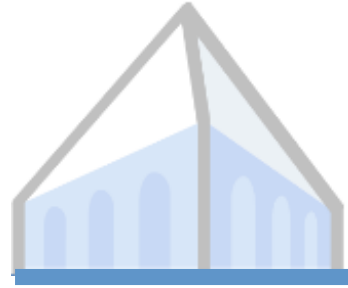


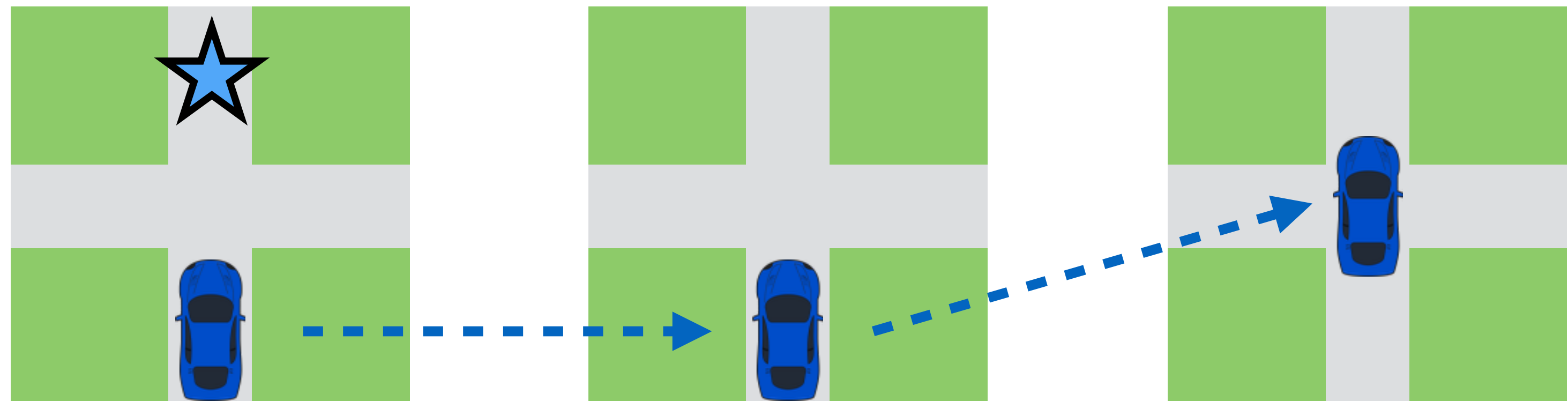
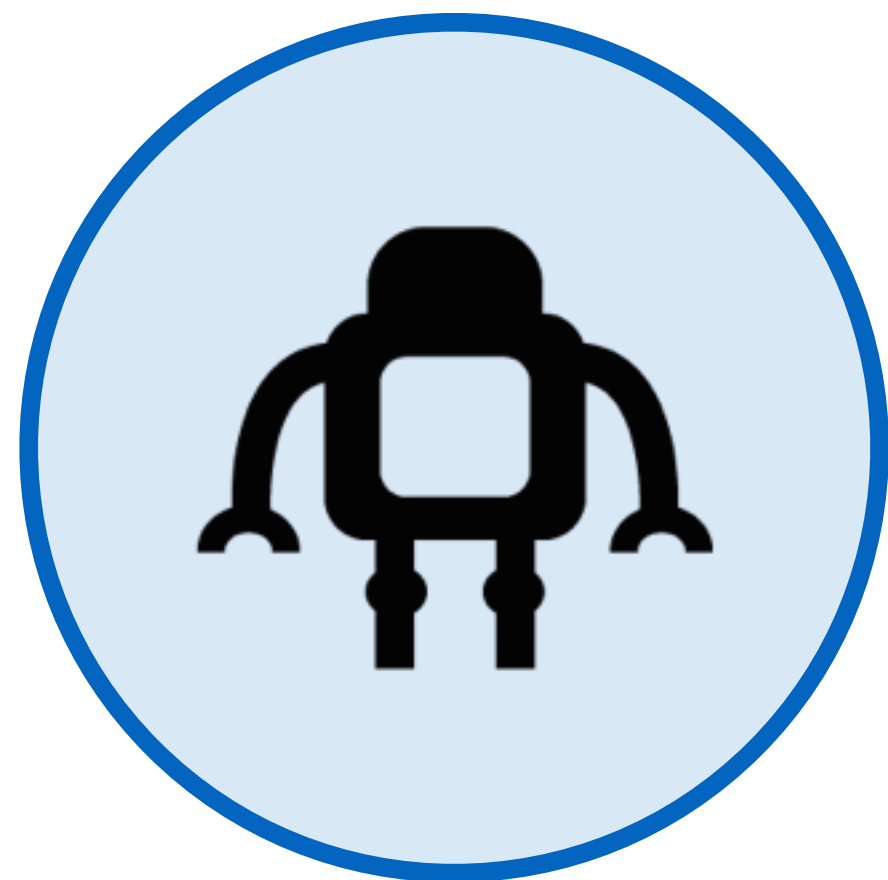
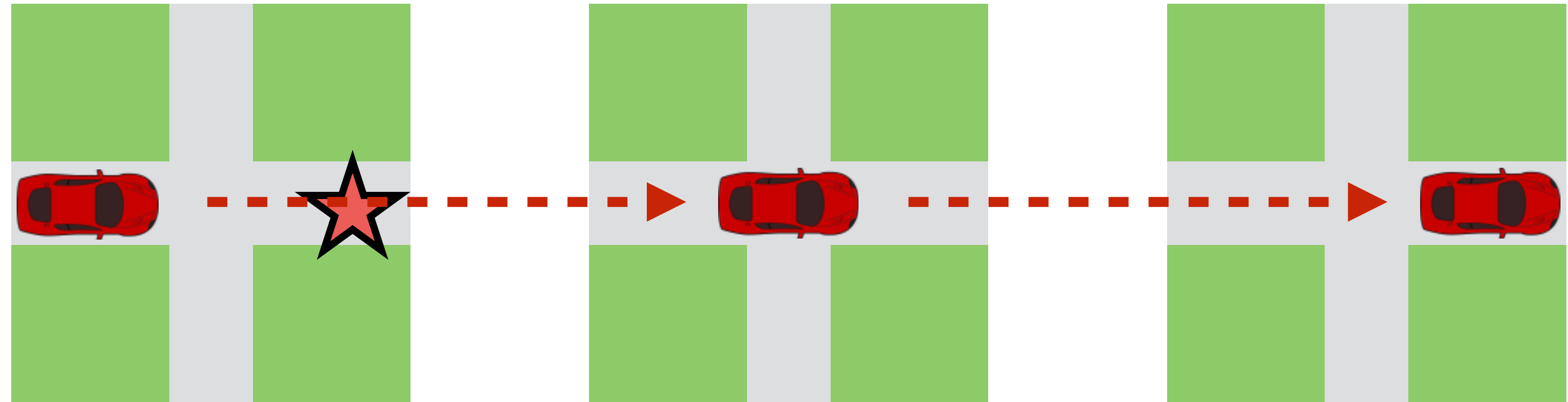
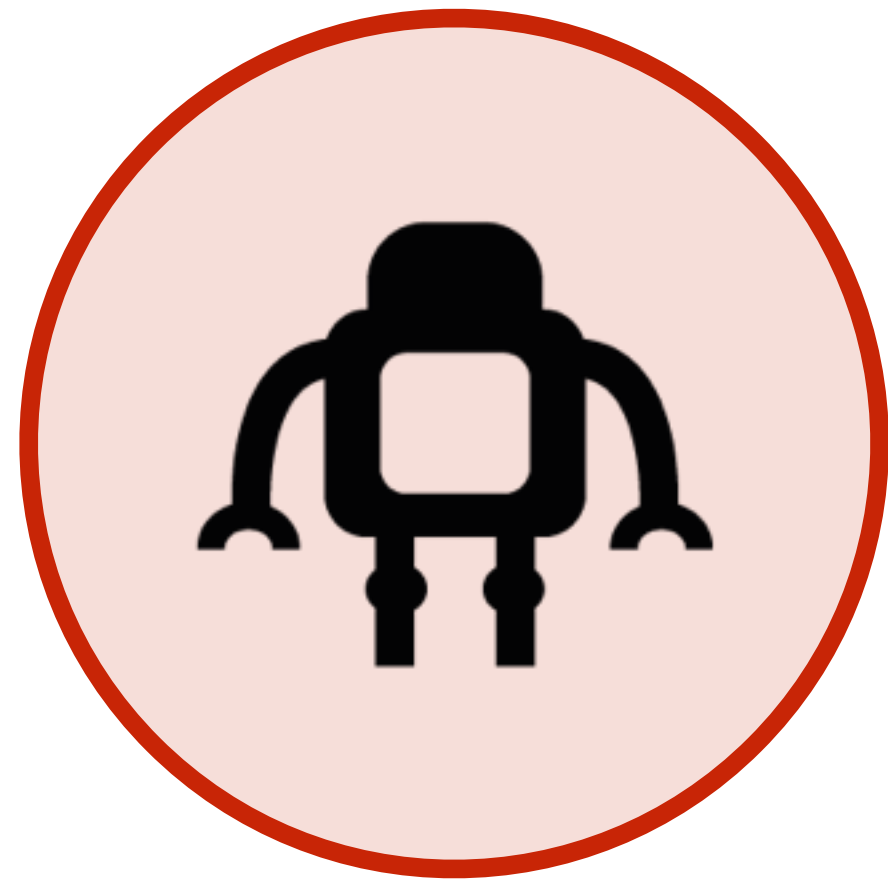
Translating Neuralese



