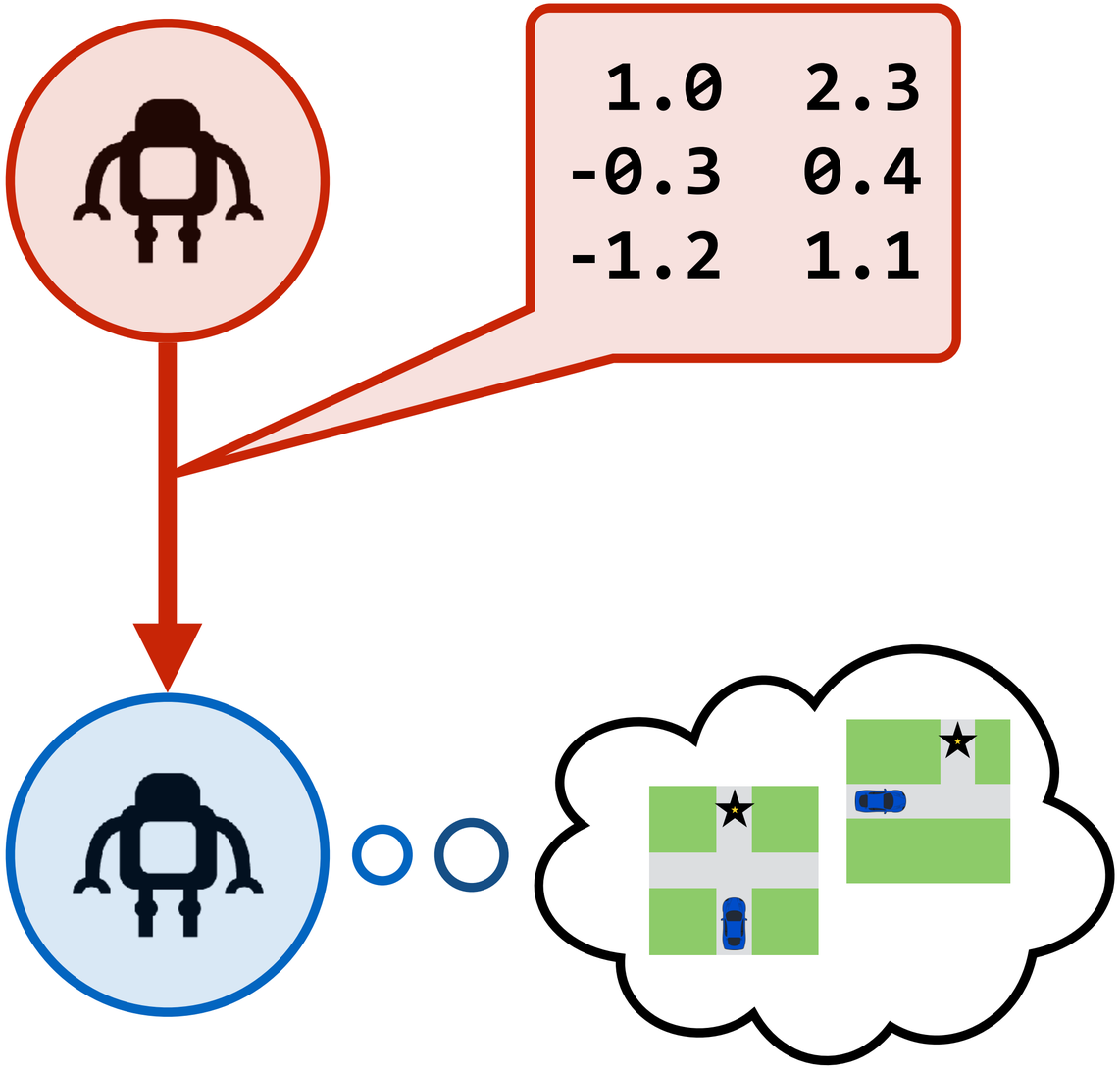
The meaning of an utterance is given by

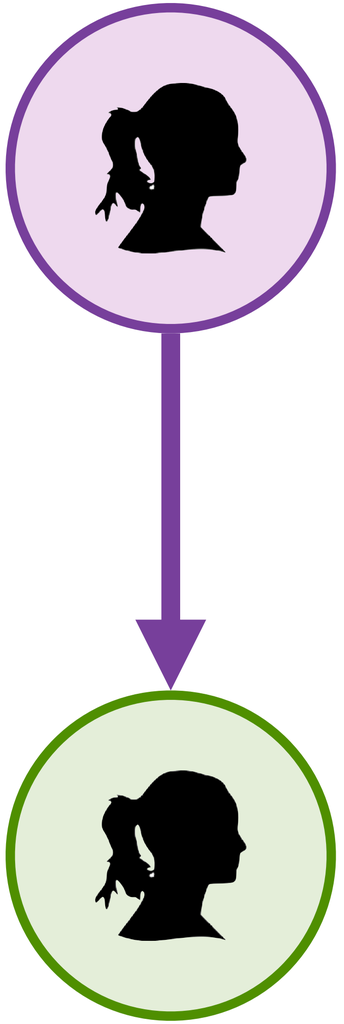
the **distribution over states** in which it is uttered

or equivalently, the **belief** it induces in listeners

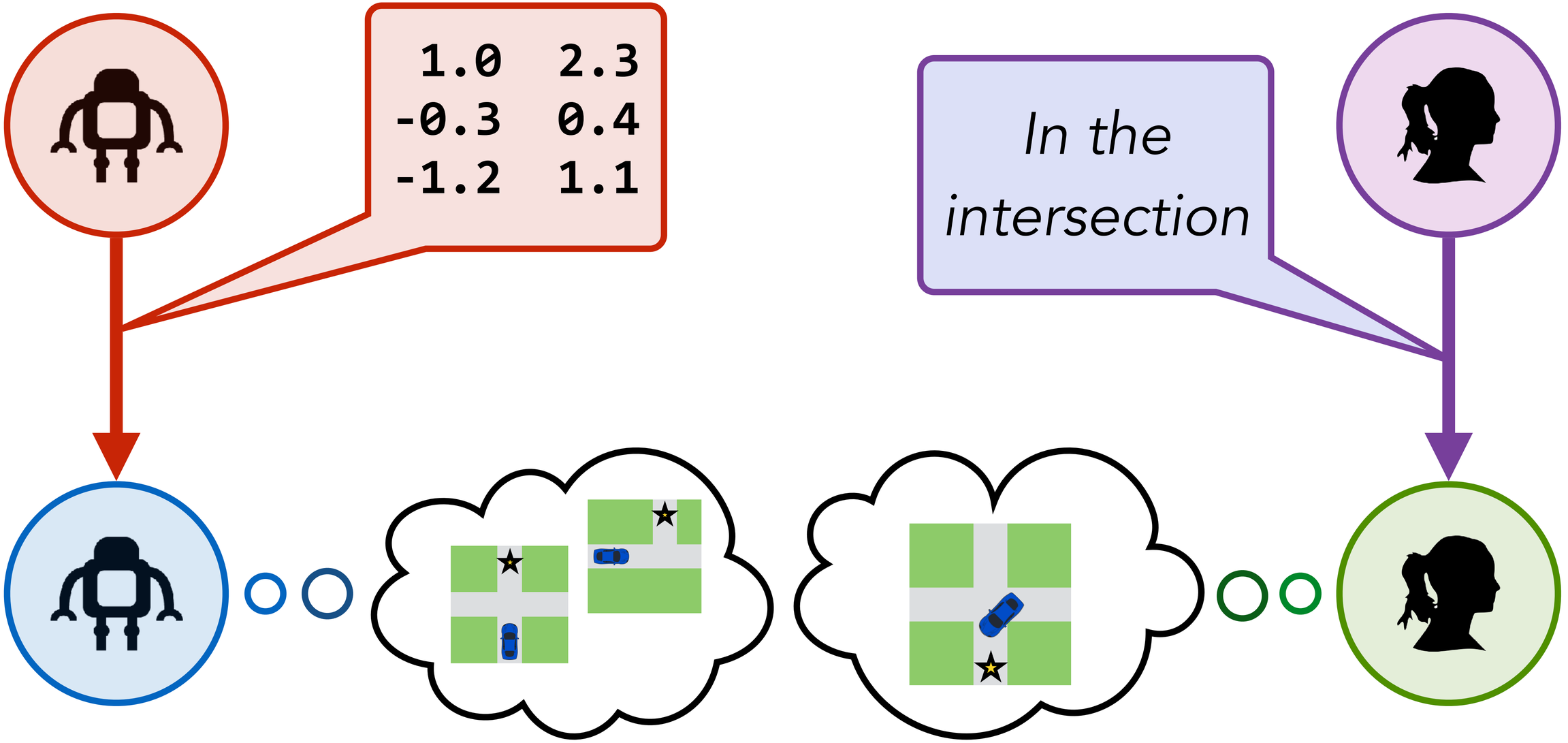
This distribution is well-defined even if the “utterance” is a vector rather than a sequence of tokens.

Translating with meaning

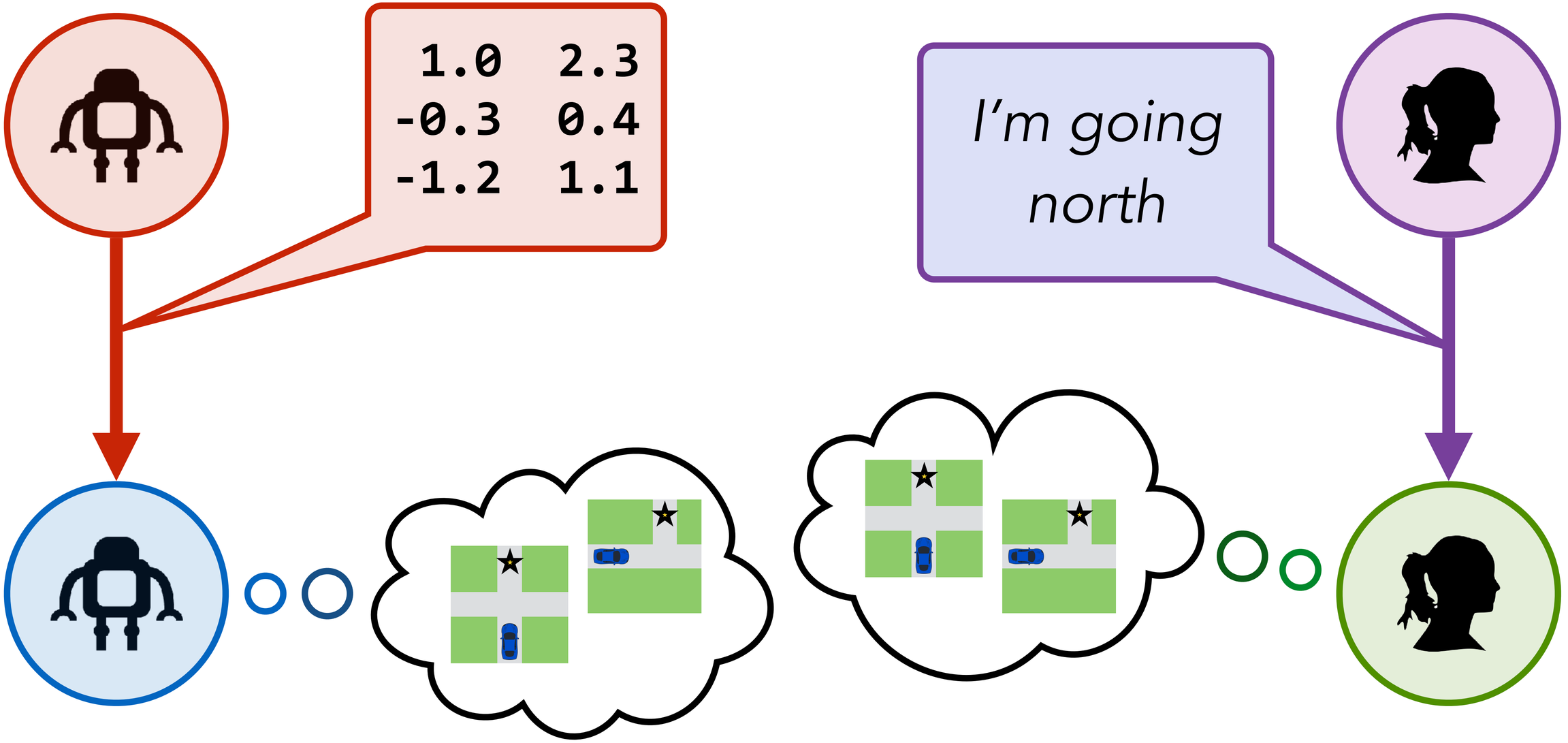




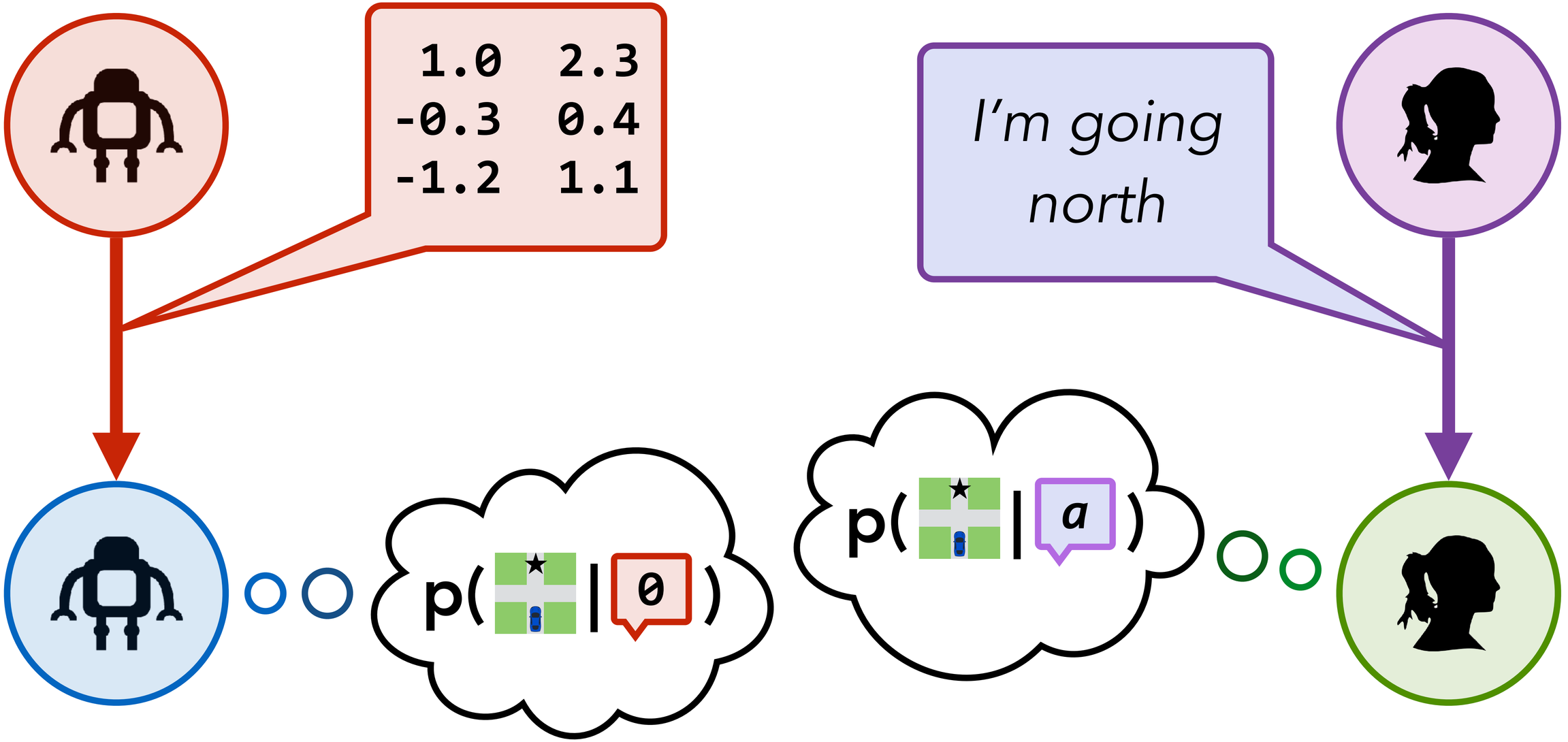
Translating with meaning



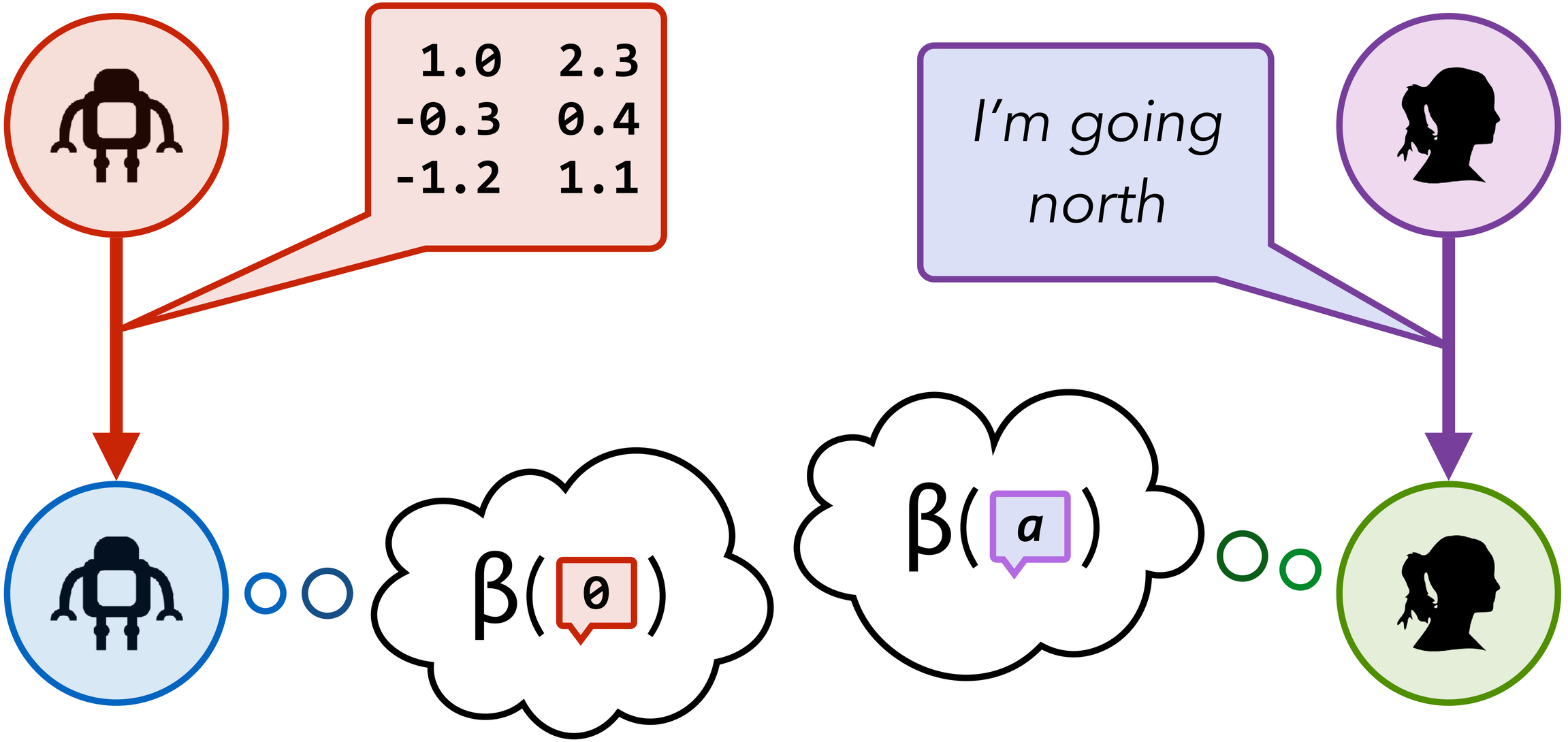
Translating with meaning



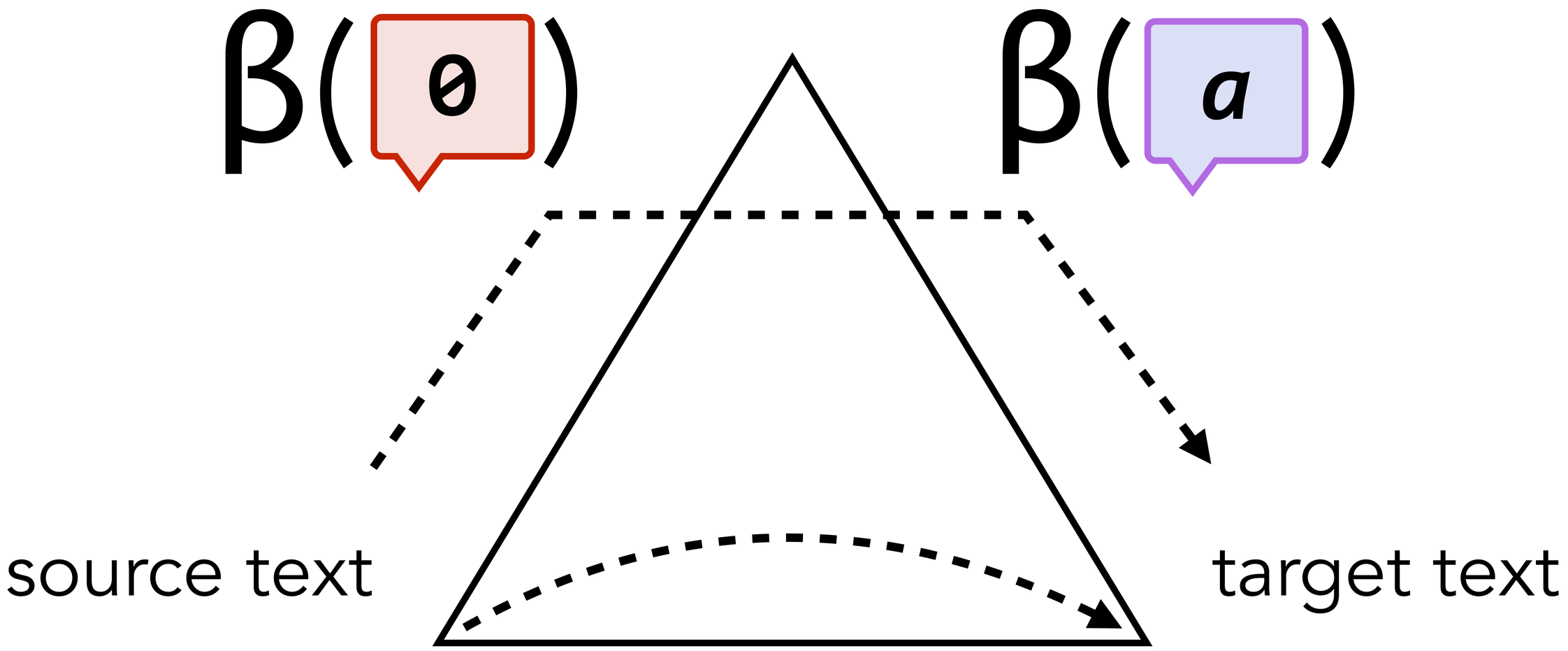
Translating with meaning



Translating with meaning



Interlingua!