Jacob Andreas, Anca Dragan, and Dan Klein

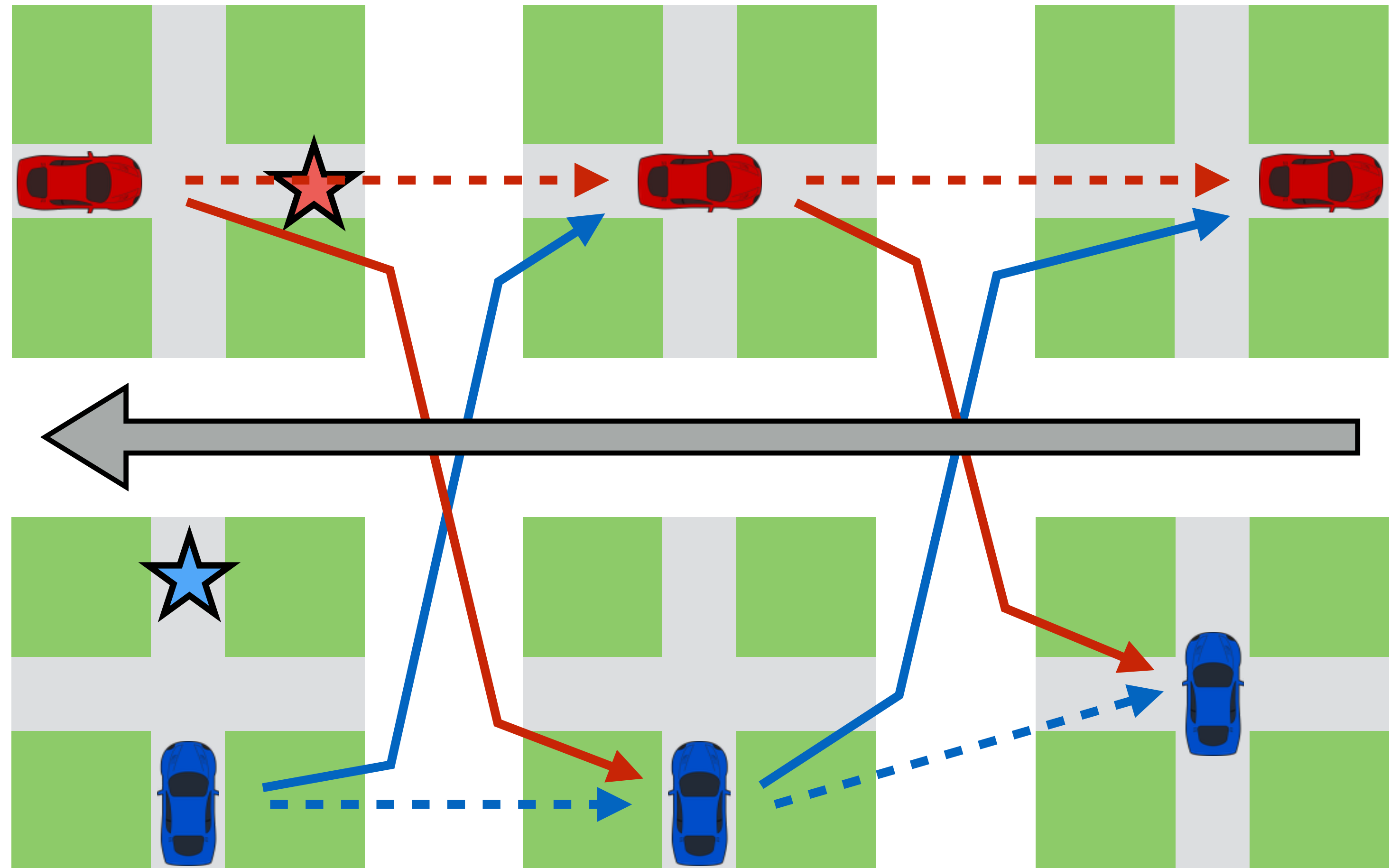
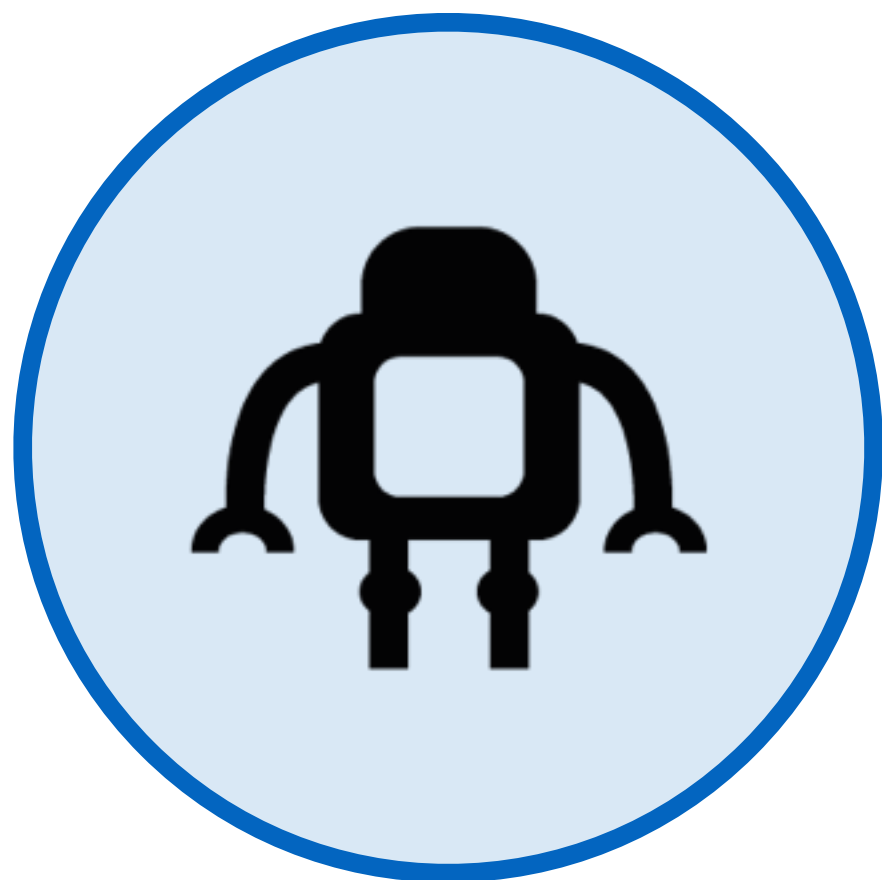
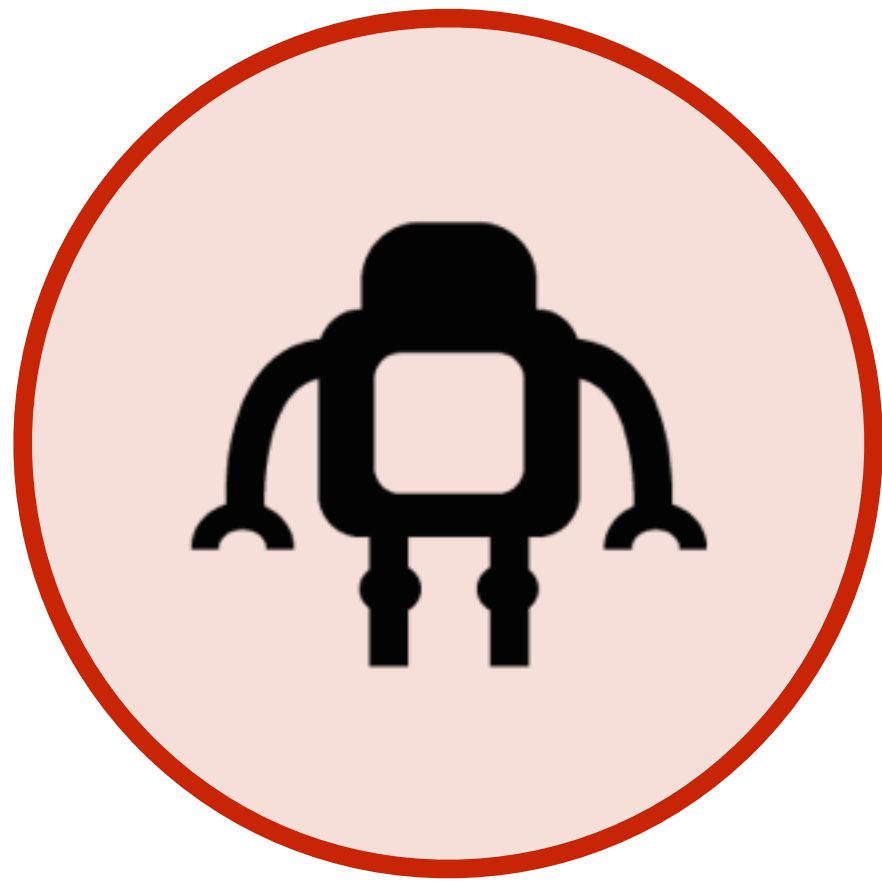