Translation criterion

$$\operatorname*{argmin}_{a} \; KL(\, \beta(\theta) \,||\, \beta(a) \,)$$

Translation criterion

$$\operatorname*{argmin}_{a} \; KL\big(\, \beta(\theta) \;||\; \beta(a) \,\big)$$

Translation criterion

$$\operatorname*{argmin}_{a} \; KL(\, \beta(\theta) \,||\, \beta(a) \,)$$

Translation criterion

$$\operatorname{argmin}_{a}$$

$$KL(\beta(\theta) \,||\, \beta(a))$$

Computing representations

$$\operatorname*{argmin}_{a} \; KL(\; \beta(\texttt{0}) \;||\; \beta(\texttt{a}) \;)$$

Computing representations: sparsity

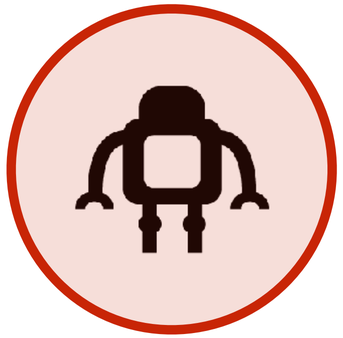
$$\operatorname{argmin}_{a} \; KL\big(\beta(0) \,\|\, \beta(a)\big)$$

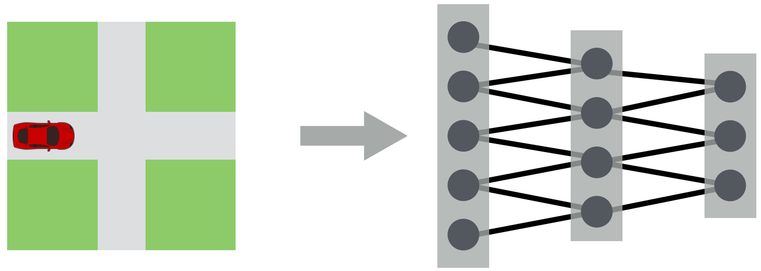
$p(\cdot \mid 0)$	$p(\cdot \mid a)$
	
	
	

Computing representations: smoothing

$$\underset{a}{\operatorname{argmin}} \; KL(\; \beta(\texttt{0}) \;||\; \beta(\texttt{a}) \;)$$

agent policy



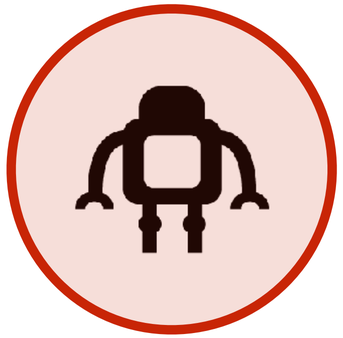


actions & messages

Computing representations: smoothing

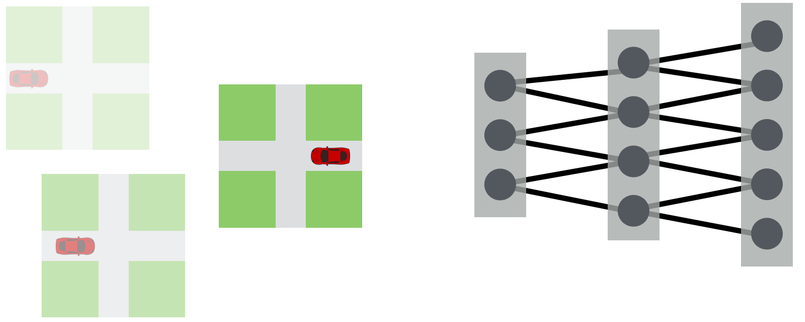
$$\text{argmin}_{a} \; KL(\; \beta(\boxed{0}) \;||\; \beta(\boxed{a}) \;)$$

agent policy



actions & messages

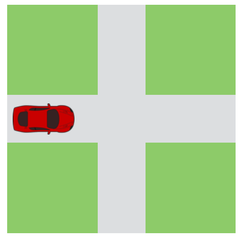
agent model



Computing representations: smoothing

$$\underset{a}{\text{argmin}}\ KL(\ \beta(\boldsymbol{0})\ ||\ \beta(\boldsymbol{a})\ )$$

human



actions & messages

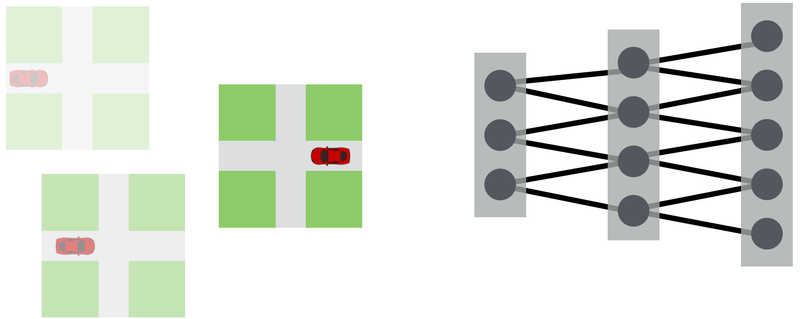
Computing representations: smoothing

$$\text{argmin}_{a} \; KL(\beta(0) \,||\, \beta(a))$$

human policy

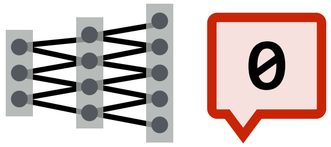
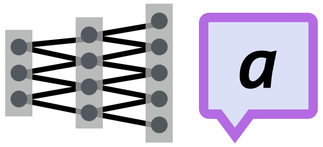
actions & messages

human model



Computing representations: smoothing

$$\text{argmin}_{a} \; KL(\beta(0) \,||\, \beta(a))$$

	 0	 a
	0.10	 0.08
	0.05	 0.01
	0.13	 0.22

Computing KL

$$\operatorname*{argmin}_{a} \; KL(\, \beta(0) \,||\, \beta(a) \,)$$

Computing KL

$$\underset{a}{\text{argmin}}\ \mathrm{KL}(\ \beta(\texttt{0})\ \|\ \beta(a)\ )$$

$$\mathrm{KL}(p \,\|\, q) = \mathbf{E}_{\text{}\ p} \frac{p(\text{  })}{q(\text{  })}$$

Computing KL: sampling

$$\underset{a}{\text{argmin}}\ \mathrm{KL}(\ \beta(\texttt{0})\ \|\ \beta(a)\ )$$

$$\mathrm{KL}(p \| q) = \sum_{i} p(\text{  }_i) \log \frac{p(\text{  }_i)}{q(\text{  }_i)}$$