Learning to Communicate



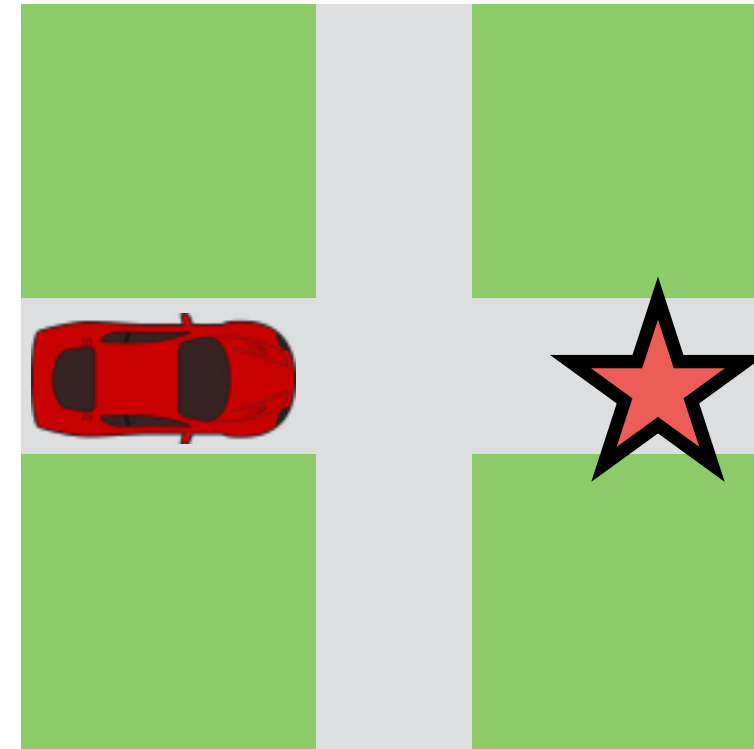
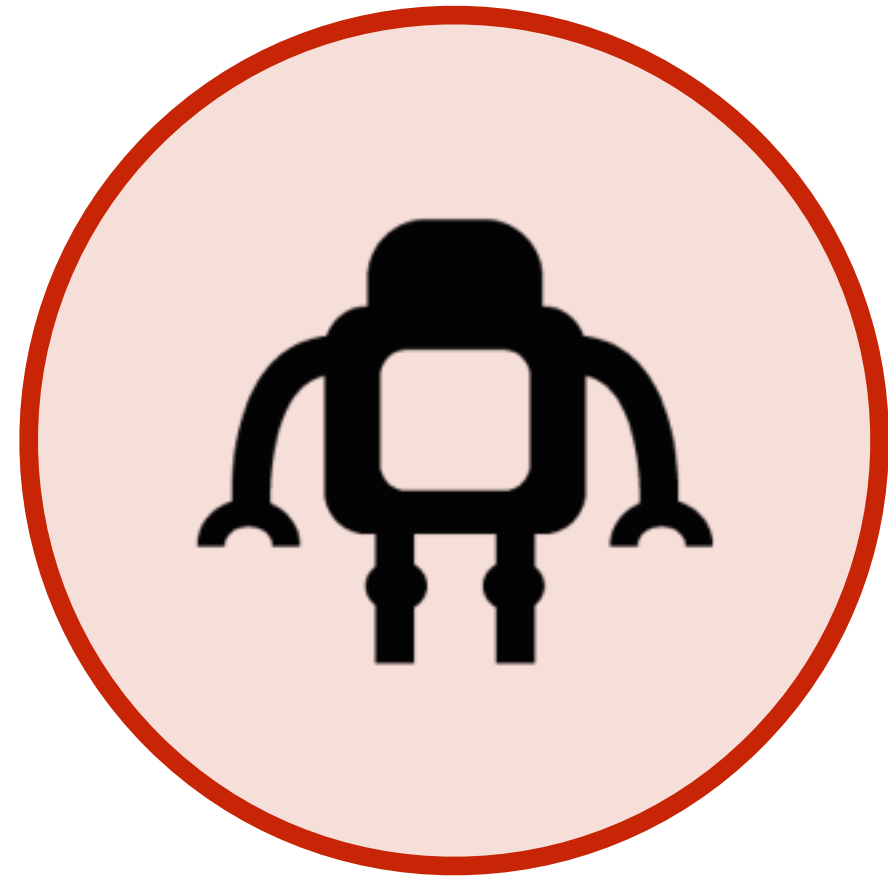