Finding translations

$$\operatorname{argmin}_{a} \; KL(\, \beta(0) \,||\, \beta(a) \,)$$

Finding translations: brute force

$$\text{argmin}_{a} \; KL(\beta(0) \,||\, \beta(a))$$

going north → 0.5

crossing the intersection → 2.3

I'm done → 0.2

after you → 9.7

Finding translations: brute force

$$\operatorname{argmin}_{a} \; KL(\beta(0) \,||\, \beta(a))$$

going north → 0.5

crossing the intersection → 2.3

I'm done → 0.2

after you → 9.7

Finding translations

$$\underset{a}{\operatorname{argmin}}\ KL(\ \beta(\theta)\ ||\ \beta(a)\ )$$

Outline

Natural language & neuralese

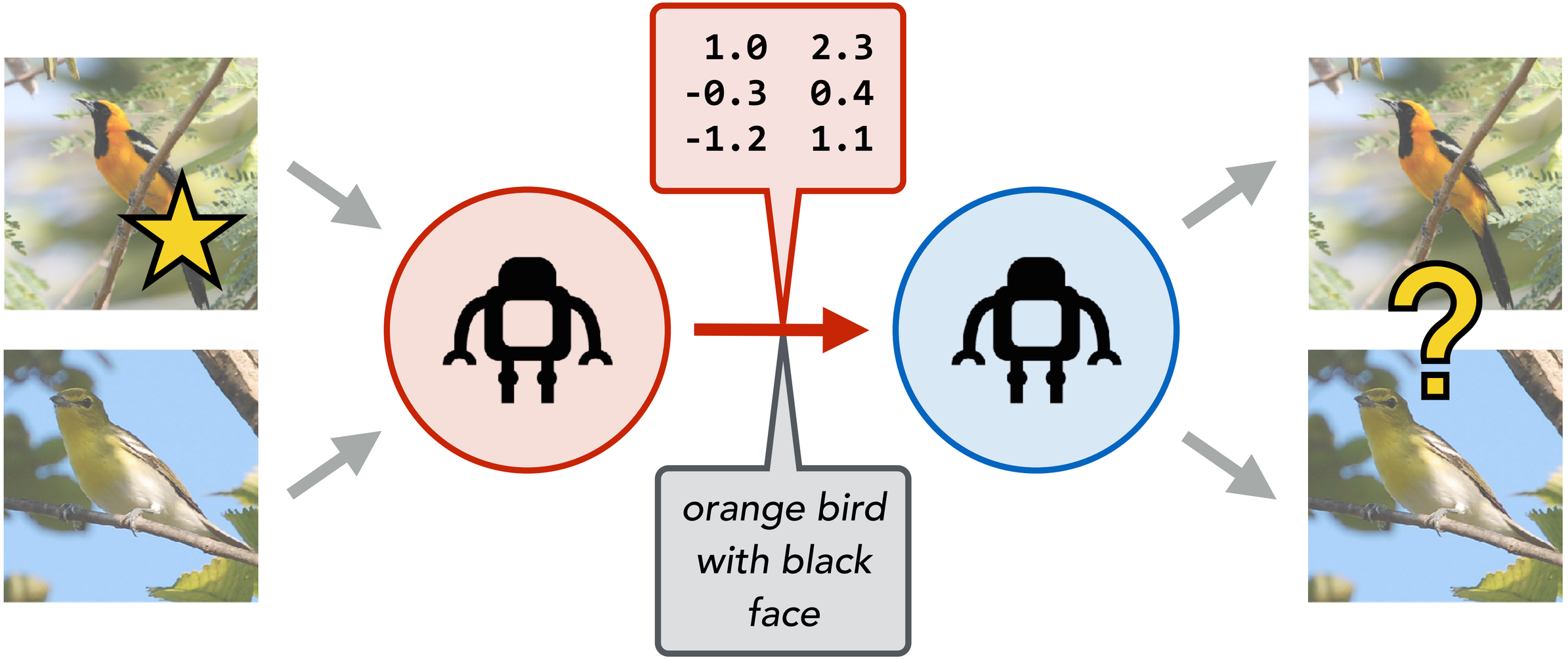
Statistical machine translation

Semantic machine translation

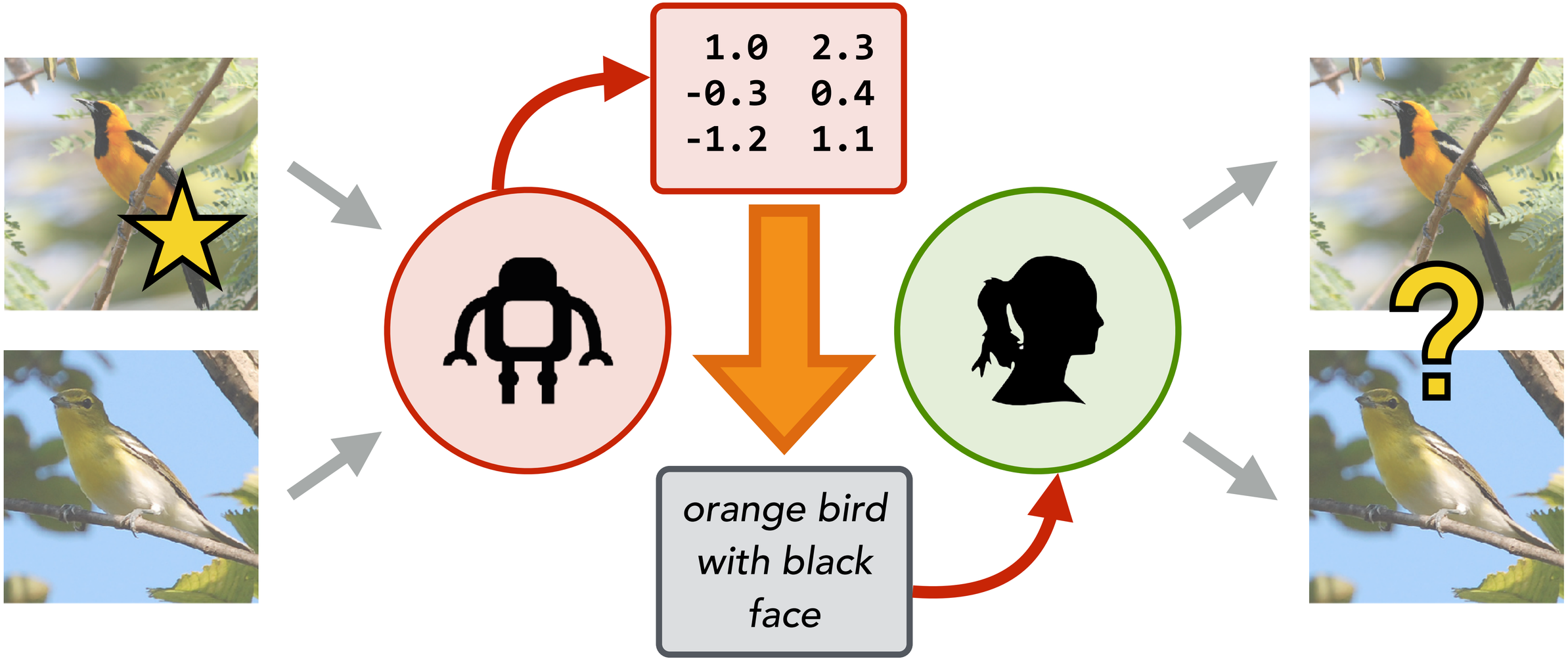
Implementation details

Evaluation

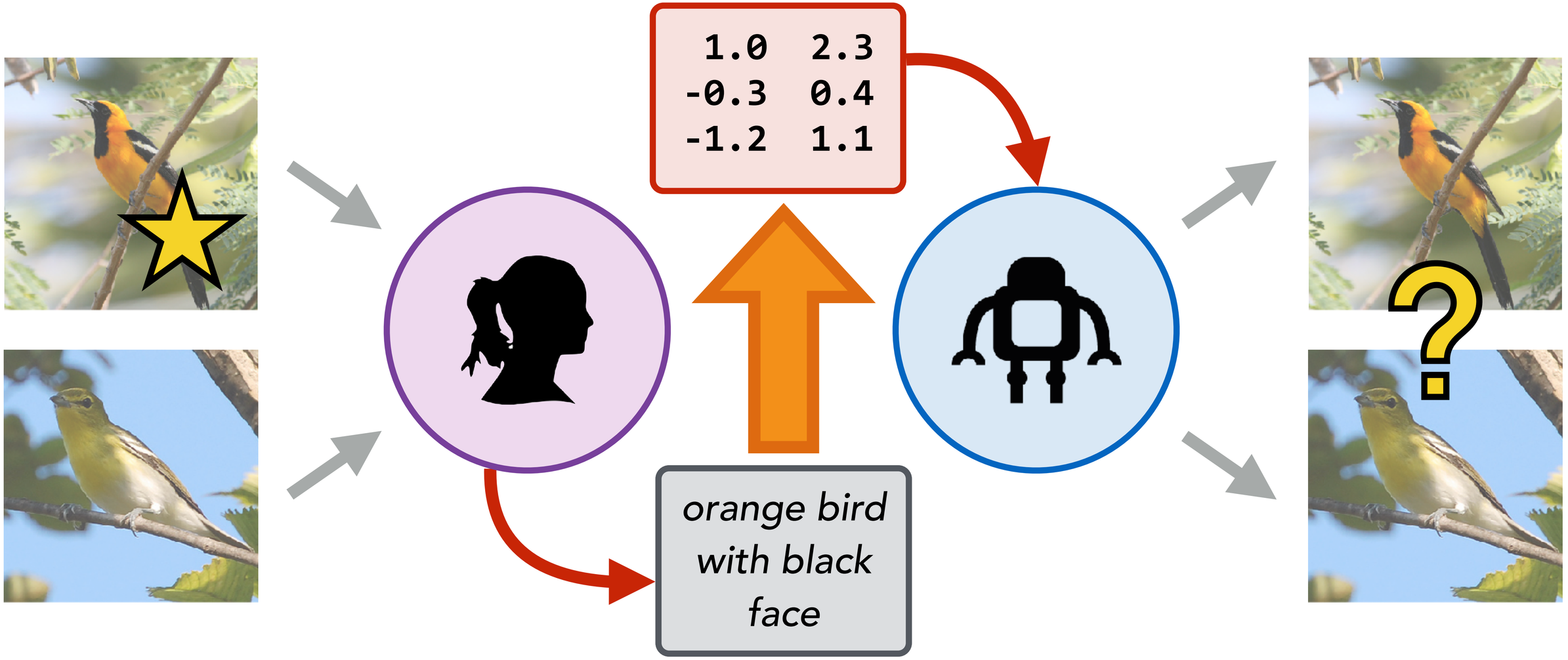
Referring expression games



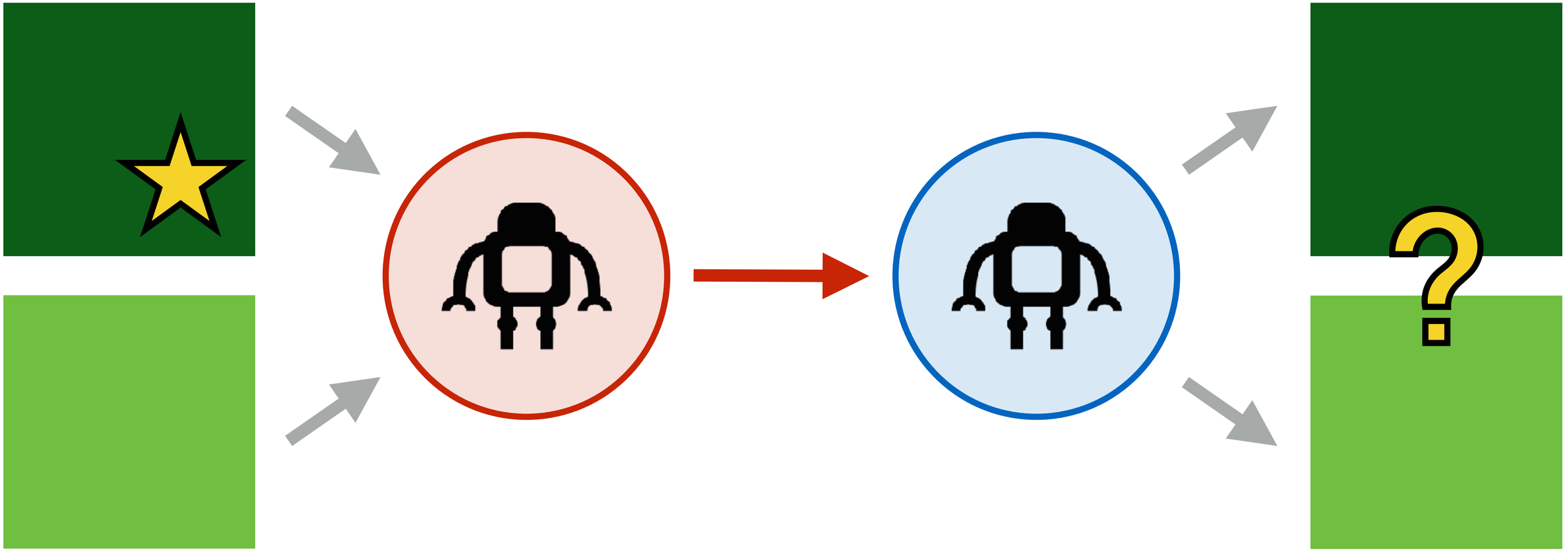
Evaluation: translator-in-the-loop



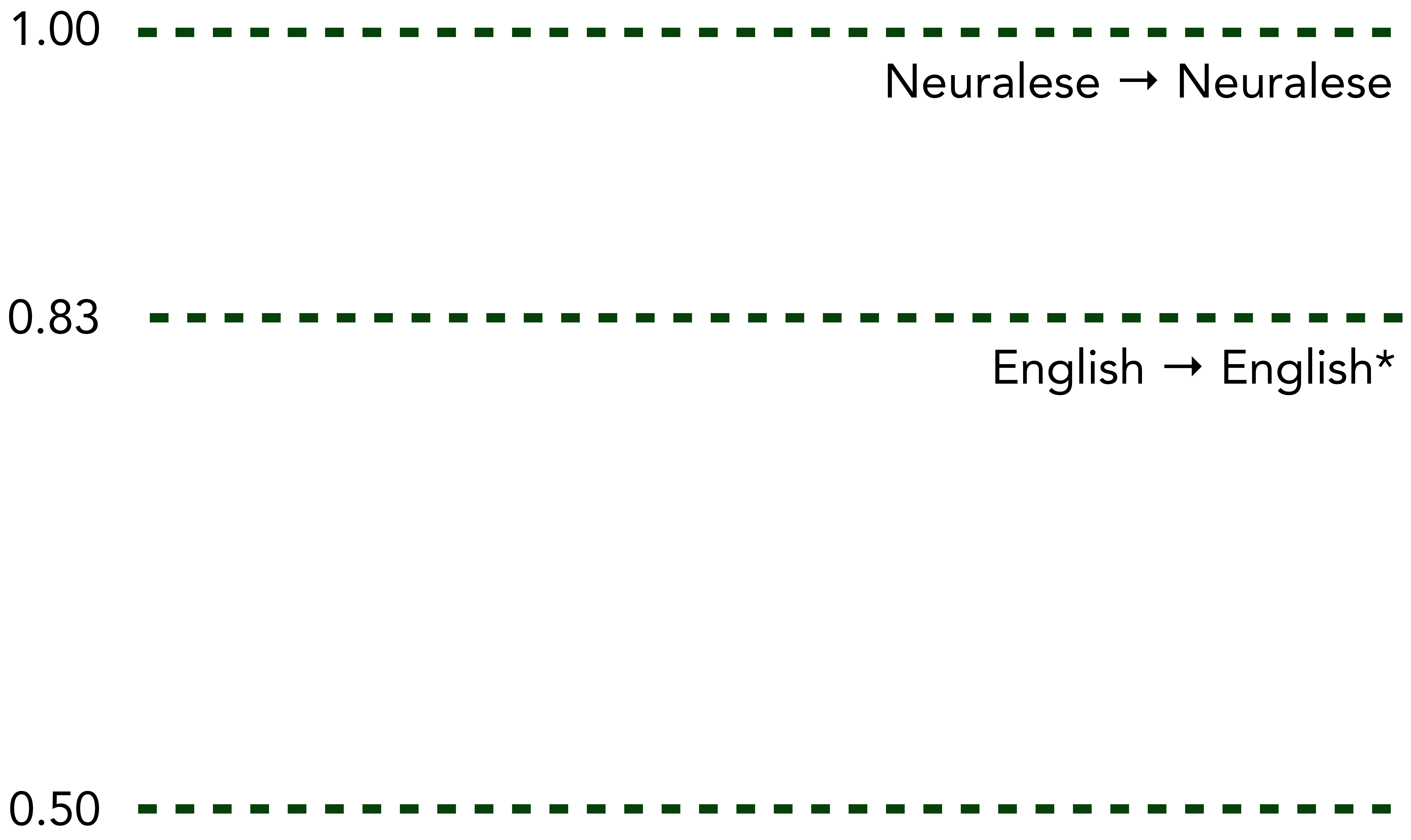
Evaluation: translator-in-the-loop



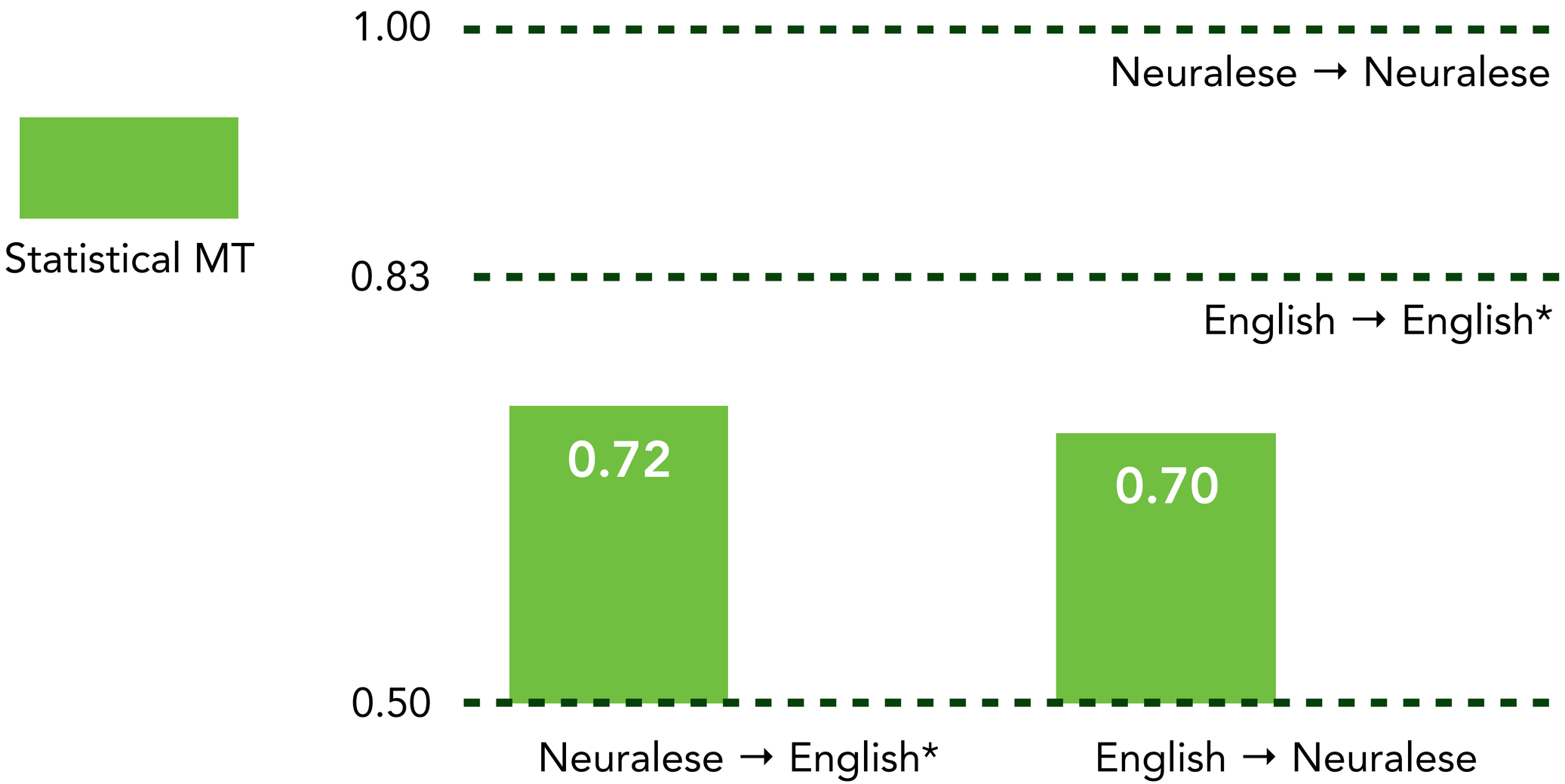
Experiment: color references



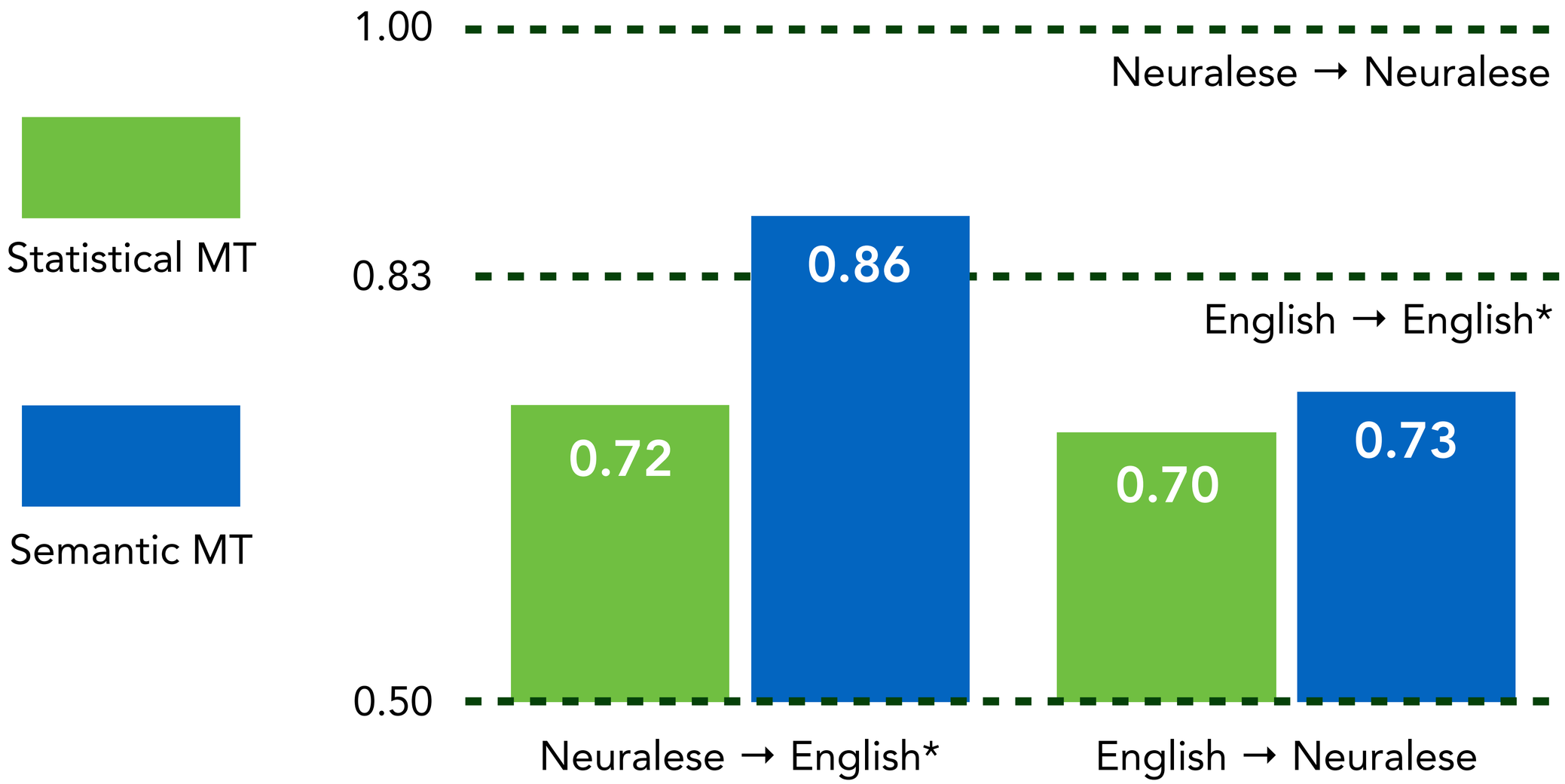
Experiment: color references



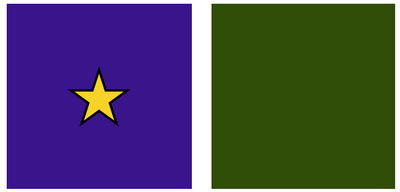
Experiment: color references



Experiment: color references

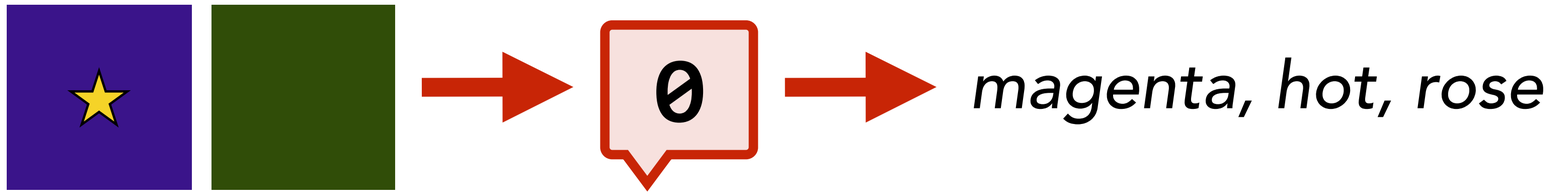


Experiment: color references

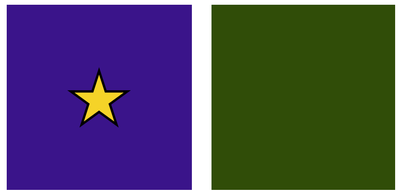


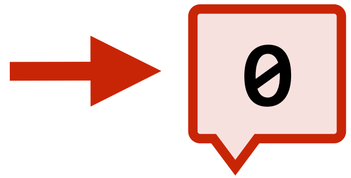
Experiment: color references

Experiment: color references

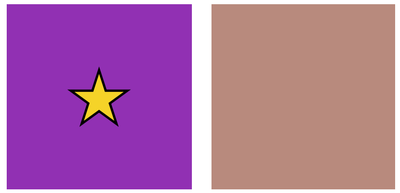


Experiment: color references



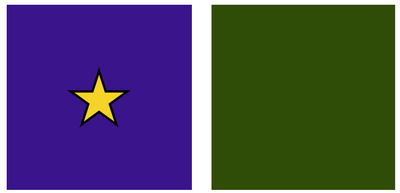


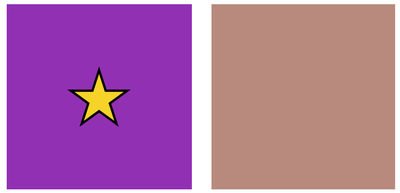
magenta, hot, rose

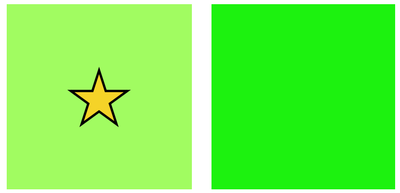


magenta, hot, violet

Experiment: color references





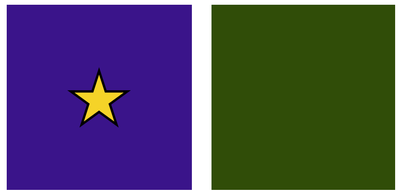


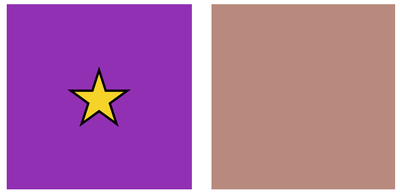
magenta, hot, rose

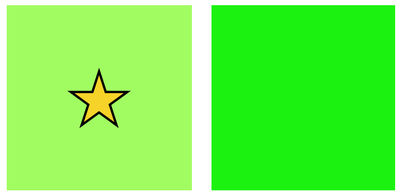
magenta, hot, violet

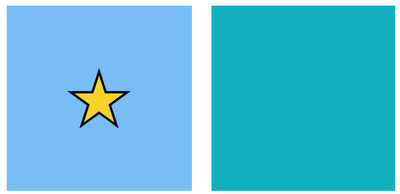
olive, puke, pea

Experiment: color references









0

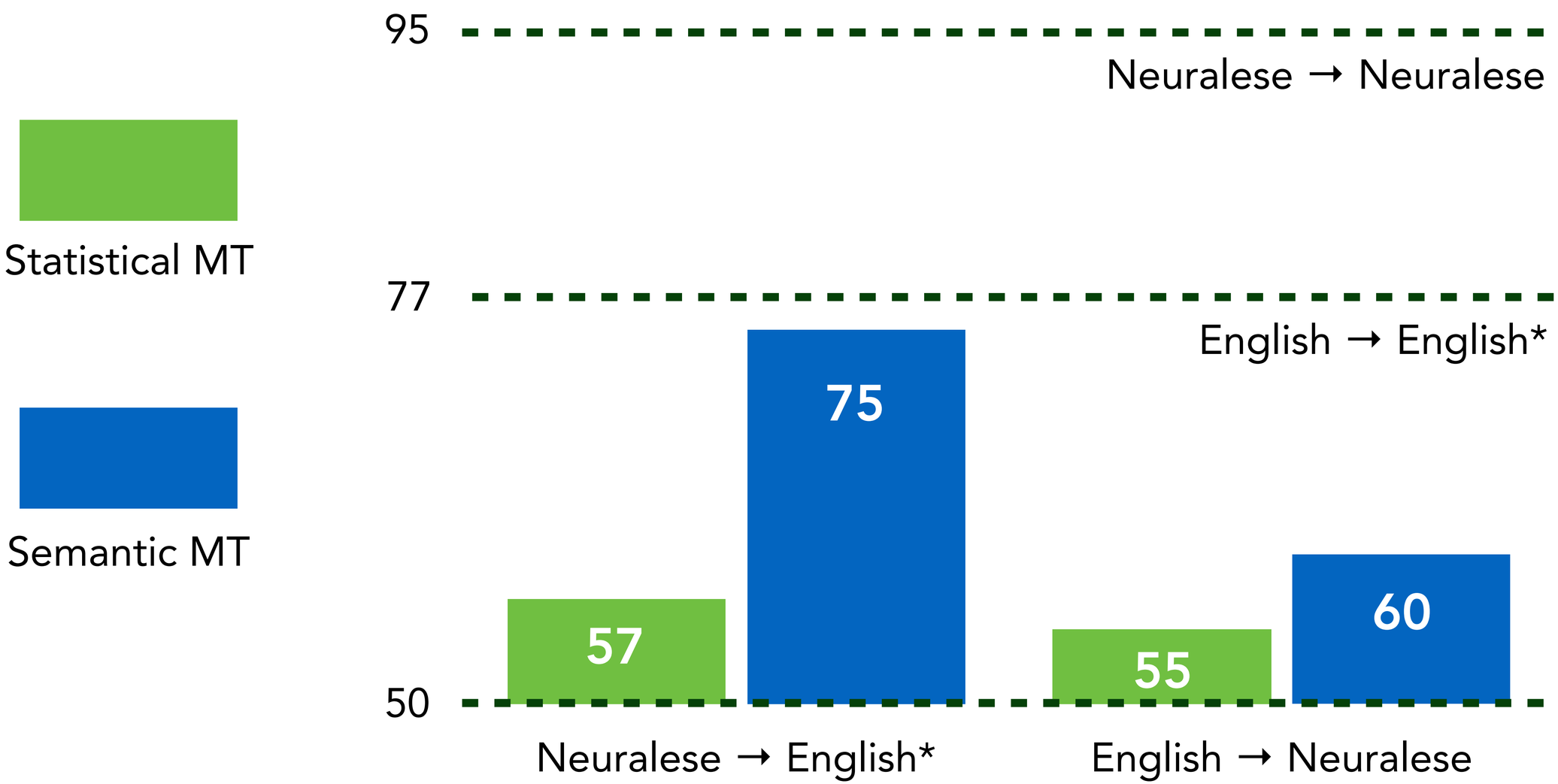
magenta, hot, rose

magenta, hot, violet

olive, puke, pea

pinkish, grey, dull

Experiment: image references

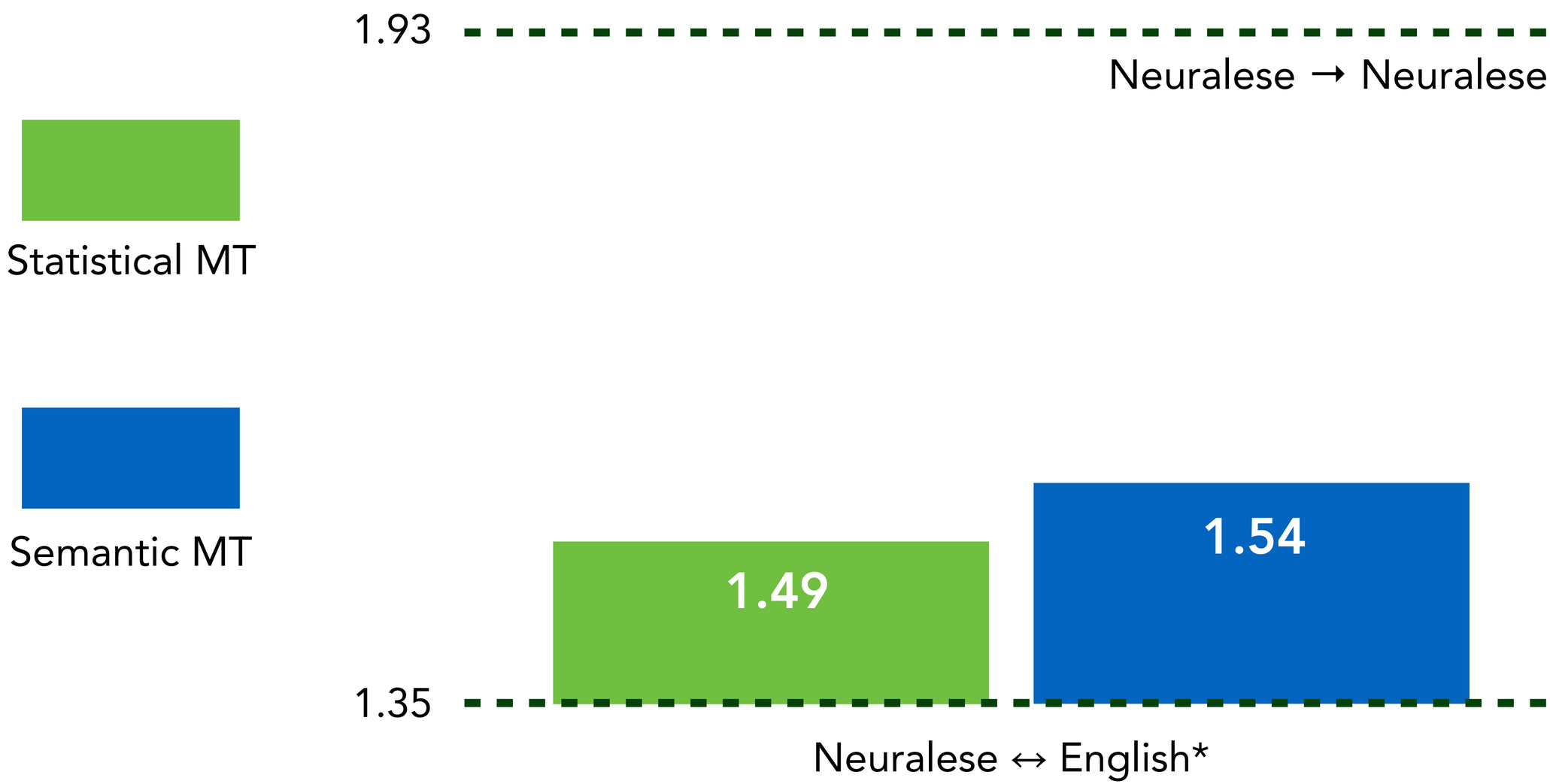


Experiment: image references

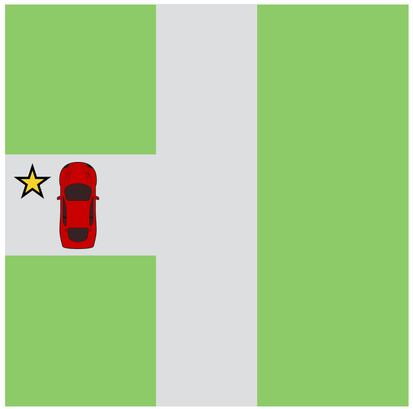

large bird, black wings, black crown


small brown, light brown, dark brown