Learning to Communicate

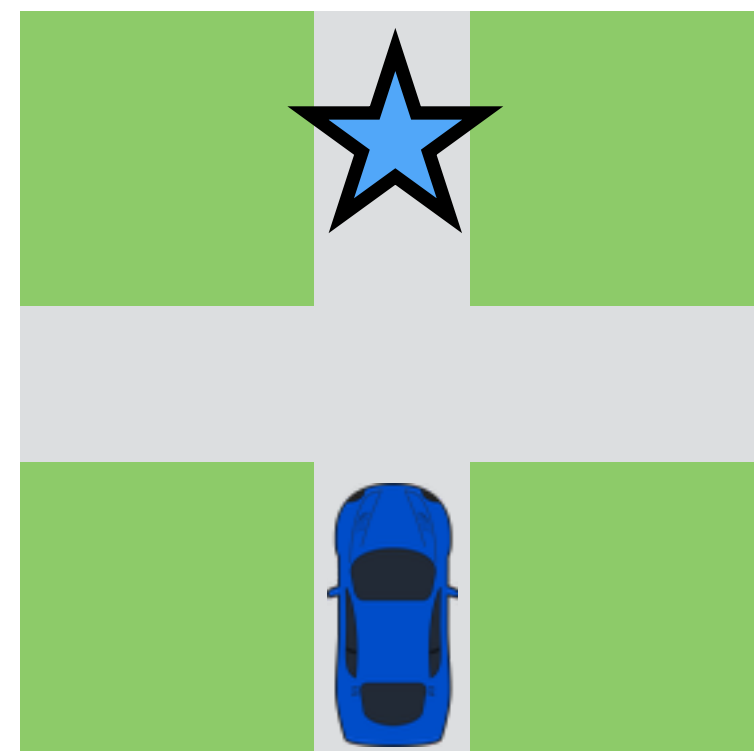
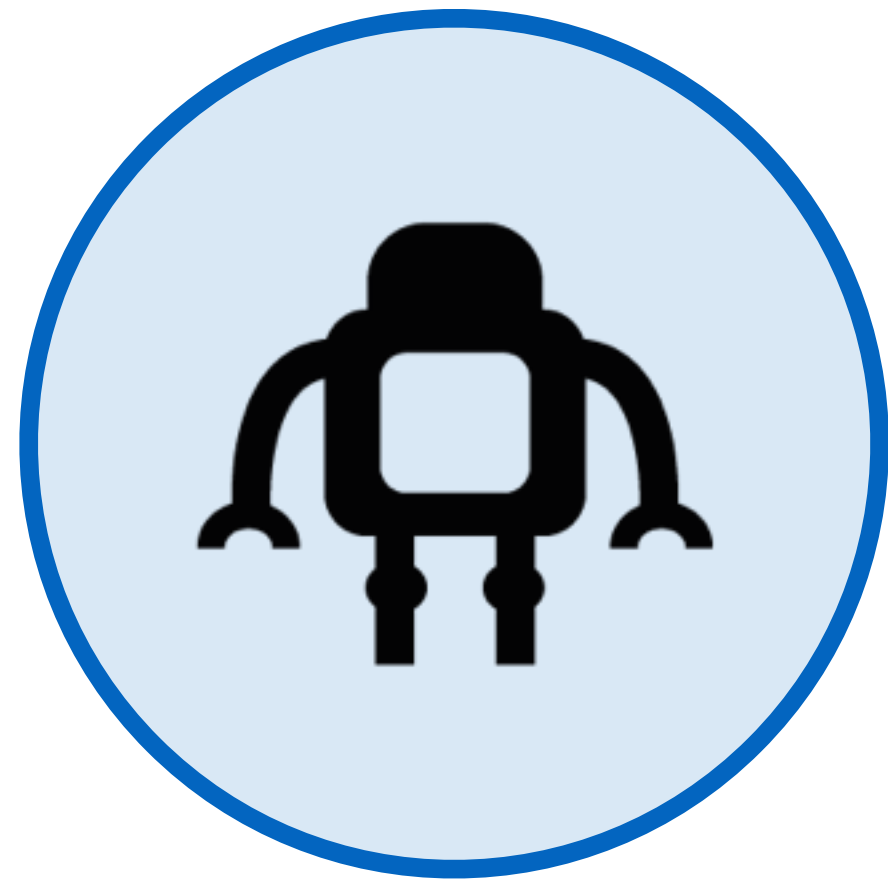
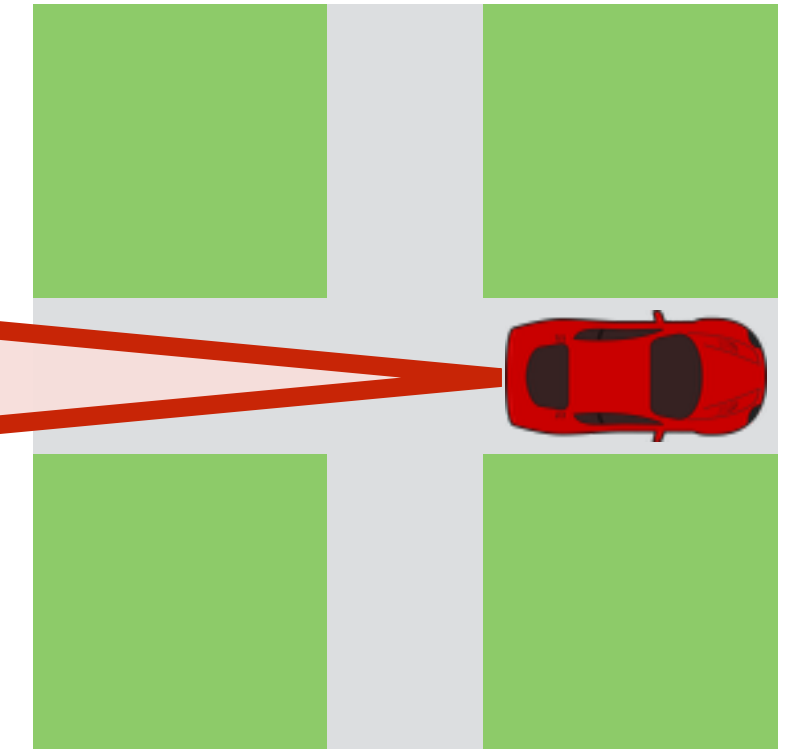




Neurales

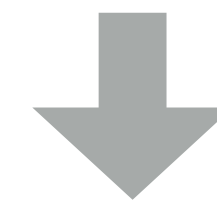
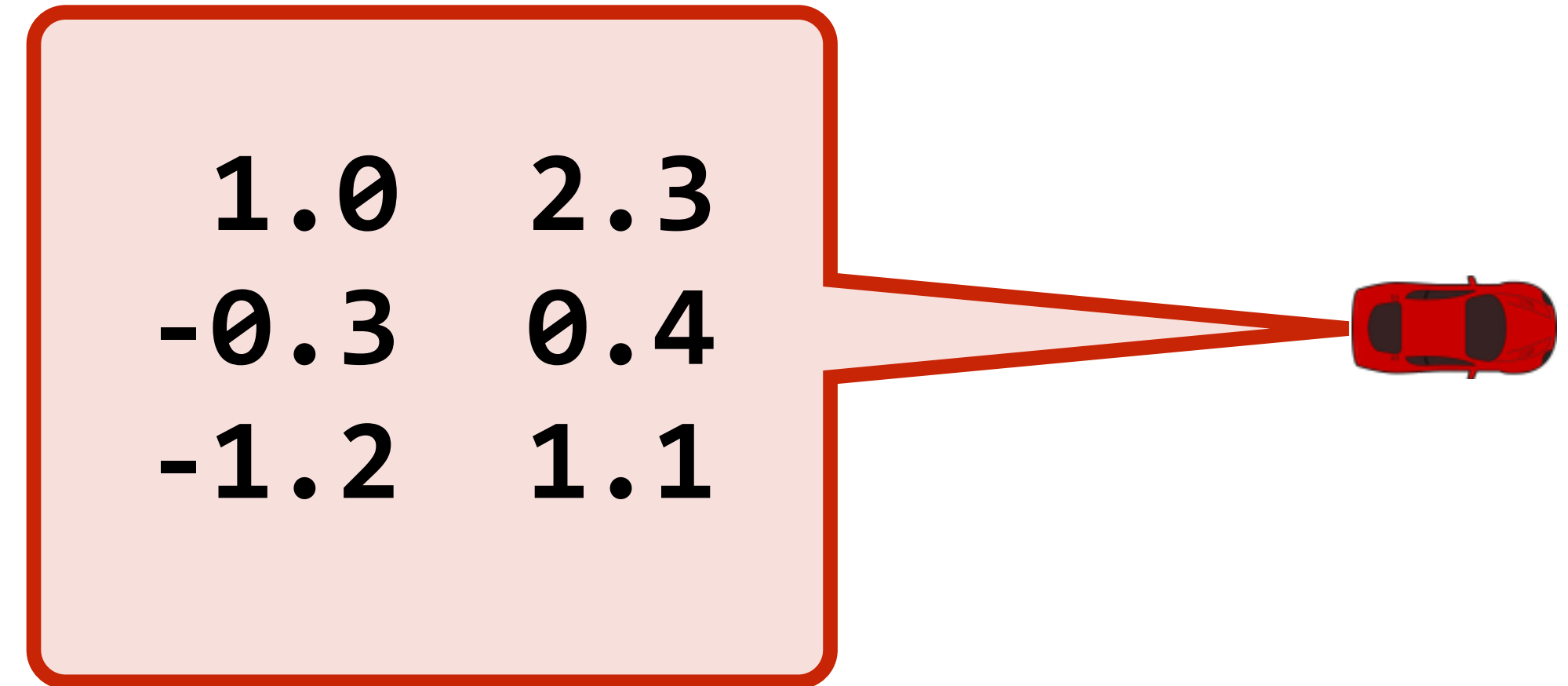
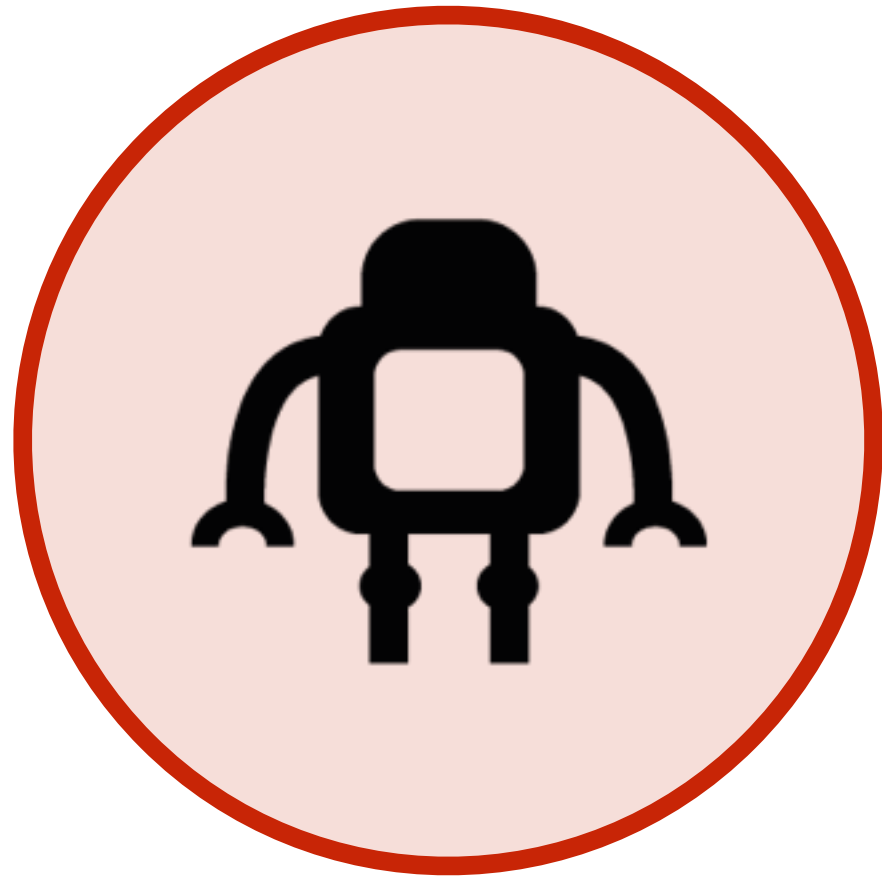