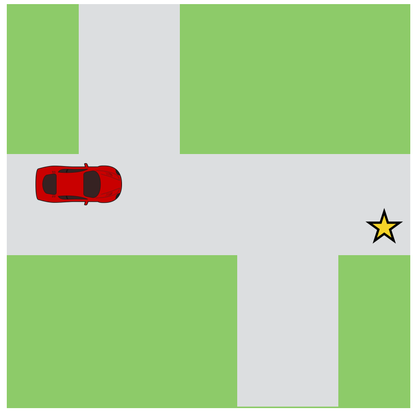
Experiment: driving game



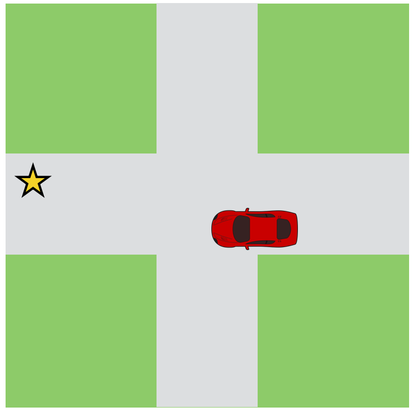
How to translate



at goal
done
left to top



you first
following
going down



going in intersection
proceed
going

Conclusions so far

- Classical notions of "meaning" apply even to un-language-like things (e.g. RNN states)
- These meanings can be compactly represented without logical forms if we have access to world states
- Communicating policies "say" interpretable things!

Conclusions so far

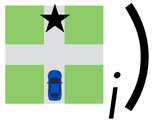
- Classical notions of "meaning" apply even to non-language-like things (e.g. RNN states)
- These meanings can be compactly represented without logical forms if we have access to world states
- Communicating policies "say" interpretable things!

Conclusions so far

- Classical notions of "meaning" apply even to non-language-like things (e.g. RNN states)
- These meanings can be compactly represented without logical forms if we have access to world states
- Communicating policies "say" interpretable things!

Limitations

$$\underset{a}{\operatorname{argmin}} \; KL(\beta(0) \,||\, \beta(a))$$

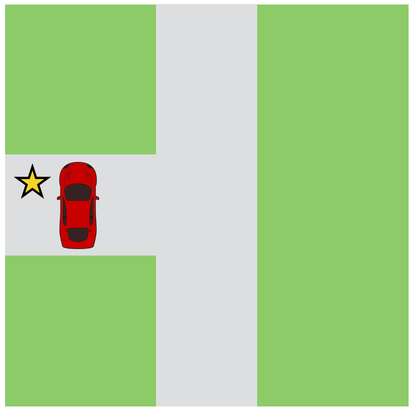
$$KL(p \,||\, q) = \sum_i p(\text{}_i) \log \frac{p(\text{  }_i)}{q(\text{  }_i)}$$

but what about compositionality?

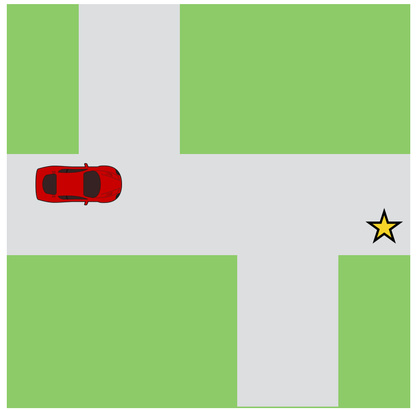
Analogs of linguistic structure in deep representations

Jacob Andreas and Dan Klein

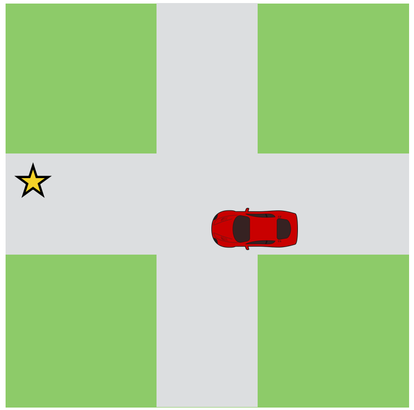
"Flat" semantics



at goal
done

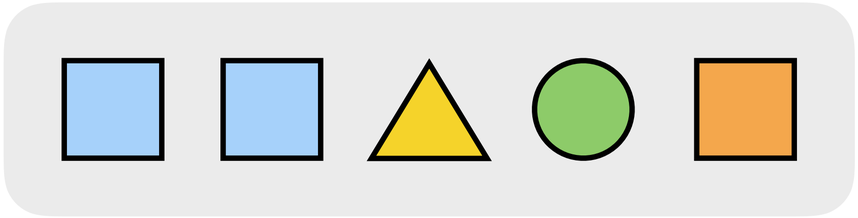


you first
following

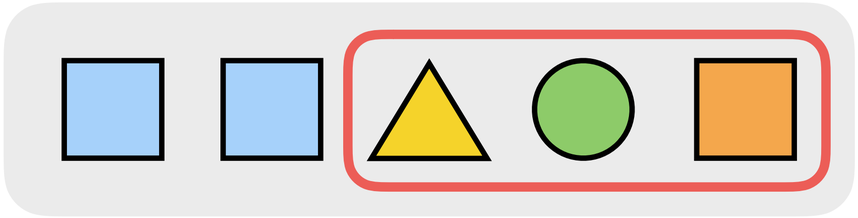


going in intersection
proceed
going

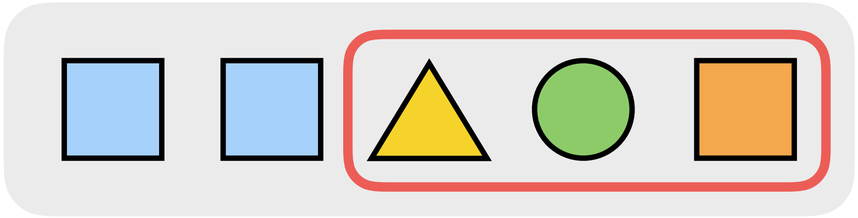
Compositional semantics



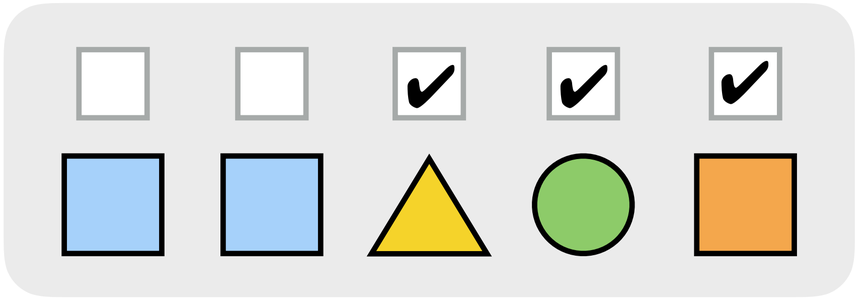
Compositional semantics



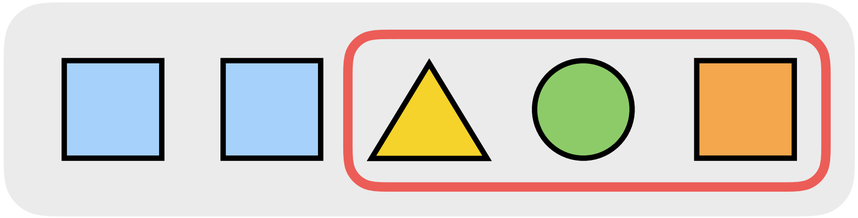
Compositional semantics



message

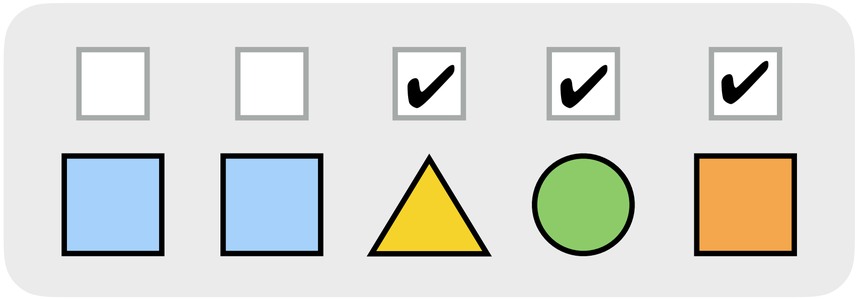


Compositional semantics



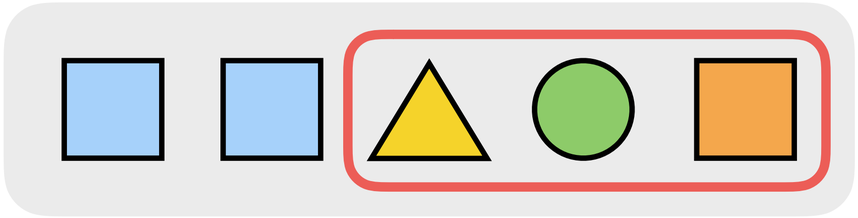
everything but the blue shapes

orange square and non-squares



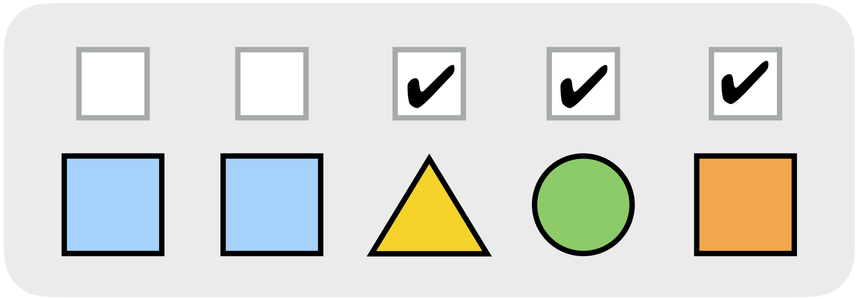
[FitzGerald et al. 2013]

Compositional semantics



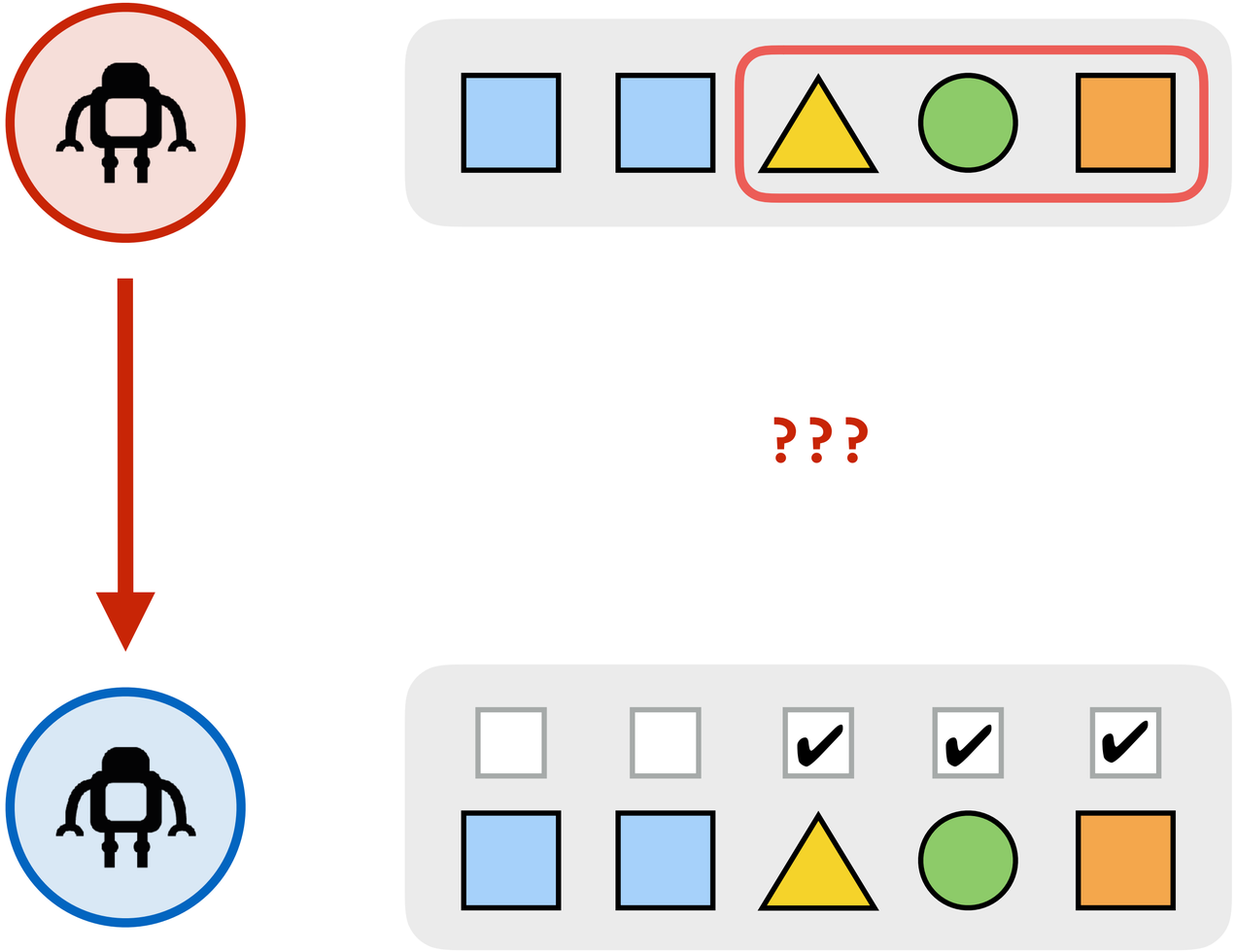
```
lambda x: not(blue(x))
```

```
lambda x: or(orange(x), not(square(x))
```

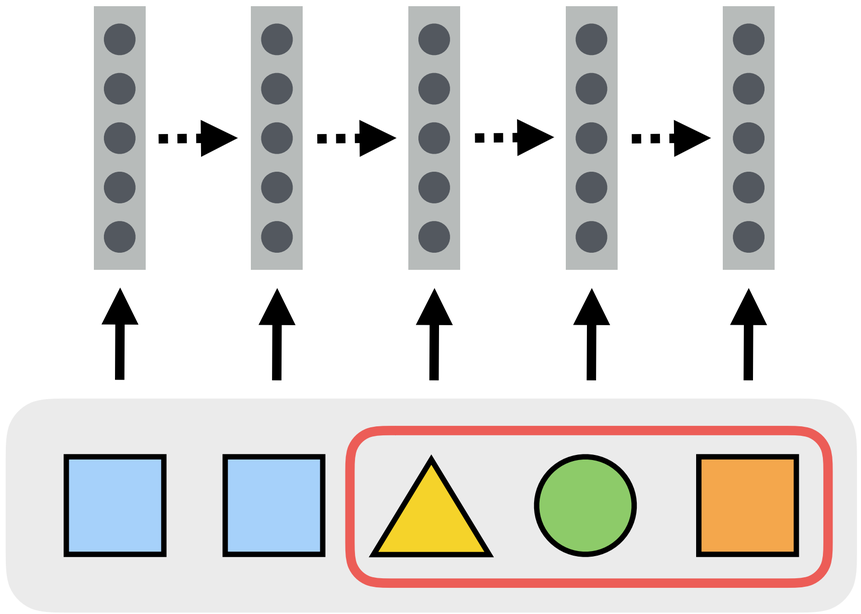


[FitzGerald et al. 2013]