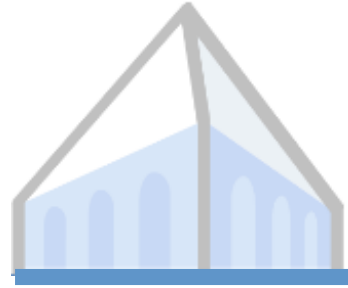
1.0	2.3
-0.3	0.4
-1.2	1.1





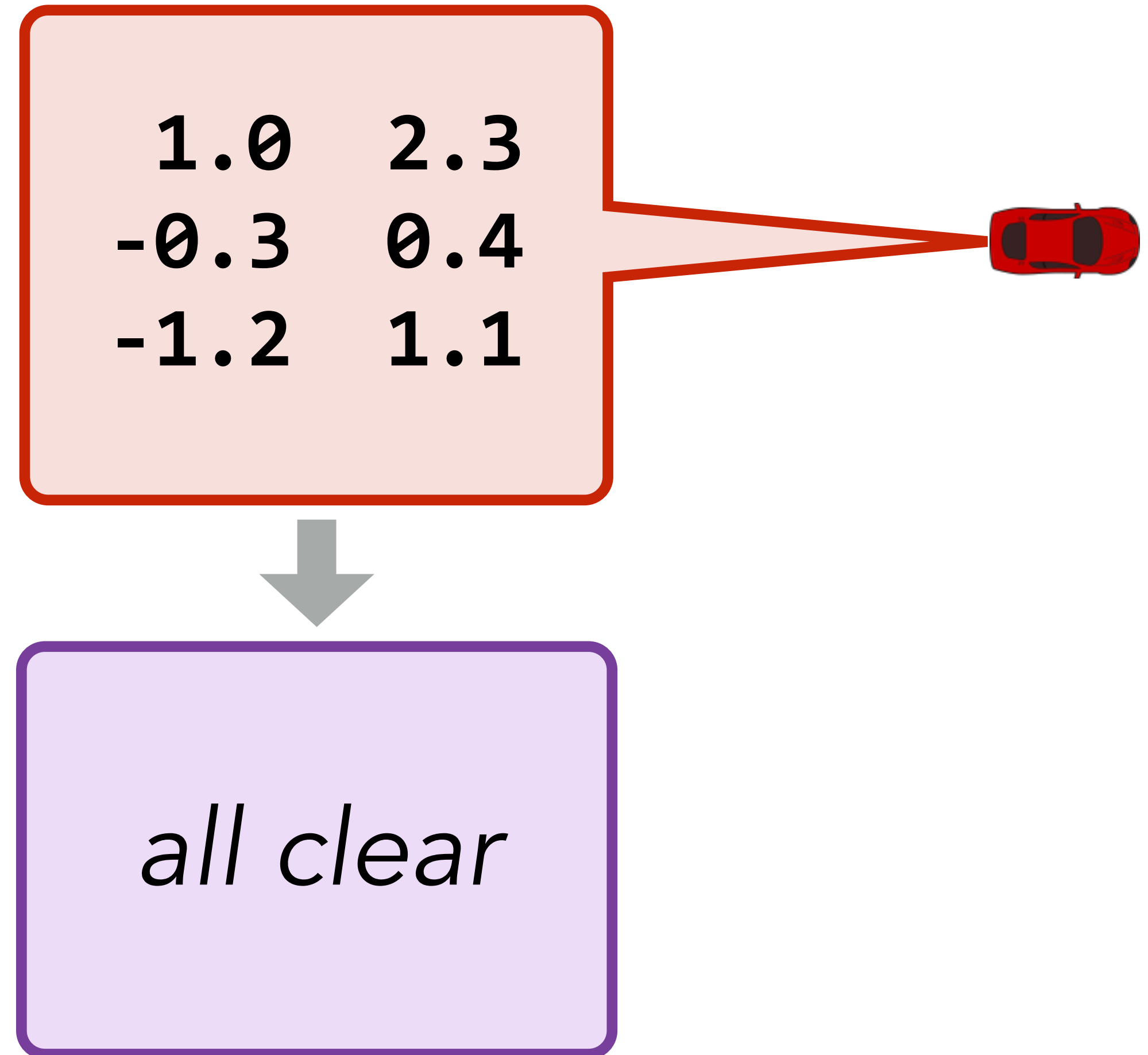
Translating neuralesse





Translating neuralese

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





Outline

Natural language & neuralese

Statistical machine translation

Semantic machine translation

Implementation details

Evaluation



Outline

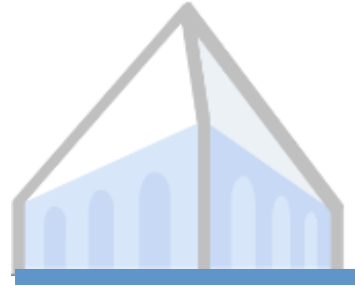
Natural language & neuralese

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