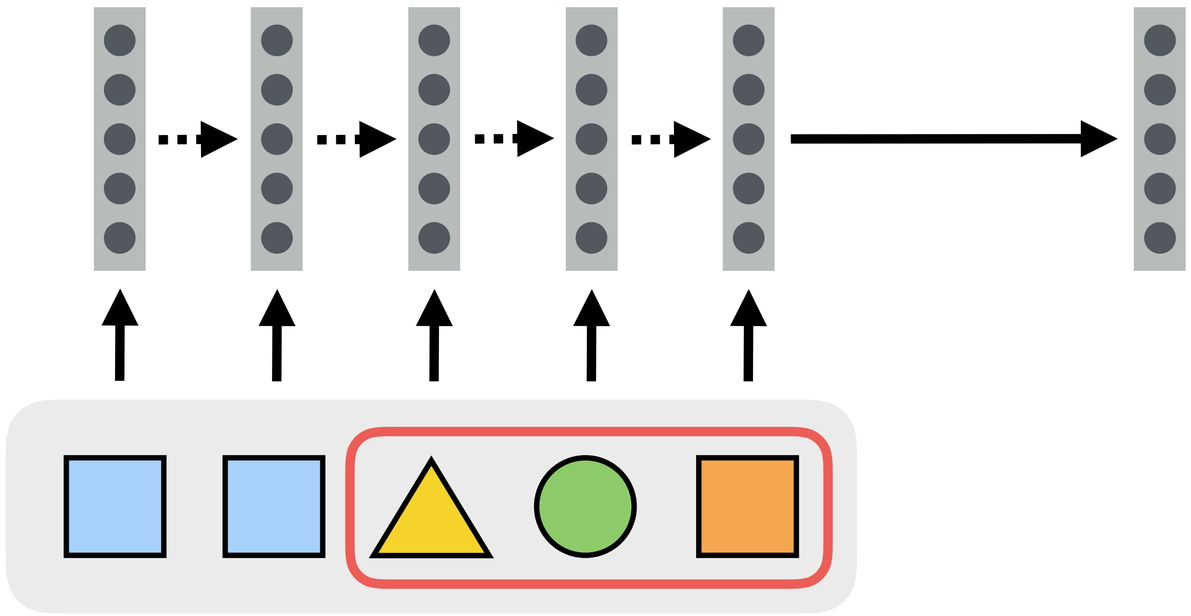
Compositional semantics



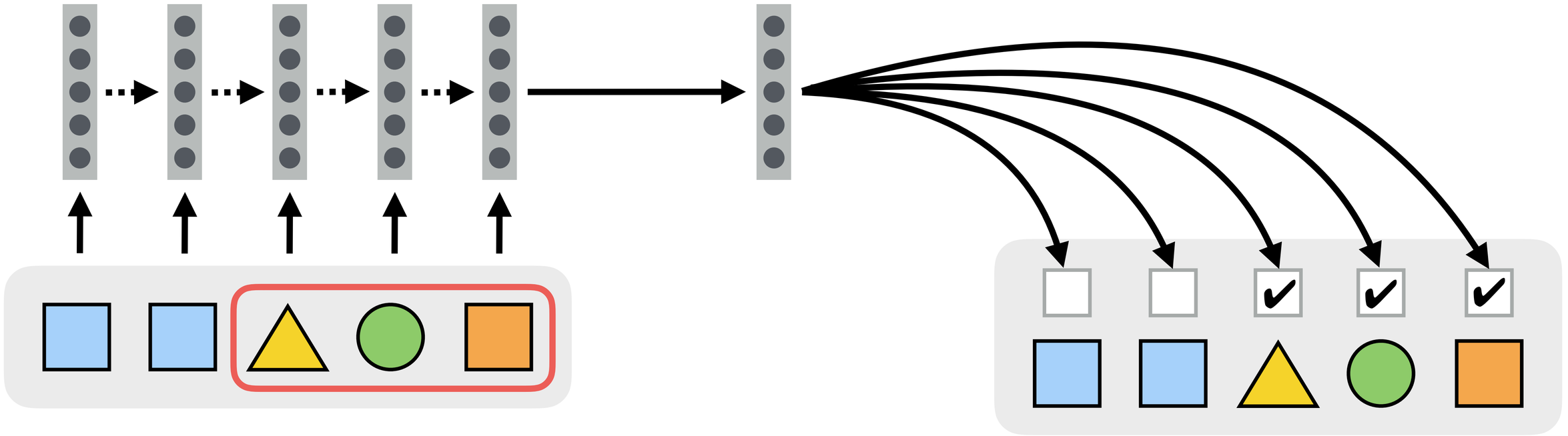
Model architecture



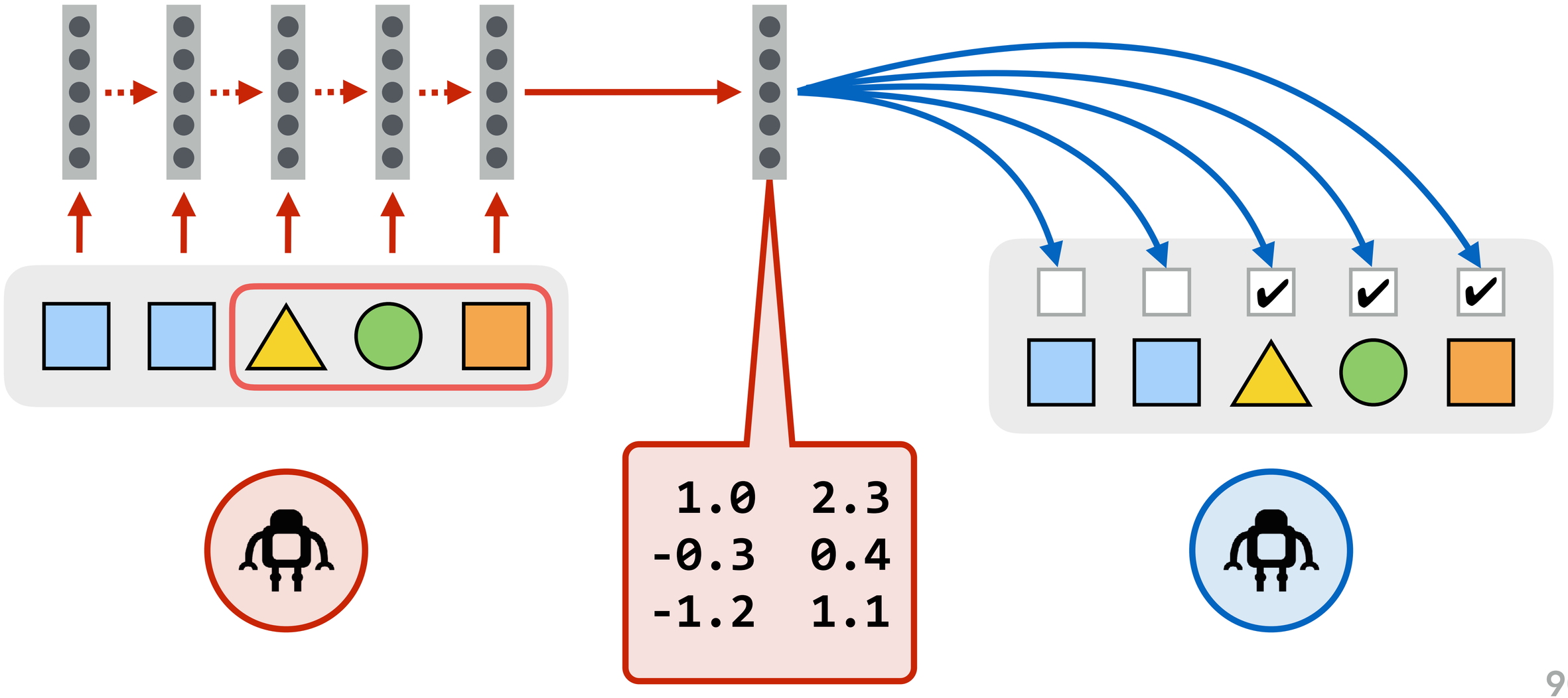
Model architecture



Model architecture



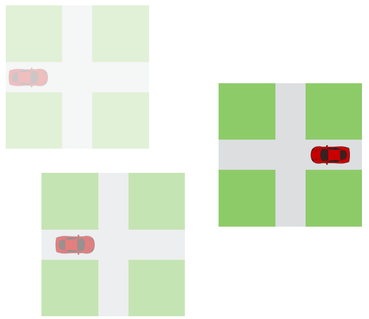
Model architecture

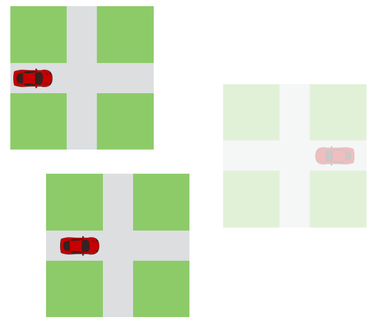


Computing meaning representations

```
 1.0  2.3
-0.3  0.4
-1.2  1.1
```

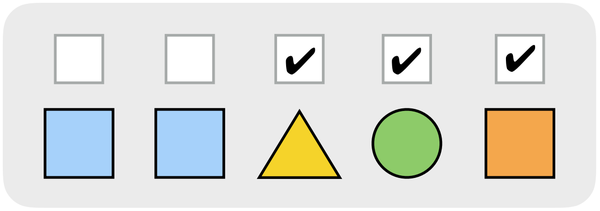
on the left

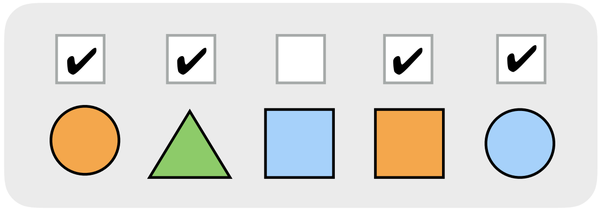




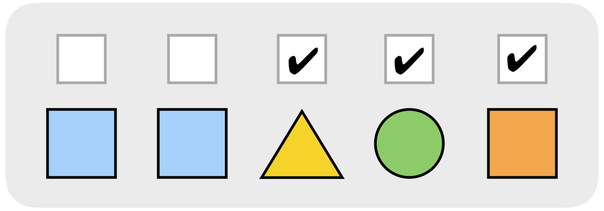
Computing meaning representations

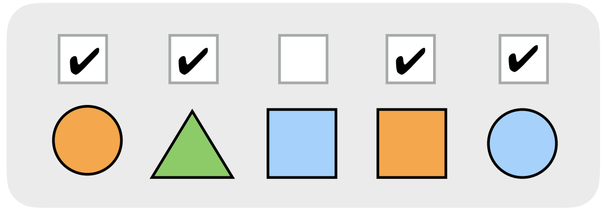
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```





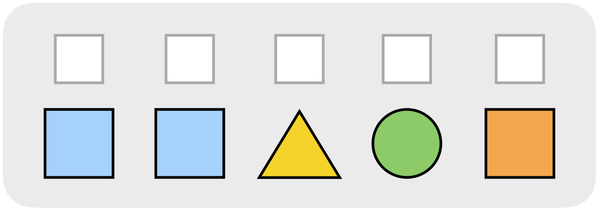
everything but squares

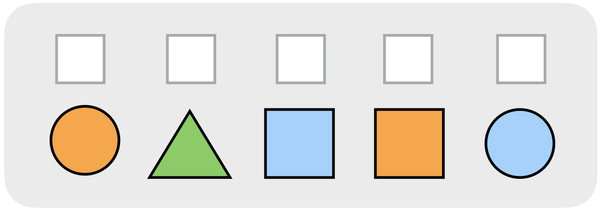




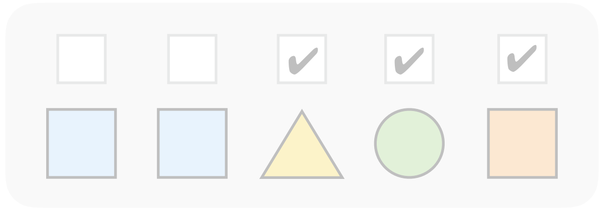
Computing meaning representations

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

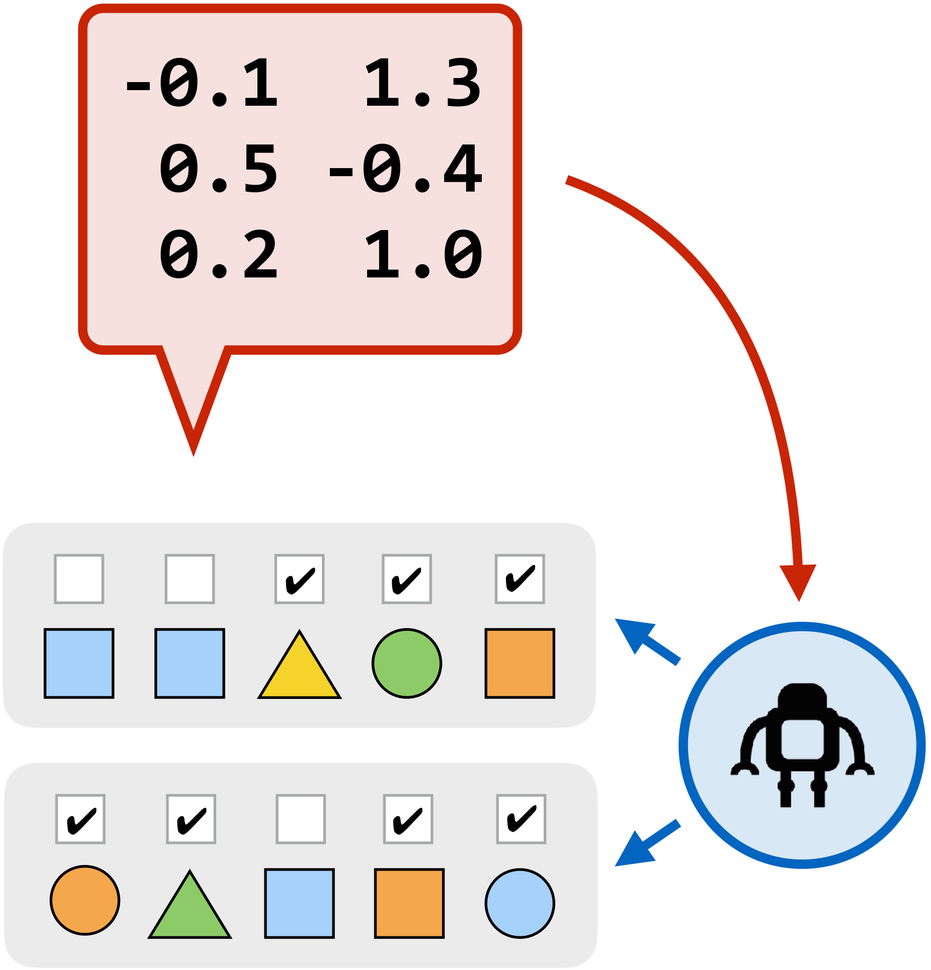


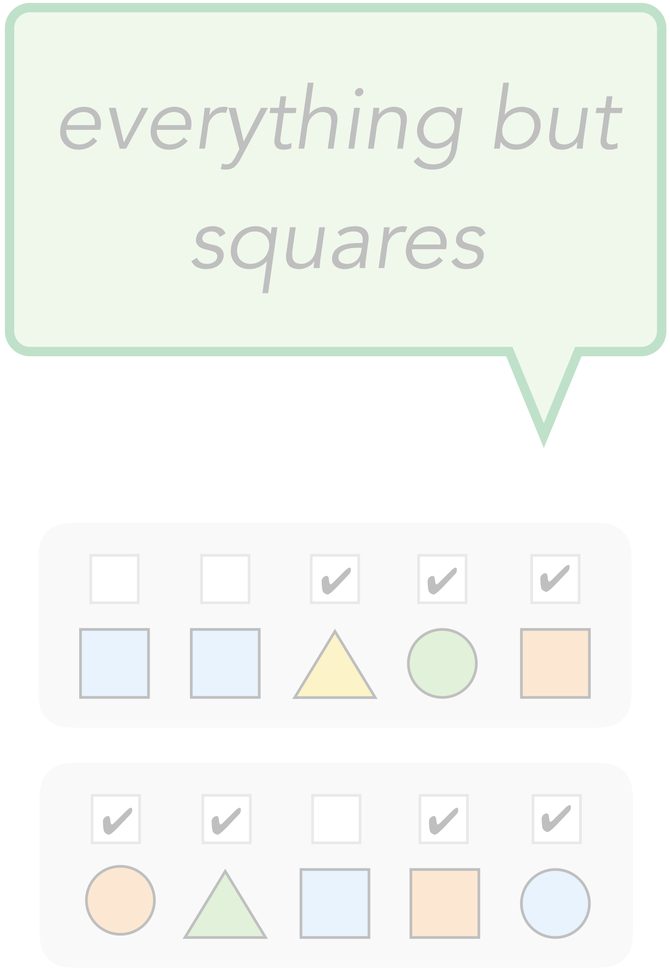


everything but squares

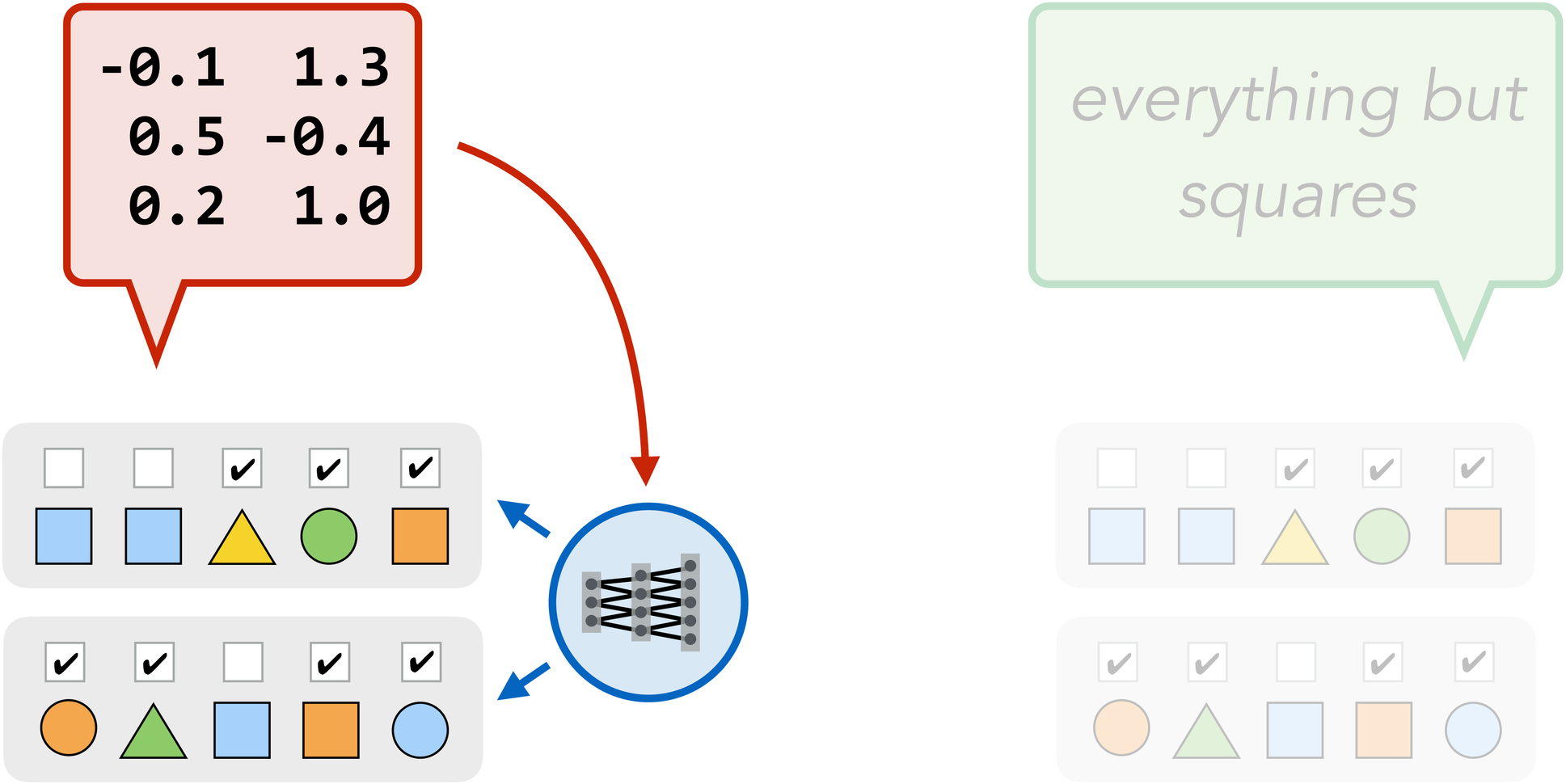


Computing meaning representations





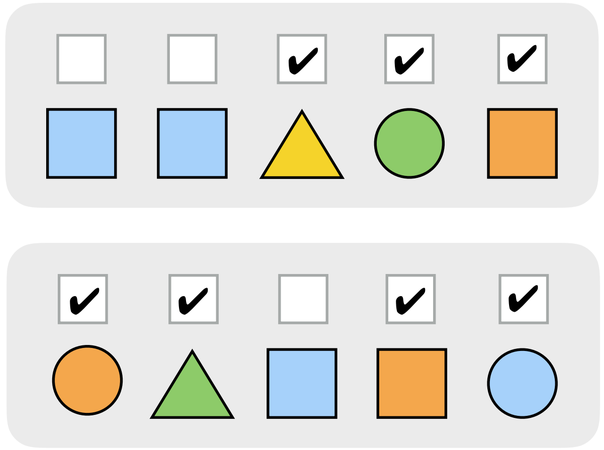
Computing meaning representations



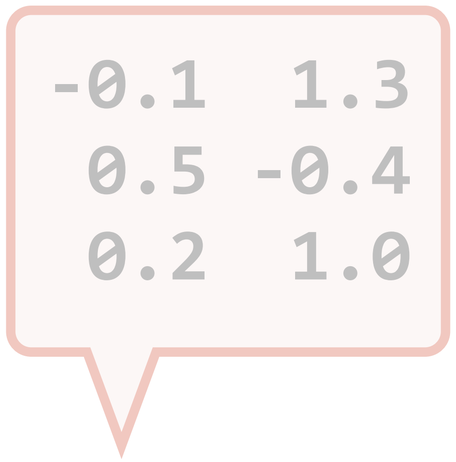
Computing meaning representations

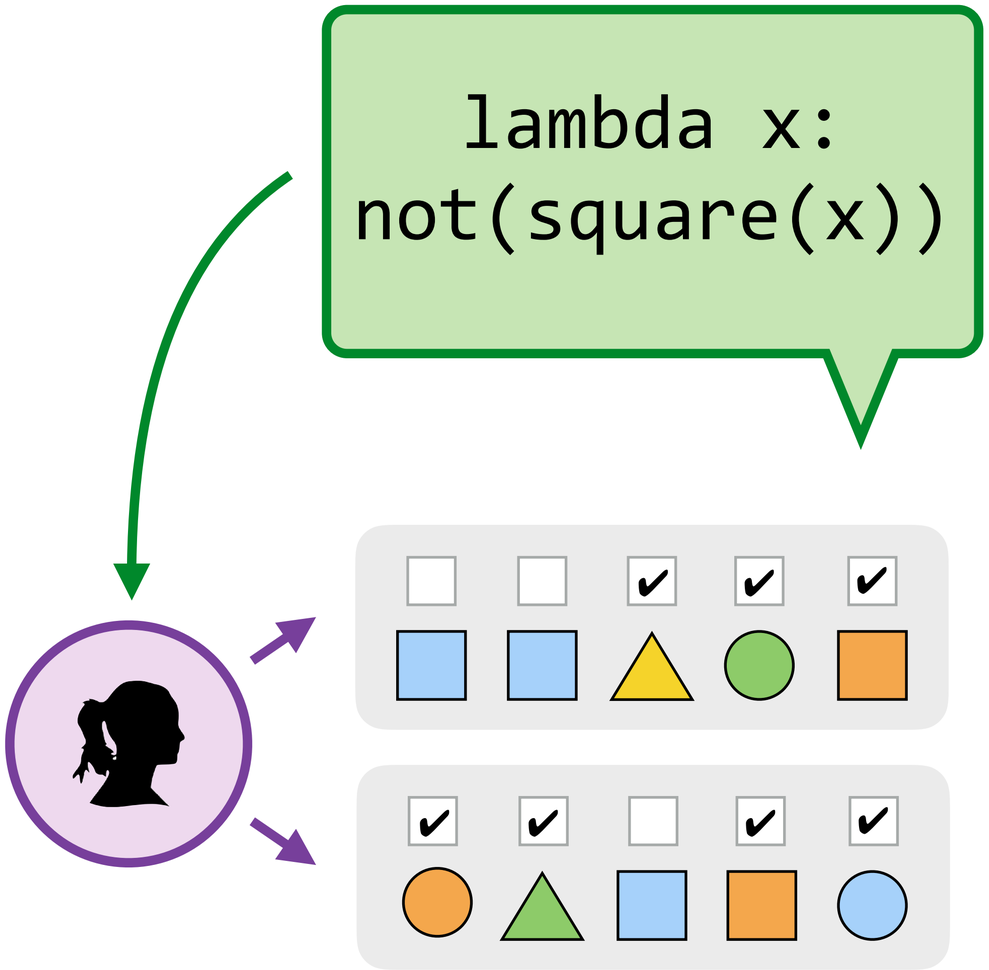
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

everything but squares

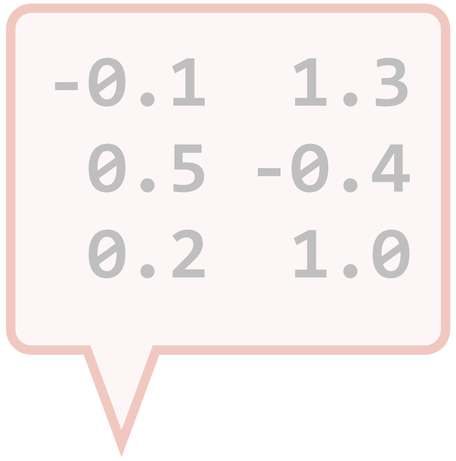


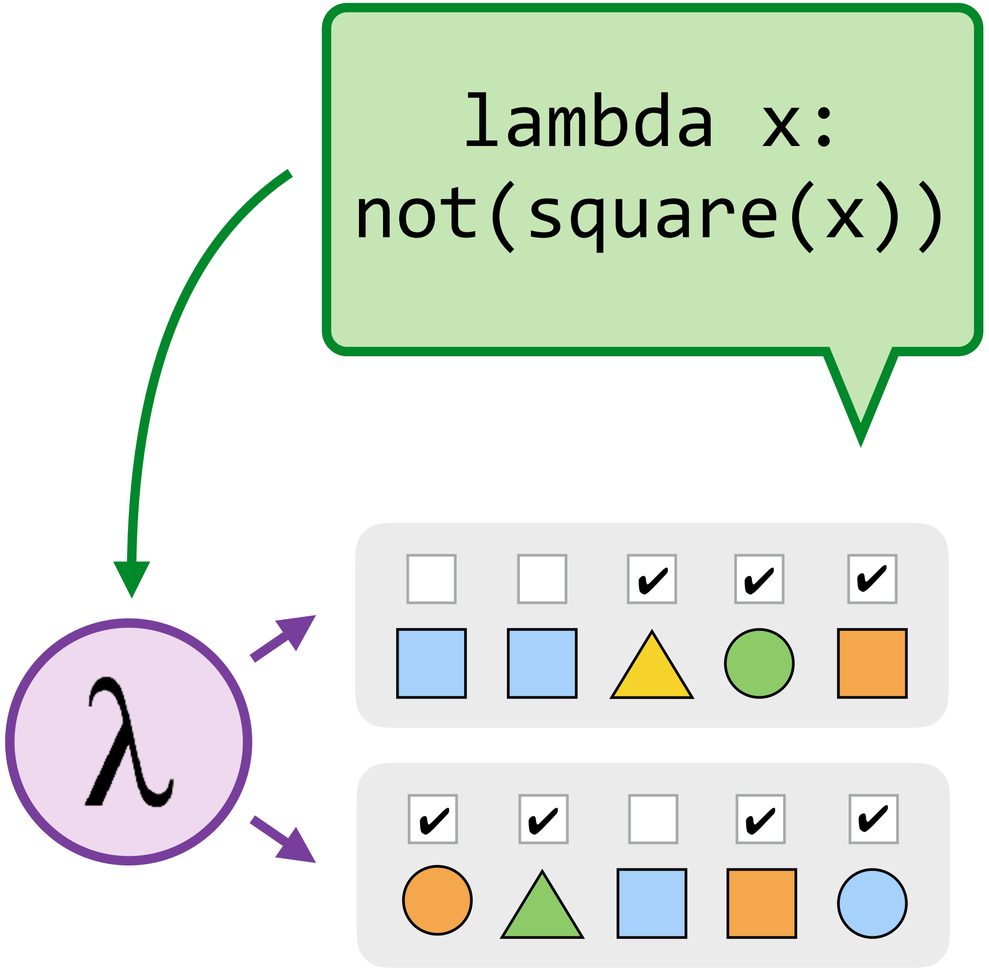
Computing meaning representations





Computing meaning representations





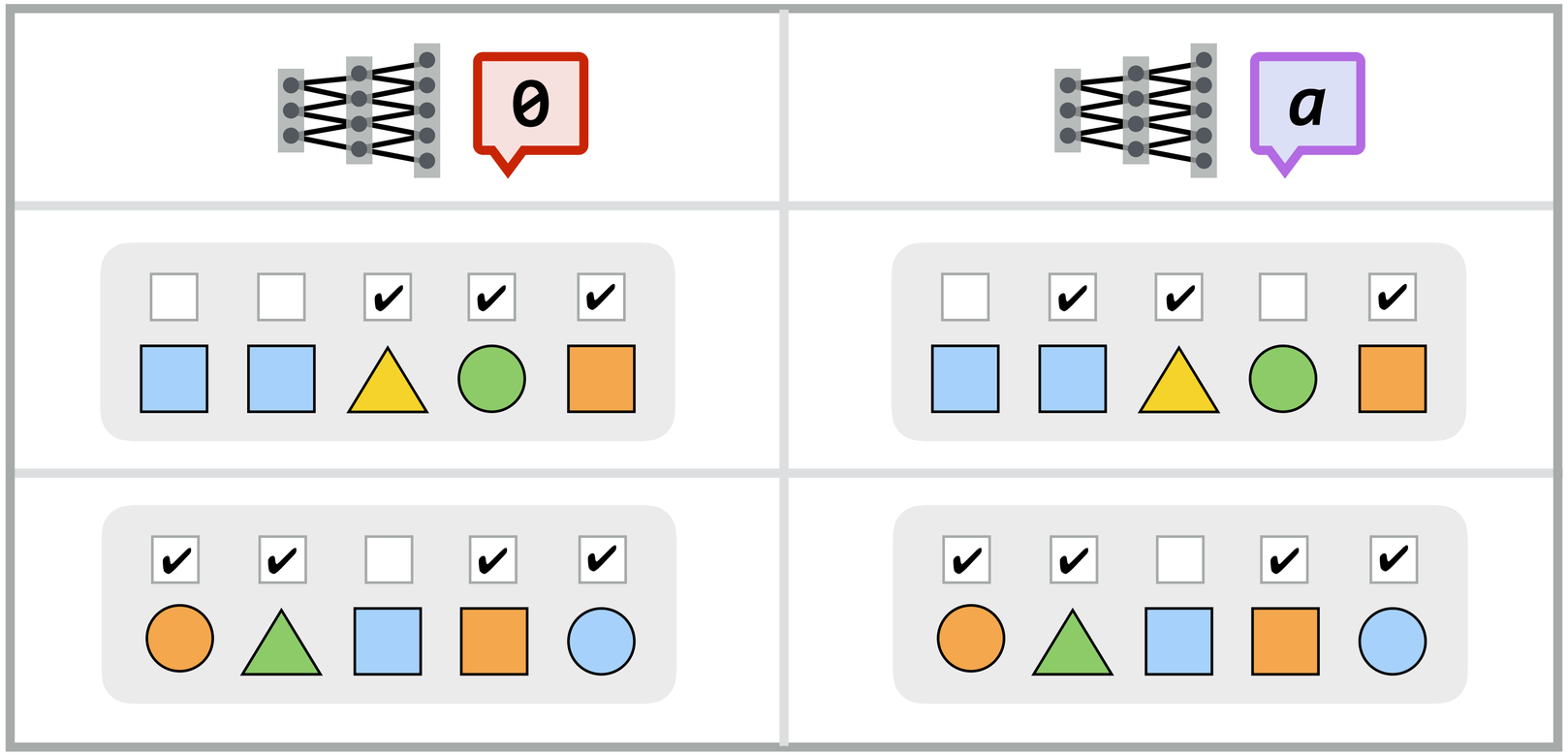
Translation criterion

$$q(o, a) = KL(\beta(o) \,\|\, \beta(a))$$

	o	a
	0.10	0.08
	0.05	0.01
	0.13	0.22

Translation criterion

$$q(\text{0}, \textit{a}) = \mathbf{E}[\,\beta(\text{0}) = \beta(\textit{a})]$$



Experiments

"High-level" communicative behavior

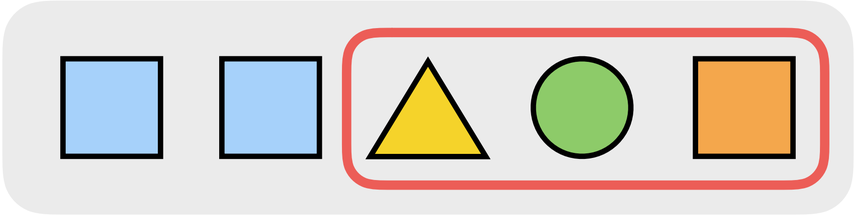
"Low-level" message structure

Experiments

"High-level" communicative behavior

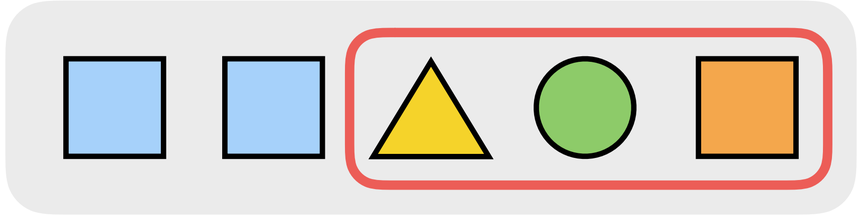
"Low-level" message structure

Comparing strategies



Comparing strategies

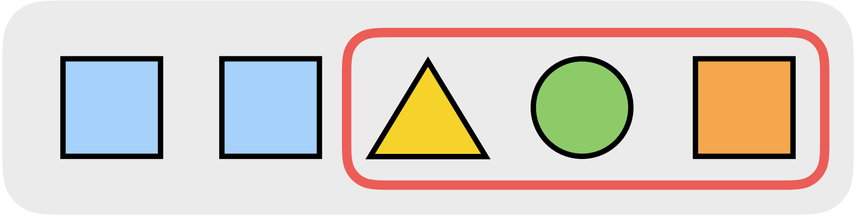
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```



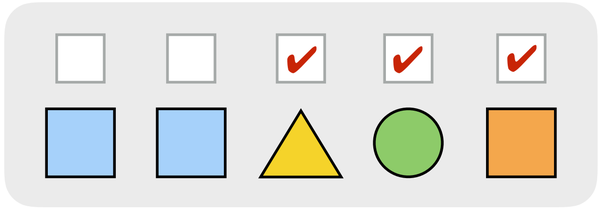
everything but squares

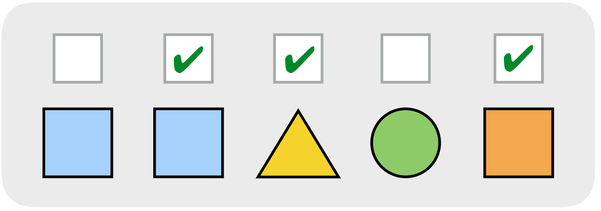
Comparing strategies

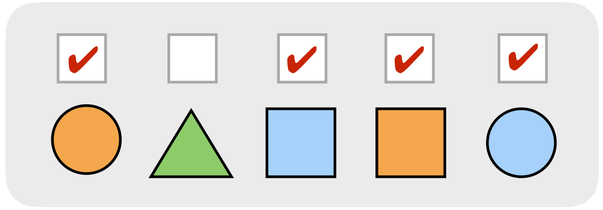
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

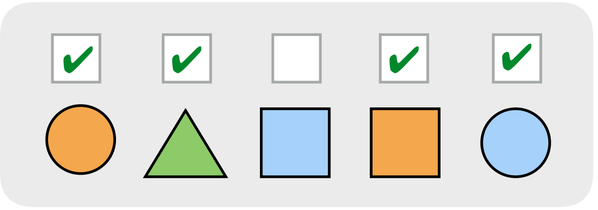


everything but squares



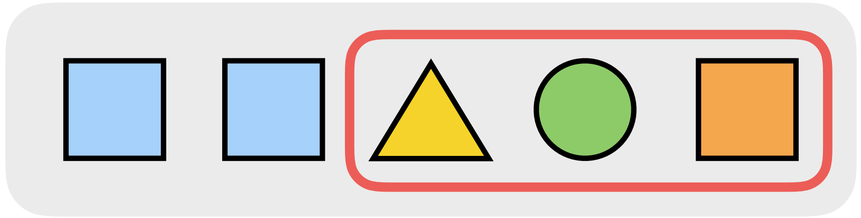




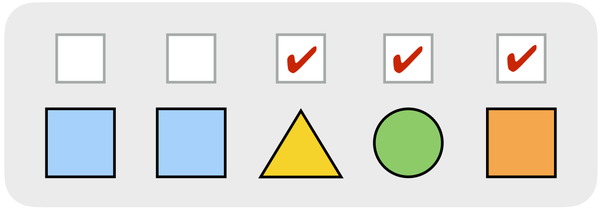


Comparing strategies

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

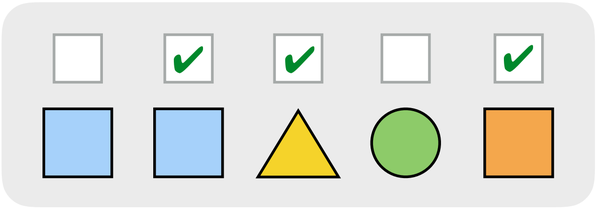


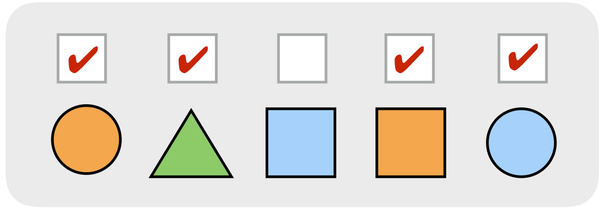
everything but squares

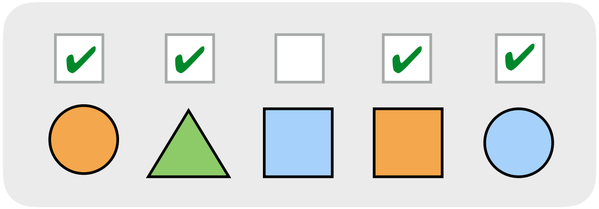


?

=

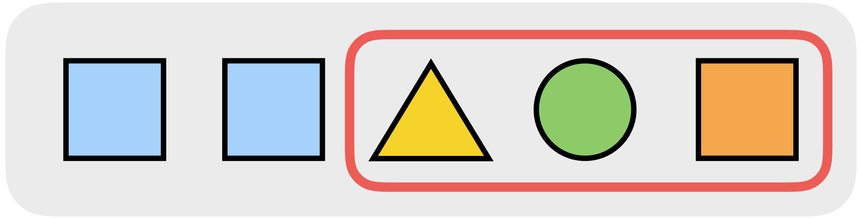


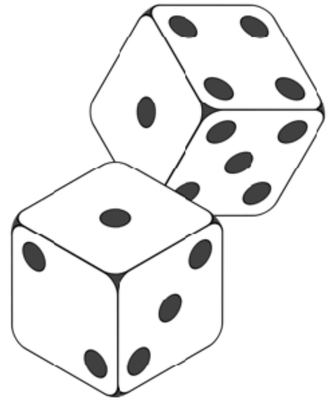


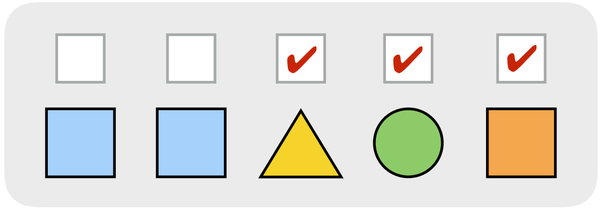


Theories of model behavior: random

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

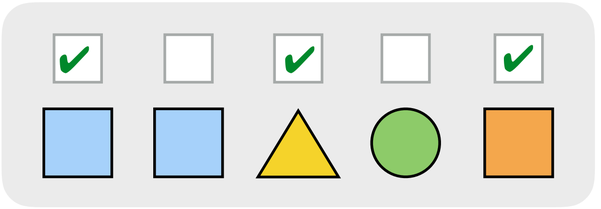


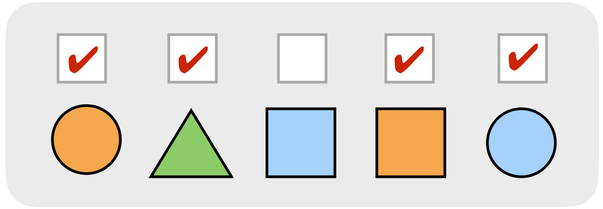


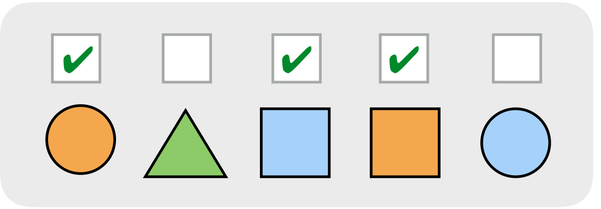


?