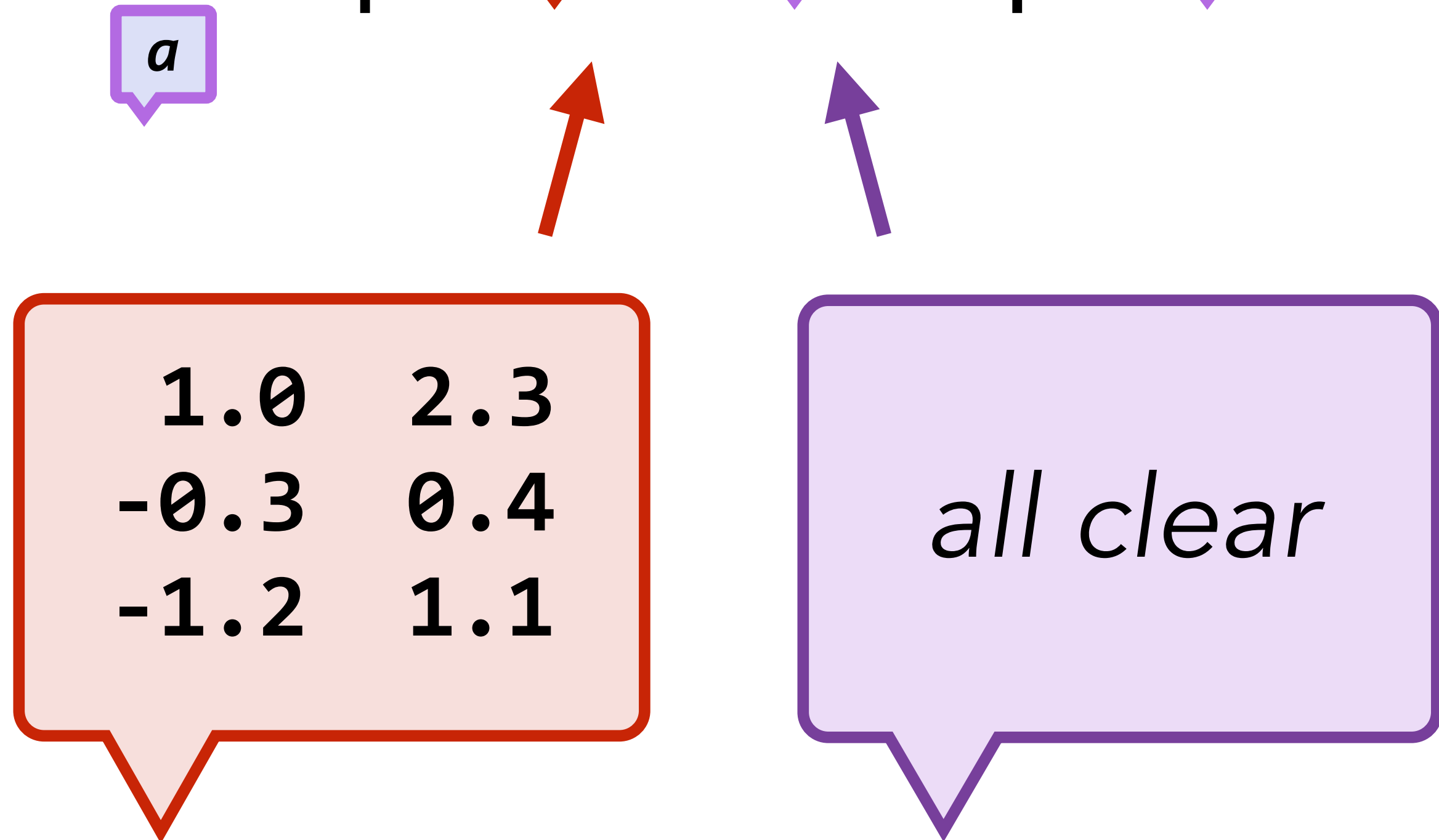
Implementation details

Evaluation



A statistical MT problem

$$\max_a p(\theta | a) p(a)$$

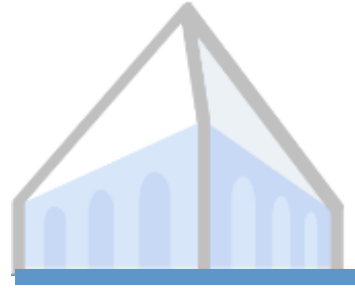




A statistical MT problem



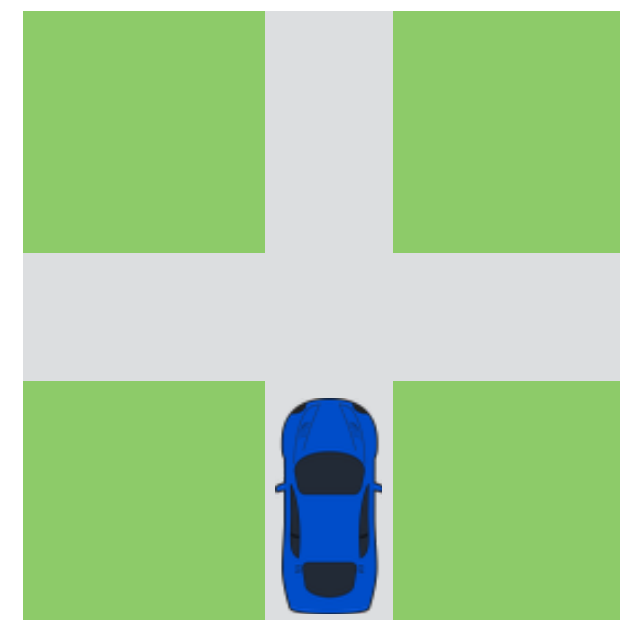
How do we induce a translation model?



A statistical MT problem

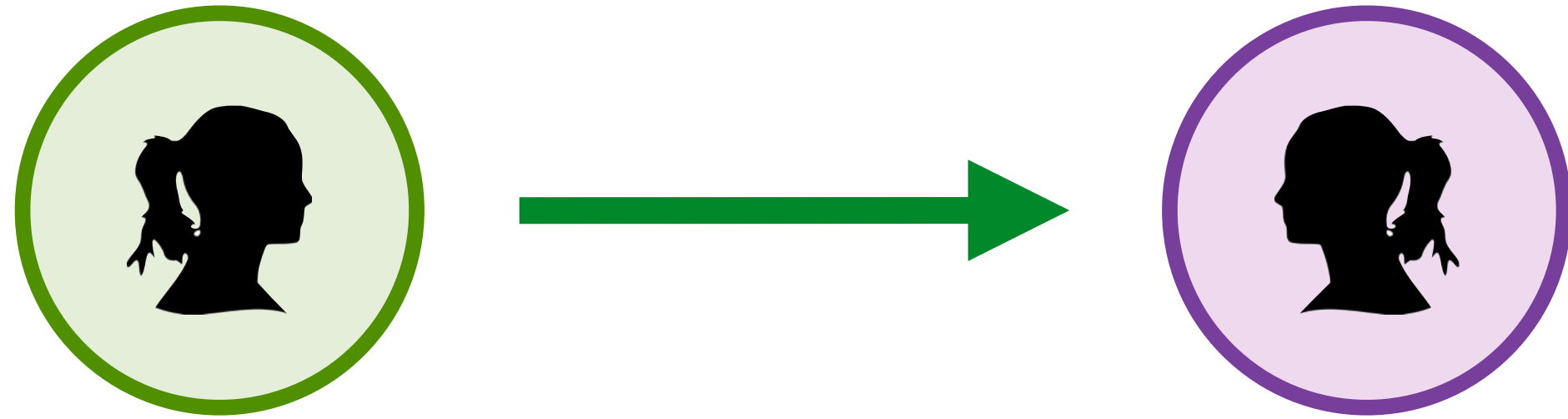
$$\max_a p(\theta | a) p(a)$$

$$\propto \max_a \sum_{\text{map}} p(\theta | \text{map}) p(a | \text{map}) p(\text{map})$$