=

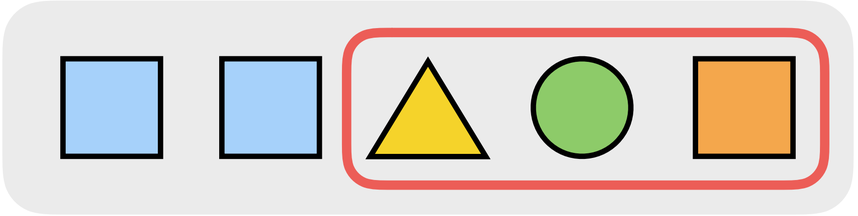


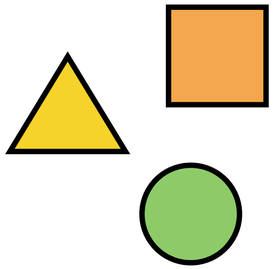


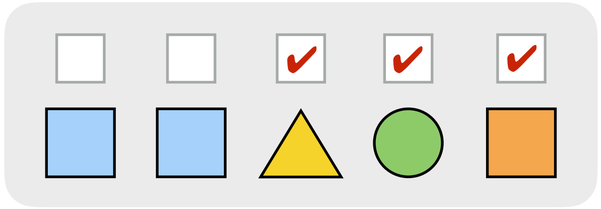


Theories of model behavior: literal

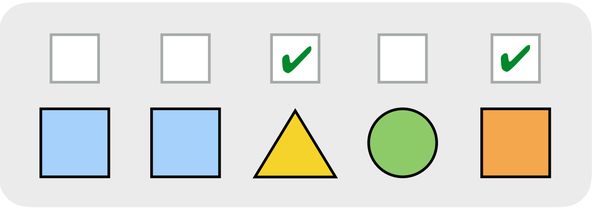
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

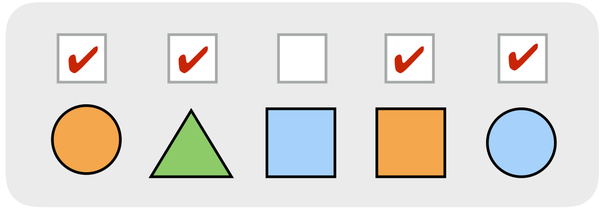


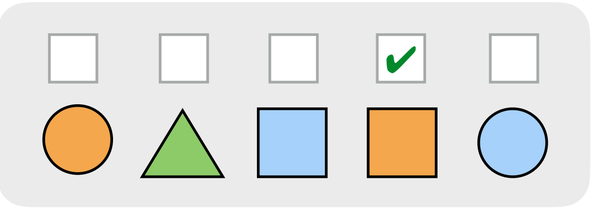




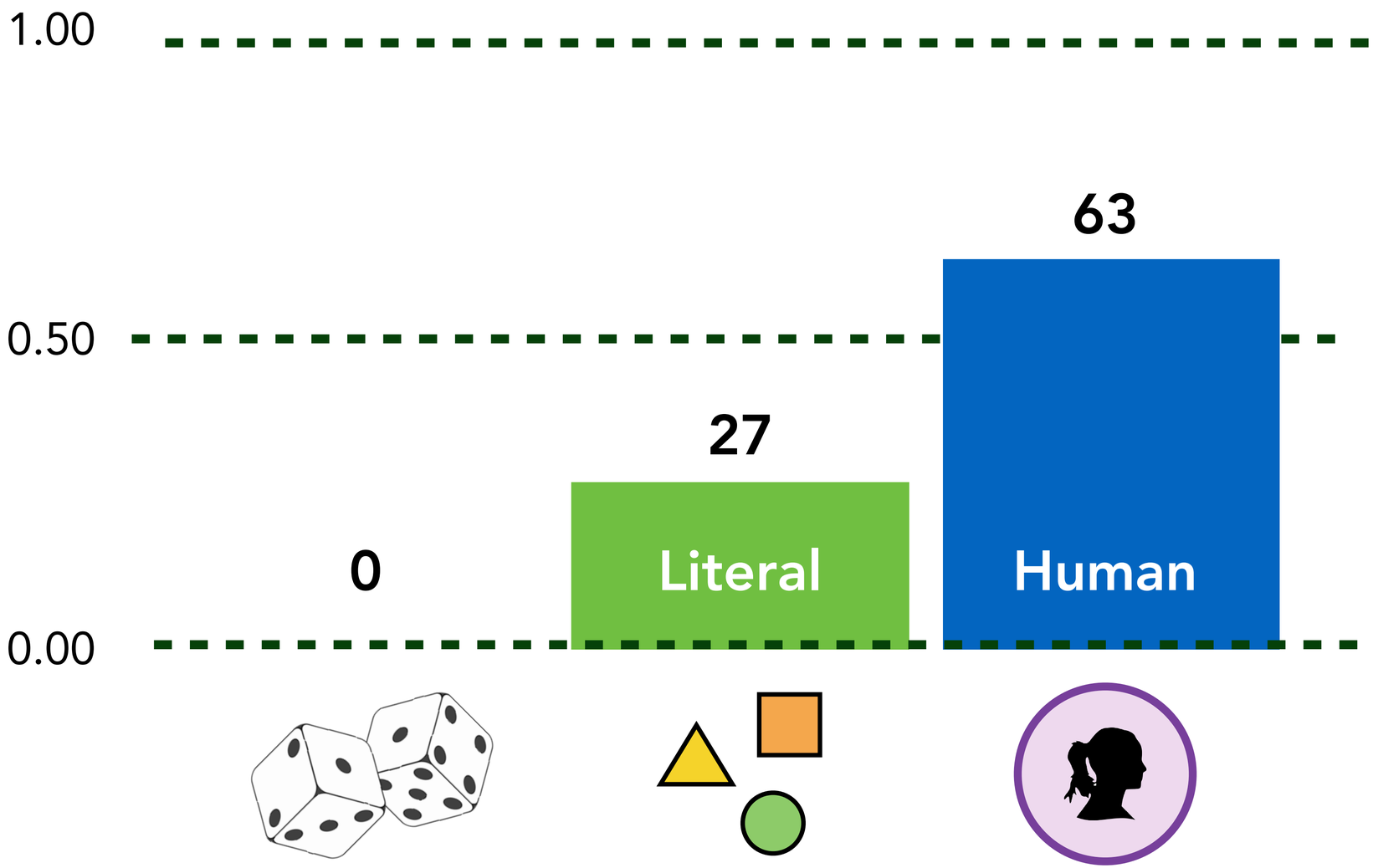
?



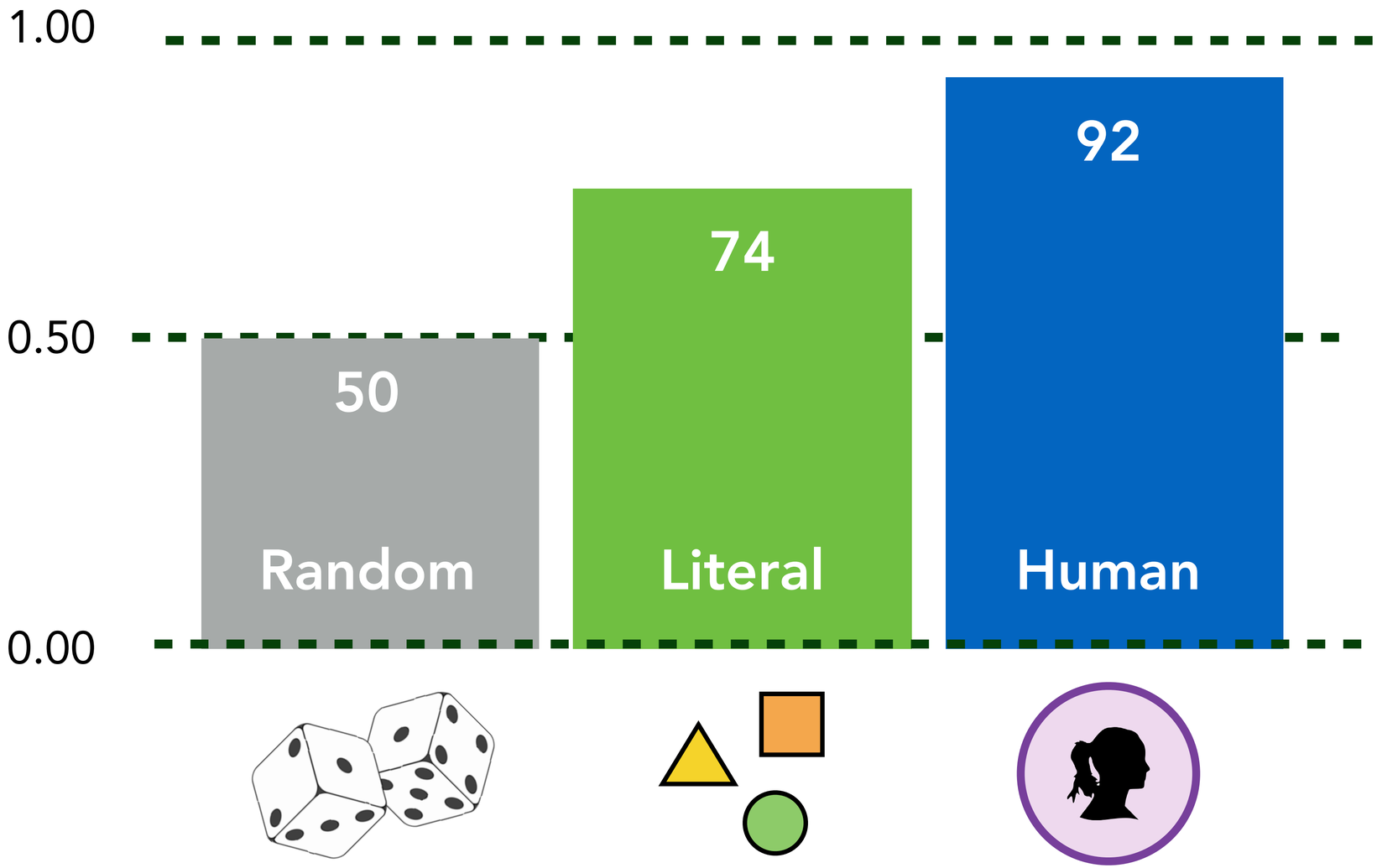




Evaluation: high-level scene agreement



Evaluation: high-level object agreement



Experiments

"High-level" communicative behavior

"Low-level" message structure

Collecting translation data

all the red shapes

blue objects

everything but red

green squares

not green squares

Collecting translation data

λx.red(x)

λx.blu(x)

λx.¬red(x)

λx.grn(x)∧sqr(x)

λx.¬(grn(x)∧sqr(x))

Collecting translation data

λx.red(x)	→	0.1 -0.3 0.5 1.1
λx.blu(x)	→	-0.3 0.2 0.1 0.1
λx.¬red(x)	→	1.4 -0.3 -0.5 0.8
λx.grn(x)∧sqr(x)	→	0.2 -0.2 0.5 -0.1
λx.¬(grn(x)∧sqr(x))	→	0.3 -1.3 -1.5 0.1

Extracting related pairs

λx.red(x)

0.1 -0.3 0.5 1.1

λx.¬red(x)

1.4 -0.3 -0.5 0.8

λx.grn(x)∧sqr(x)

0.2 -0.2 0.5 -0.1

λx.¬(grn(x)∧sqr(x))

0.3 -1.3 -1.5 0.1

Extracting related pairs

λx.red(x)

0.1 -0.3 0.5 1.1

λx.¬red(x)

1.4 -0.3 -0.5 0.8

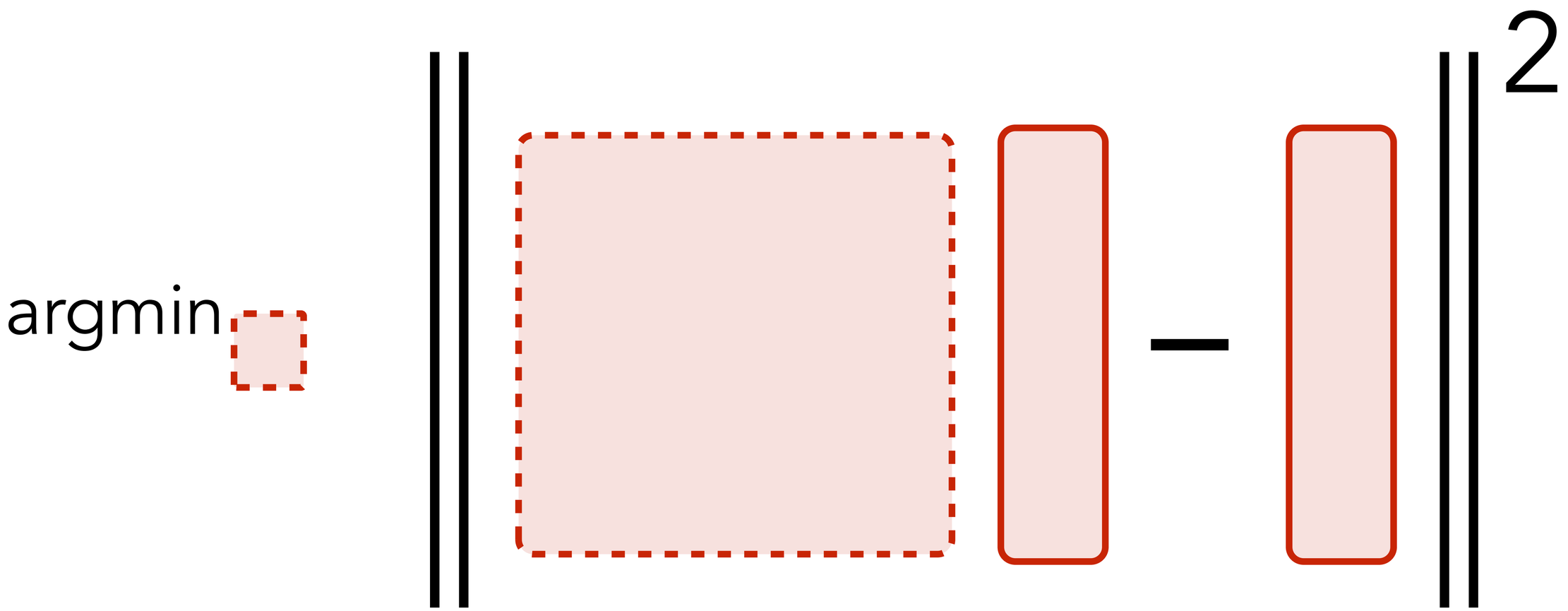
λx.grn(x)∧sqr(x)

0.2 -0.2 0.5 -0.1

λx.¬(grn(x)∧sqr(x))

0.3 -1.3 -1.5 0.1

Learning compositional operators



Evaluating learned operators

λx.red(x)

0.1 -0.3 0.5 1.1

λx.¬red(x)

1.4 -0.3 -0.5 0.8

λx.grn(x)∧sqr(x)

0.2 -0.2 0.5 -0.1

λx.¬(grn(x)∧sqr(x))

0.3 -1.3 -1.5 0.1

λx.f(x)

0.2 -0.2 0.5 -0.1

Evaluating learned operators

λx.red(x)

0.1 -0.3 0.5 1.1

λx.¬red(x)

1.4 -0.3 -0.5 0.8

λx.grn(x)∧sqr(x)

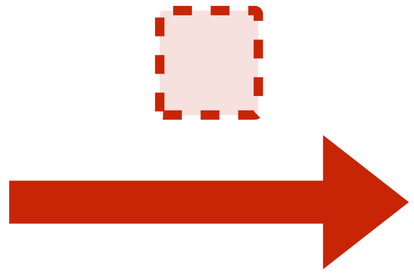
0.2 -0.2 0.5 -0.1

λx.¬(grn(x)∧sqr(x))

0.3 -1.3 -1.5 0.1

λx.f(x)

0.2 -0.2 0.5 -0.1



-0.2 0.4 -0.3 0.0

Evaluating learned operators

λx.red(x)

0.1 -0.3 0.5 1.1

λx.¬red(x)

1.4 -0.3 -0.5 0.8

λx.grn(x)∧sqr(x)

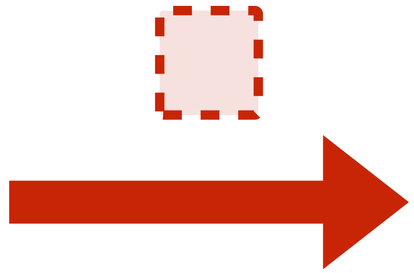
0.2 -0.2 0.5 -0.1

λx.¬(grn(x)∧sqr(x))

0.3 -1.3 -1.5 0.1

λx.f(x)

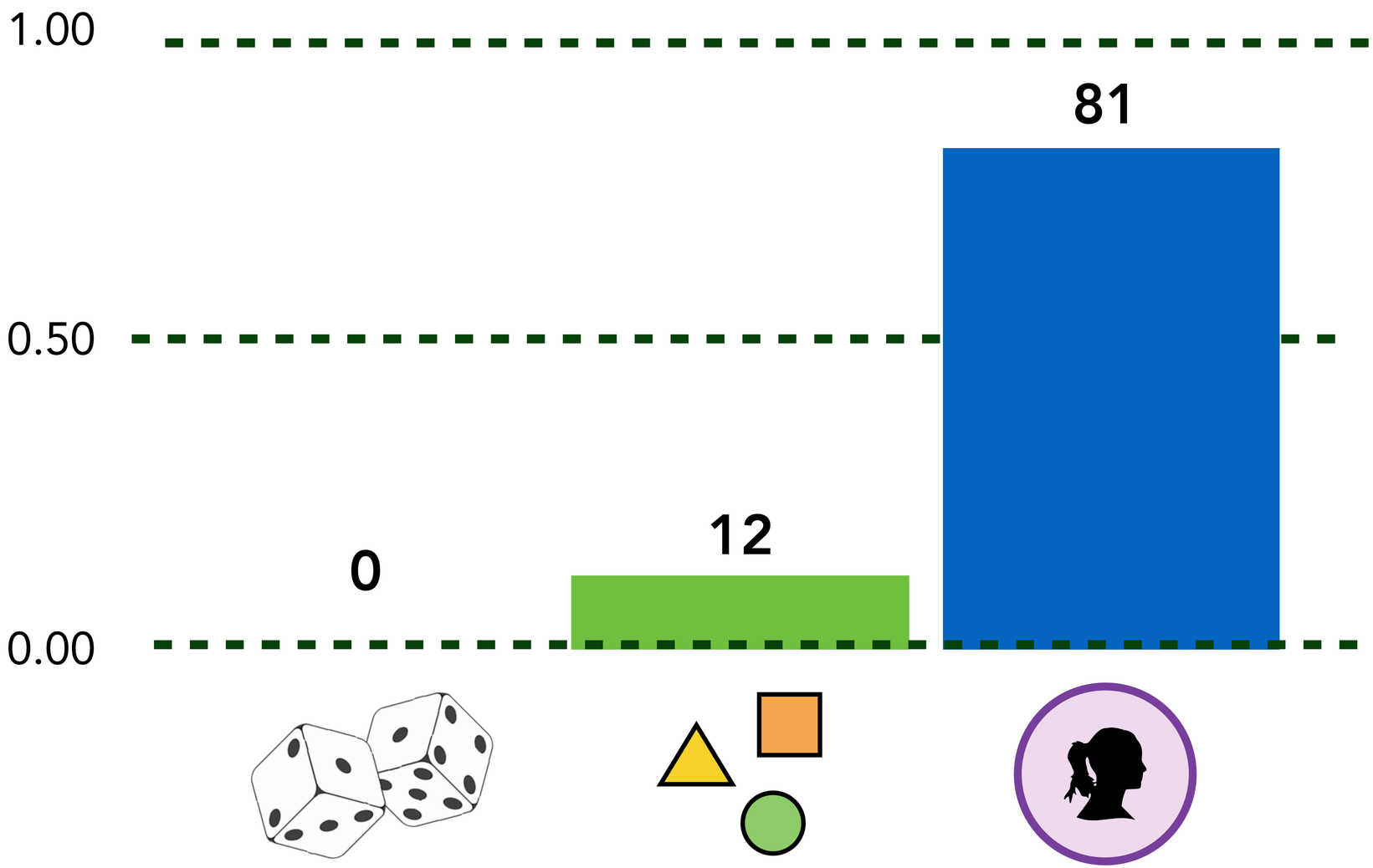
0.2 -0.2 0.5 -0.1



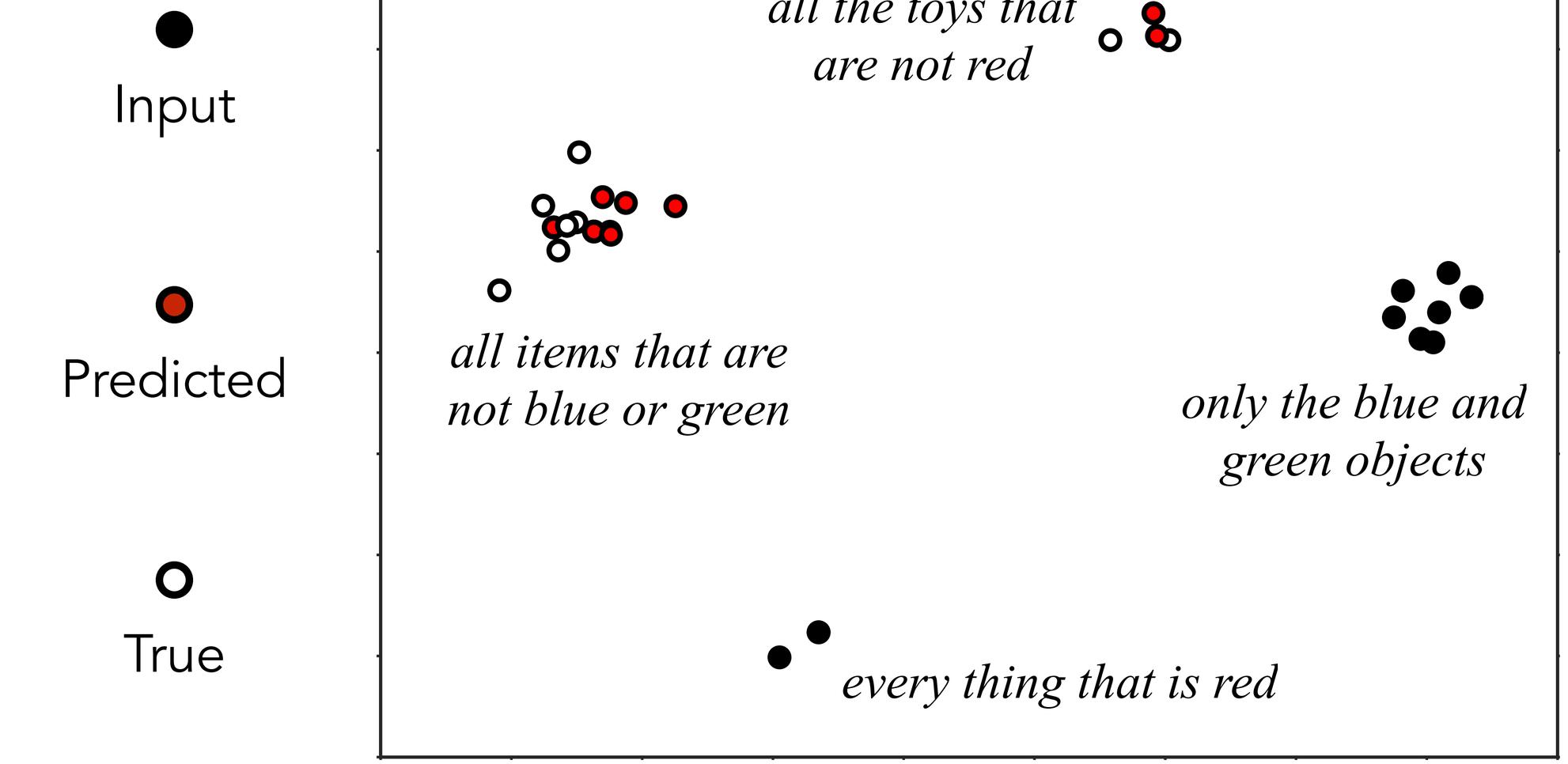
???

-0.2 0.4 -0.3 0.0

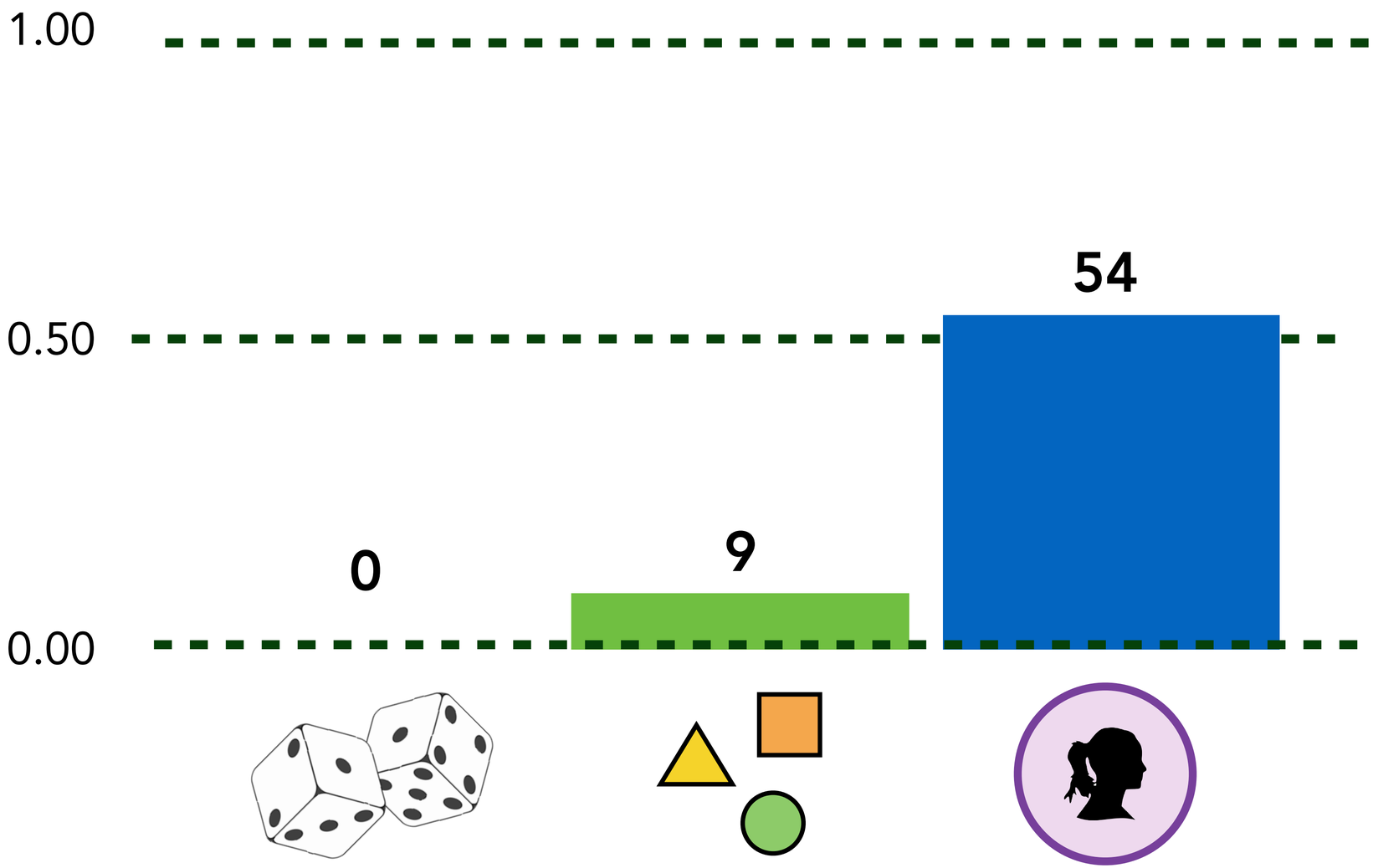
Evaluation: scene agreement for negation



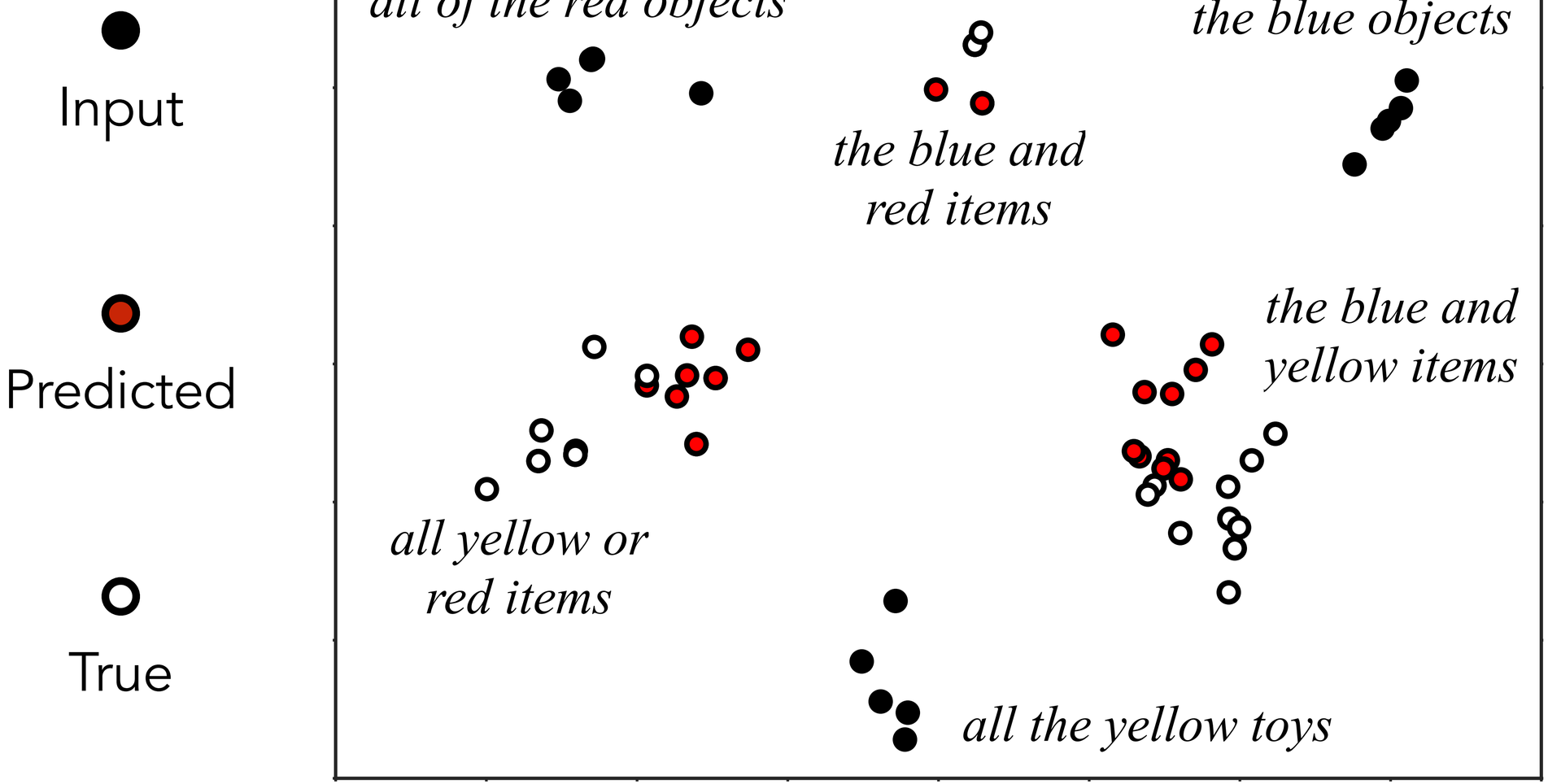
Visualizing negation



Evaluation: scene agreement for disjunction



Visualizing disjunction



Conclusions

- We can translate between neuralese and natural lang. by grounding in distributions over world states
- Under the right conditions, neuralese exhibits interpretable pragmatics & compositional structure
- Not just communication games—language might be a good general-purpose tool for interpreting deep reprs.

Conclusions

- We can translate between neuralese and natural lang. by grounding in distributions over world states

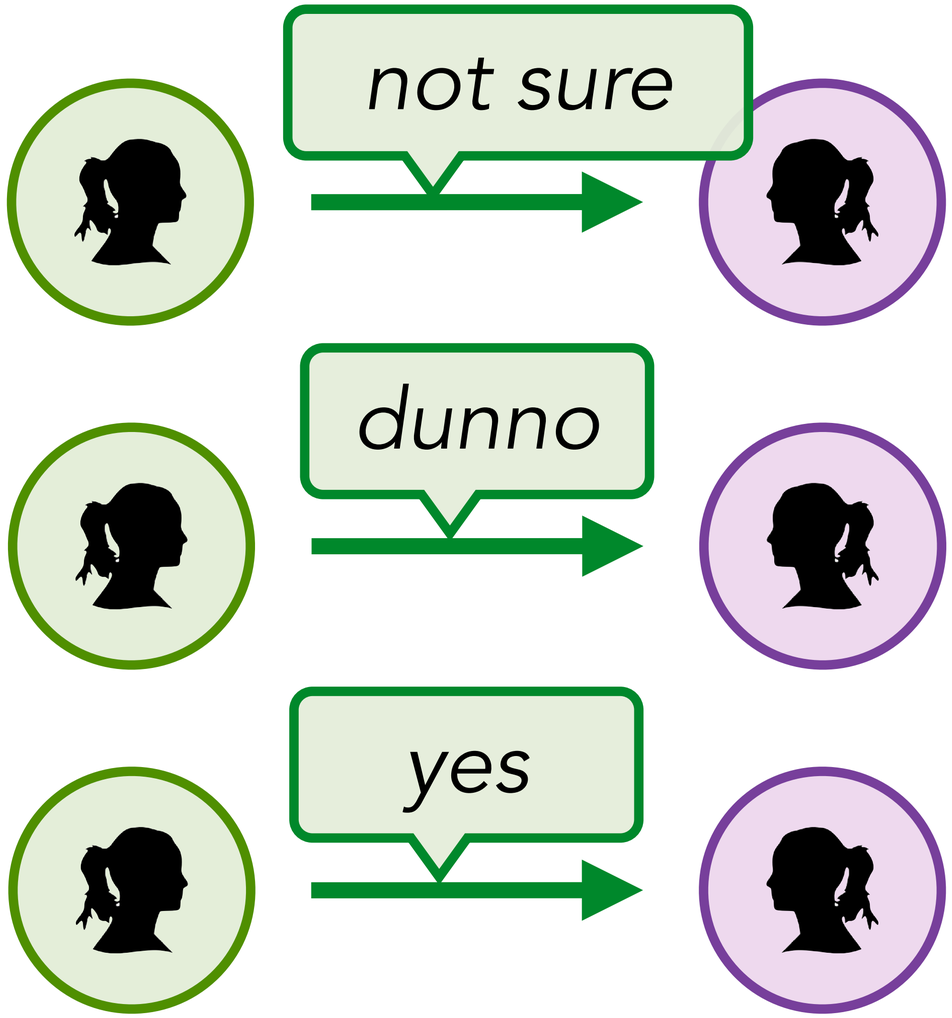
- Under the right conditions, neuralese exhibits interpretable pragmatics & compositional structure

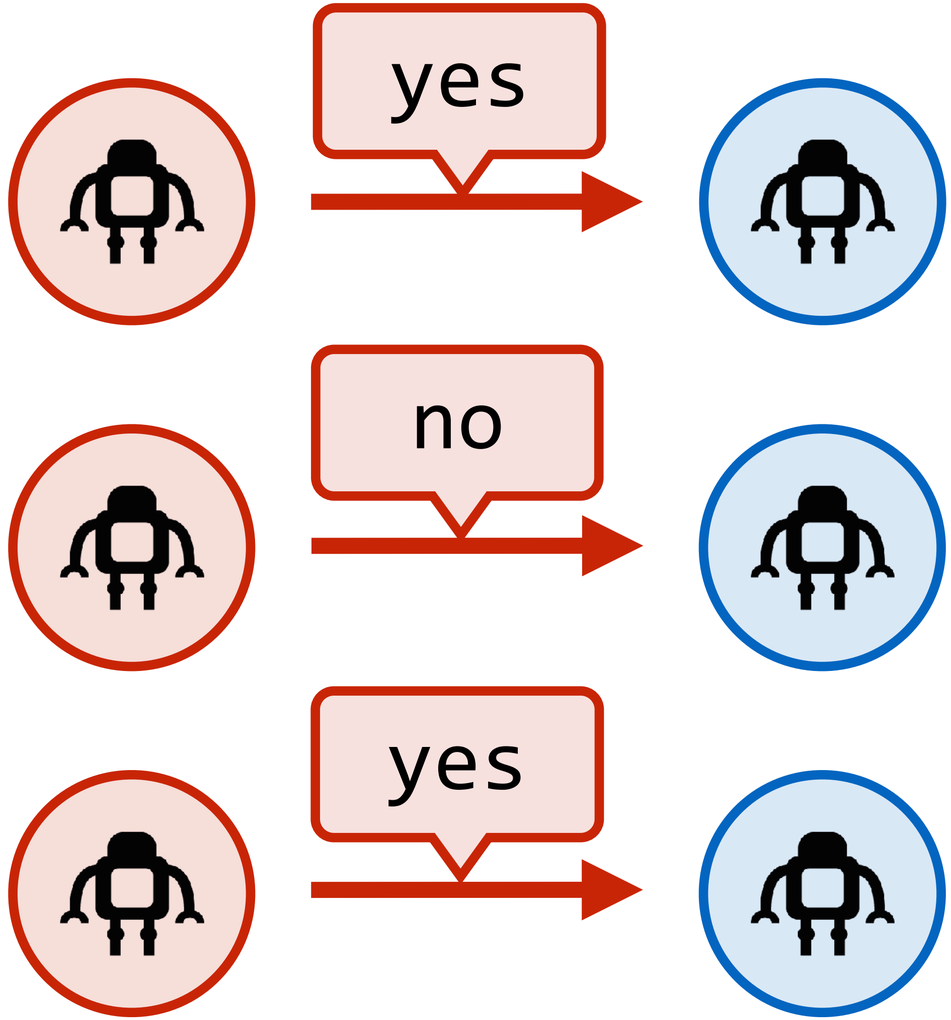
- Not just communication games—language might be a good general-purpose tool for interpreting deep reprs.

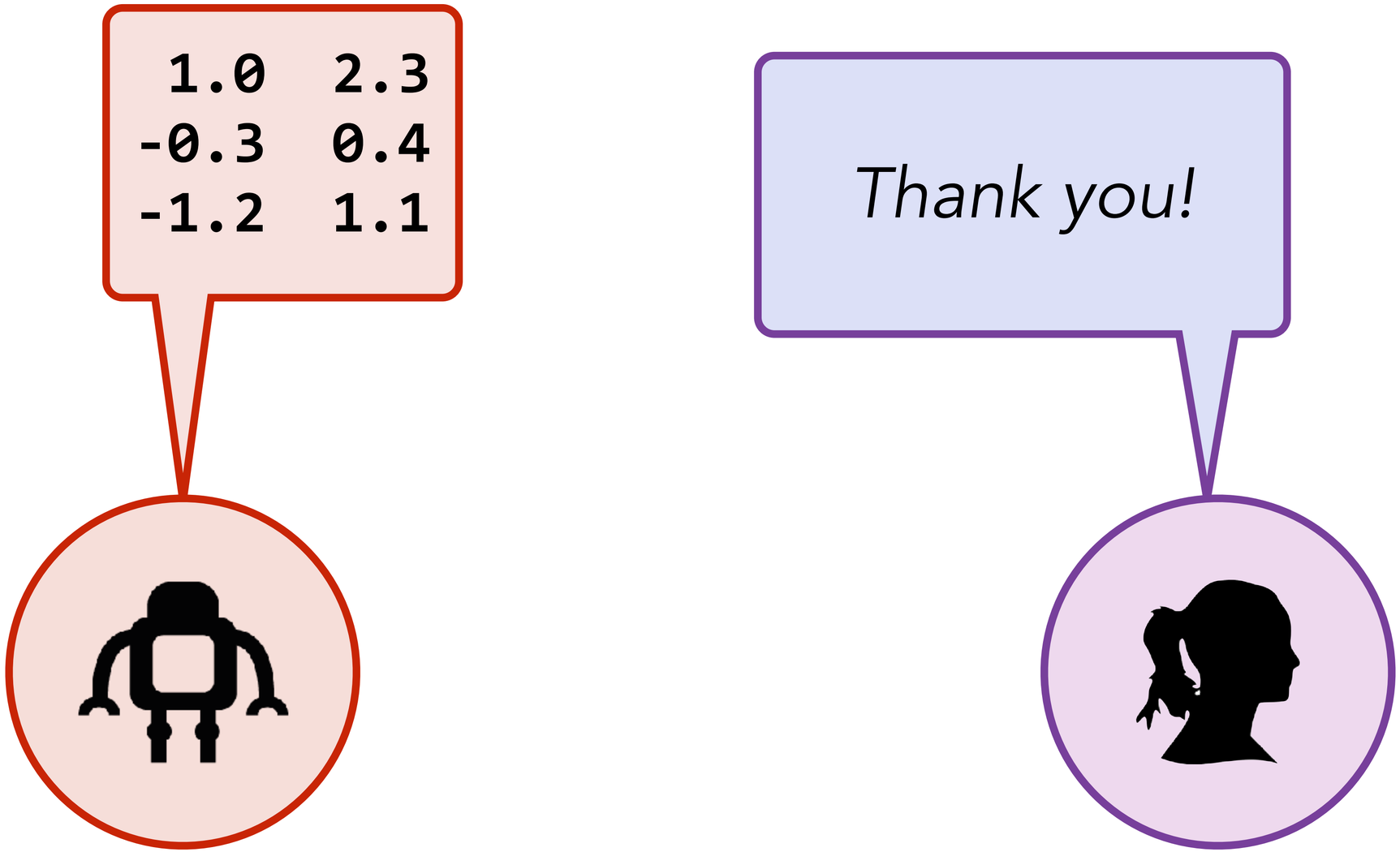
Conclusions

- We can translate between neuralese and natural lang. by grounding in distributions over world states
- Under the right conditions, neuralese exhibits interpretable pragmatics & compositional structure
- Not just communication games—language might be a good general-purpose tool for interpreting deep reprs.

Conclusions







http://github.com/jacobandreas/{neuralese,rnn-syn}