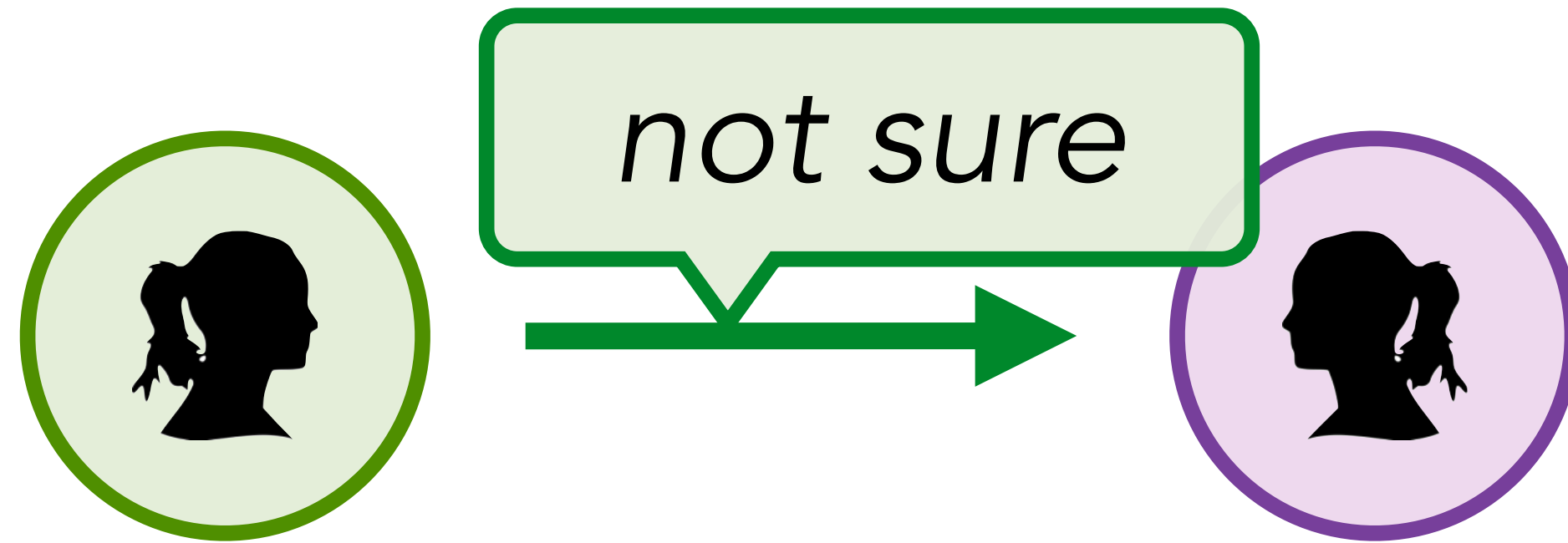
Strategy mismatch



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



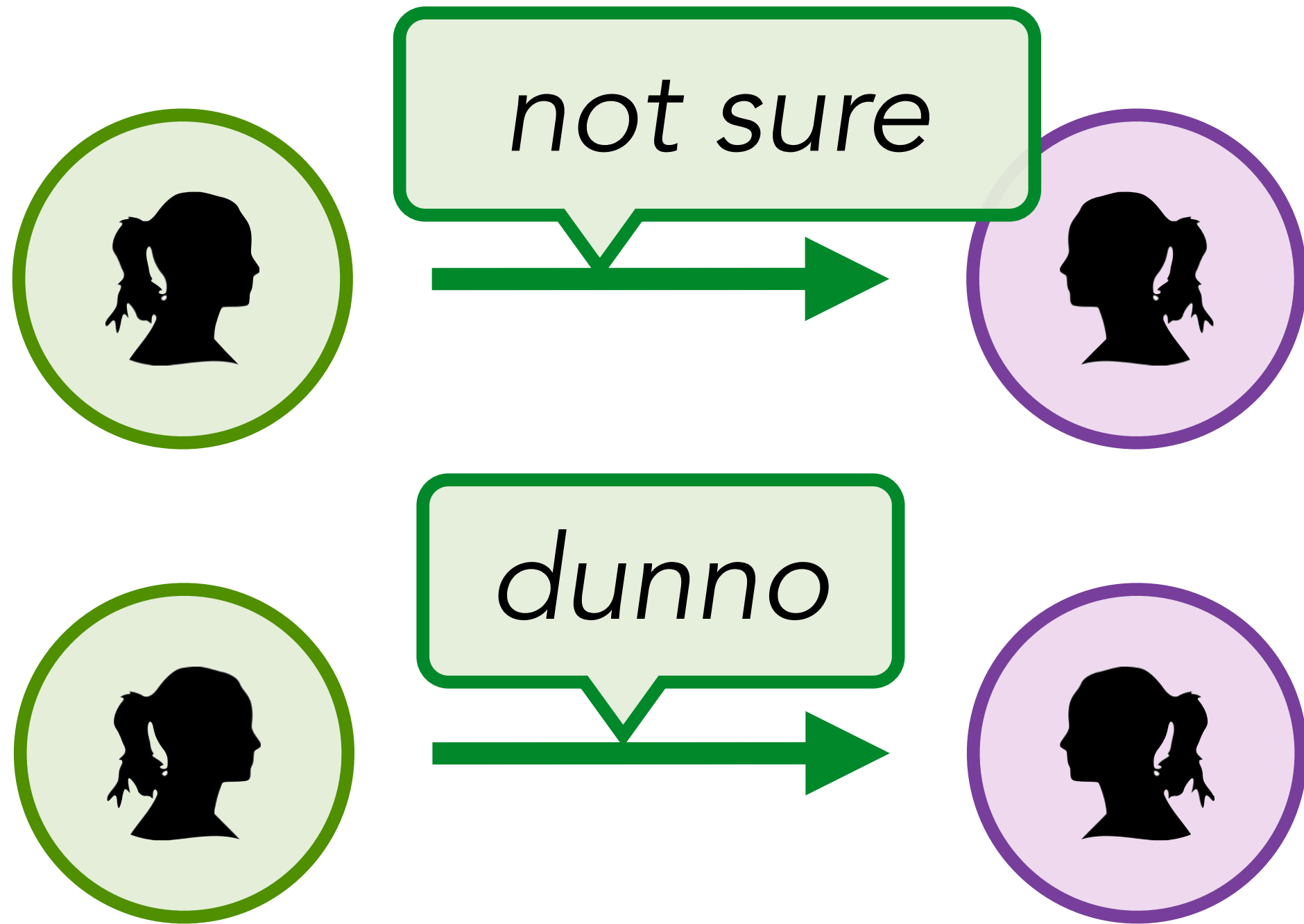
Strategy mismatch



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

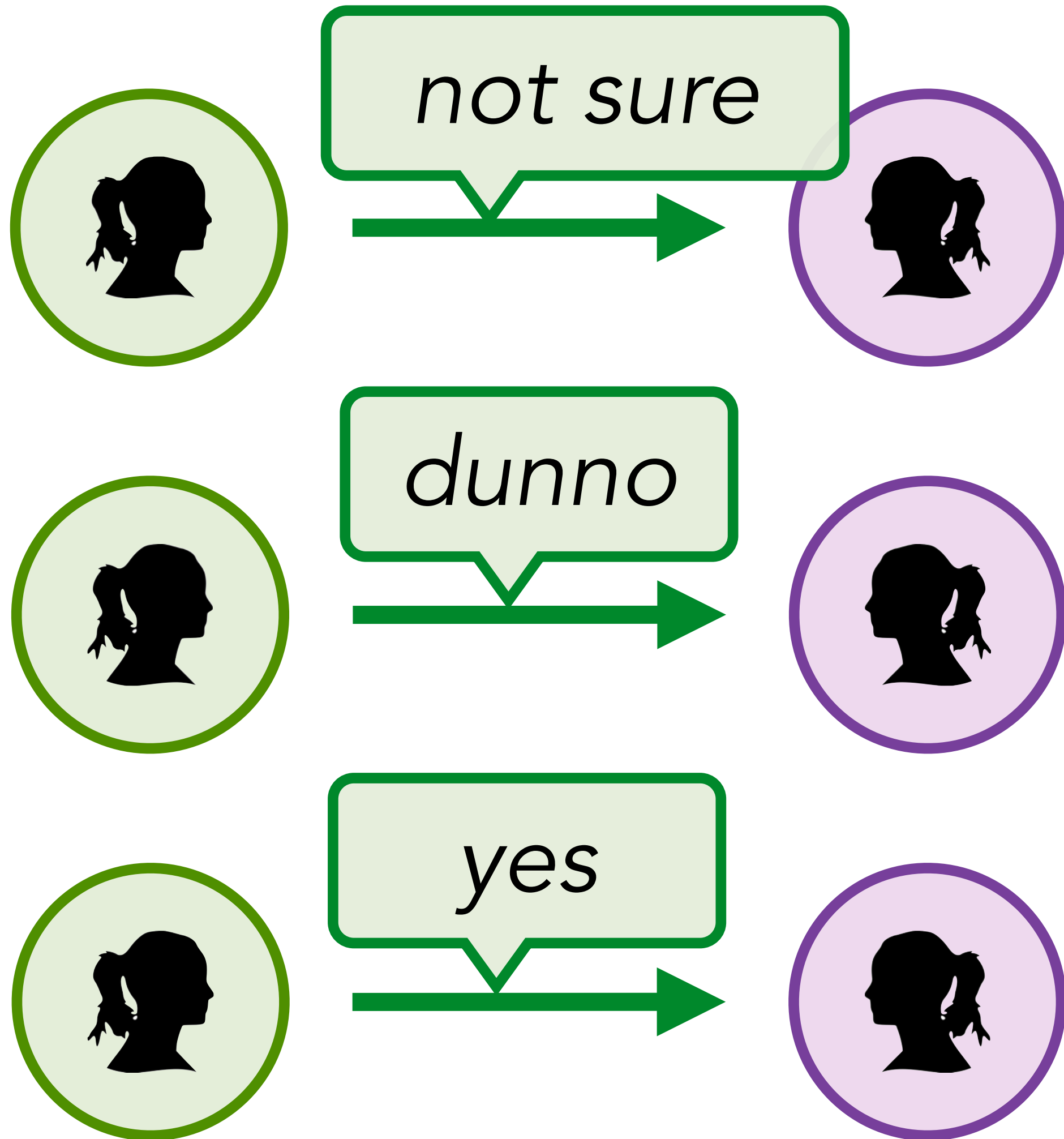


Strategy mismatch



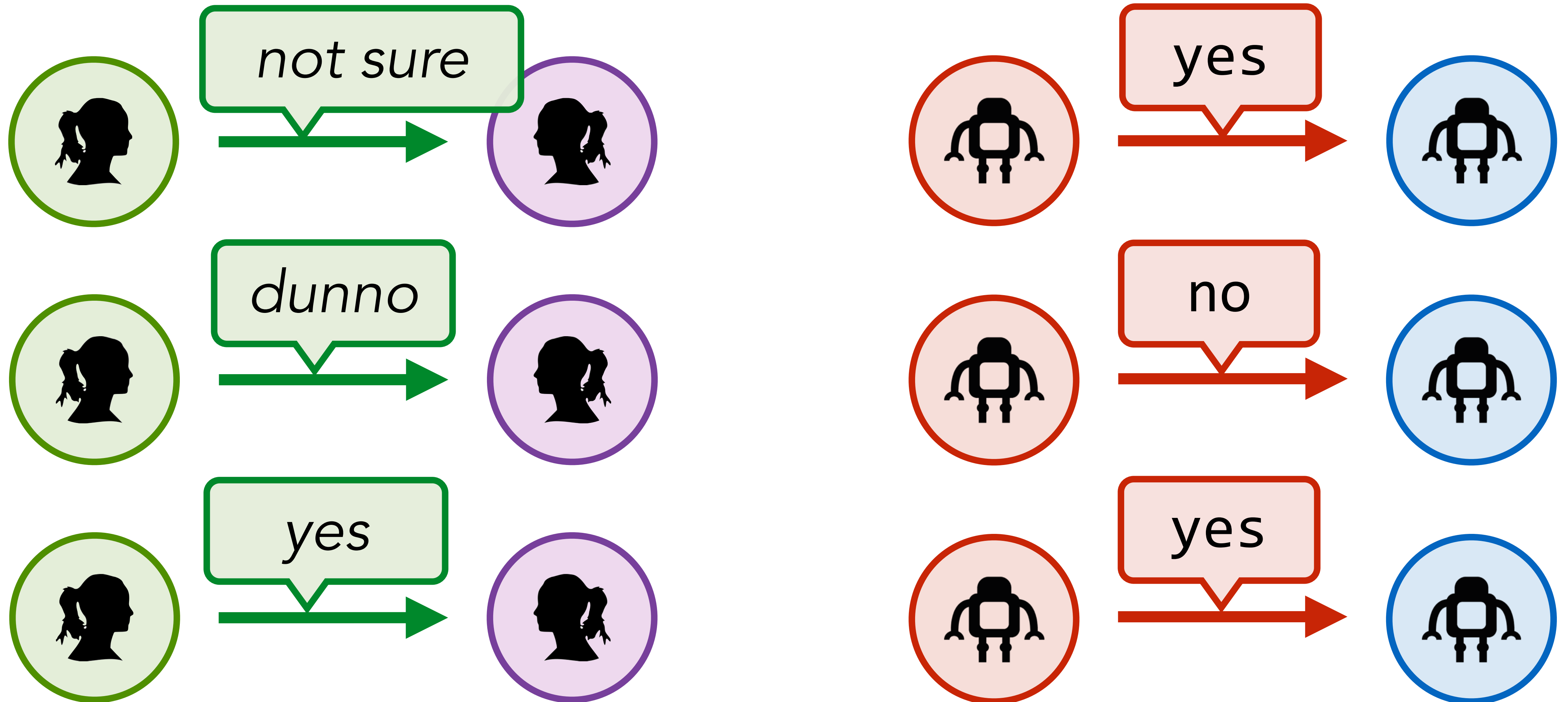


Strategy mismatch





Strategy mismatch

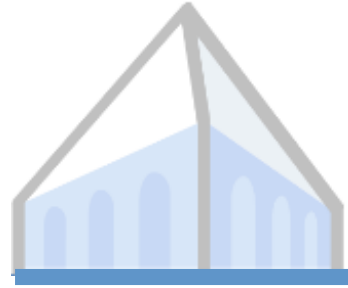




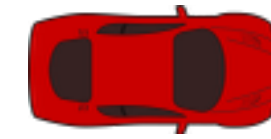
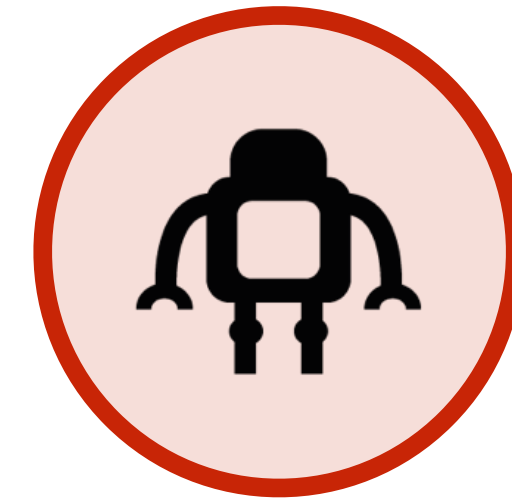
Strategy mismatch



$$\sum p(\theta, \text{map icon} \mid \text{not sure}) p(\text{not sure})$$



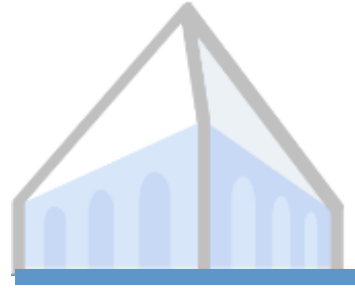
Stat MT criterion doesn't capture meaning



*In the
intersection*

moving
 $(0, 3) \rightarrow (1, 4)$





Outline

Natural language & neuralese

X **Statistical** machine translation

Semantic machine translation

Implementation details

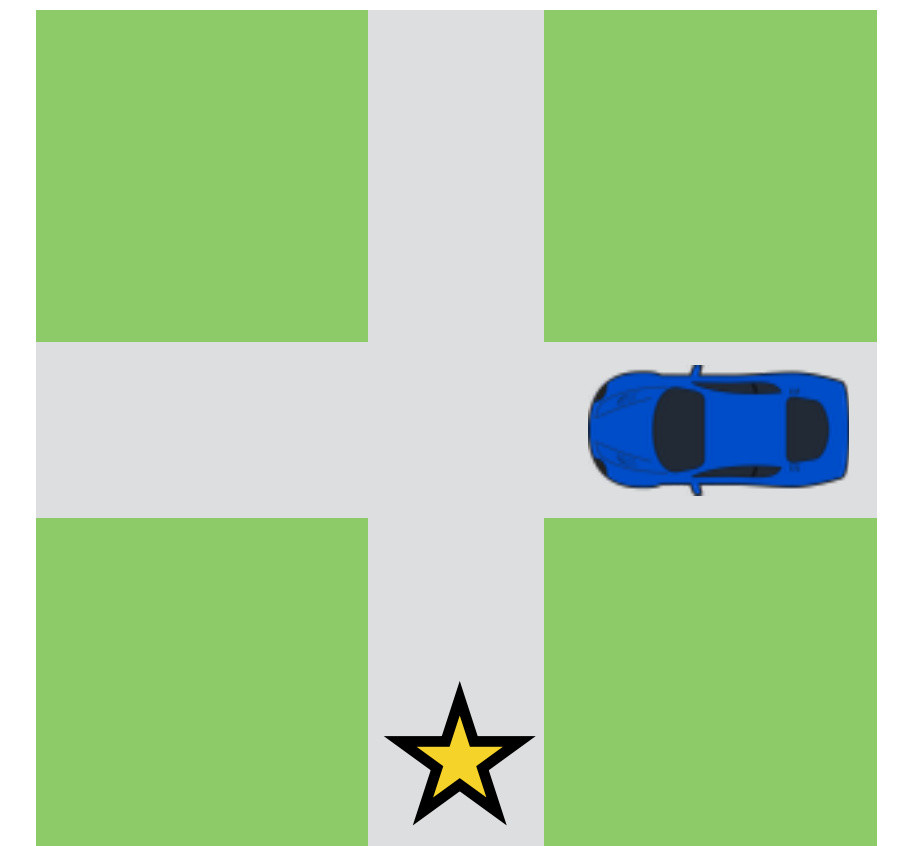
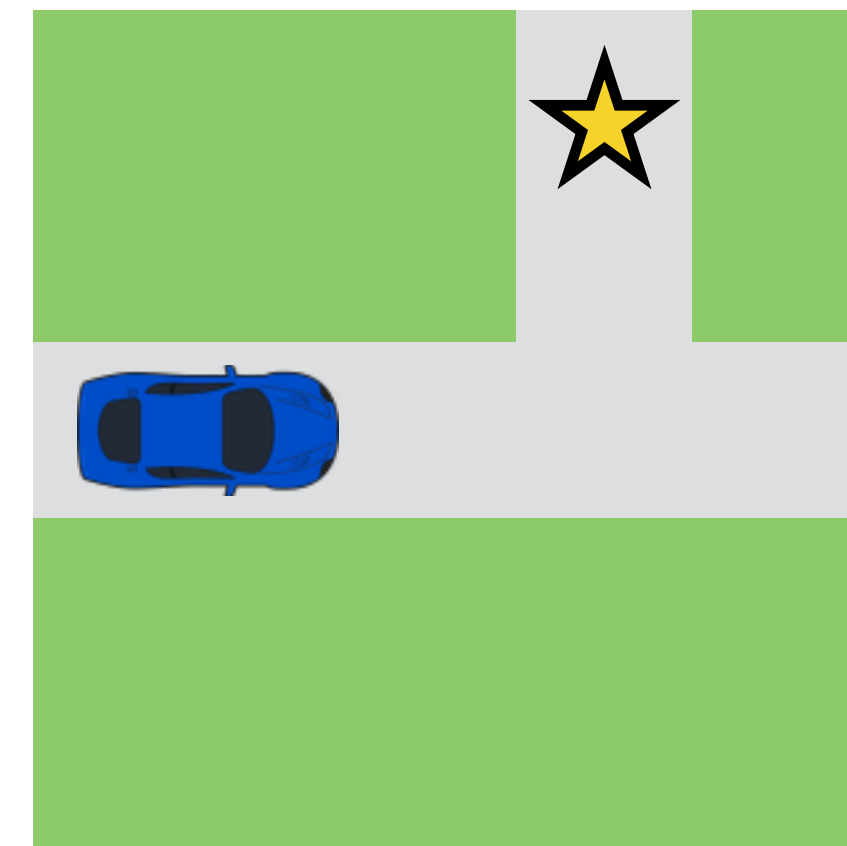
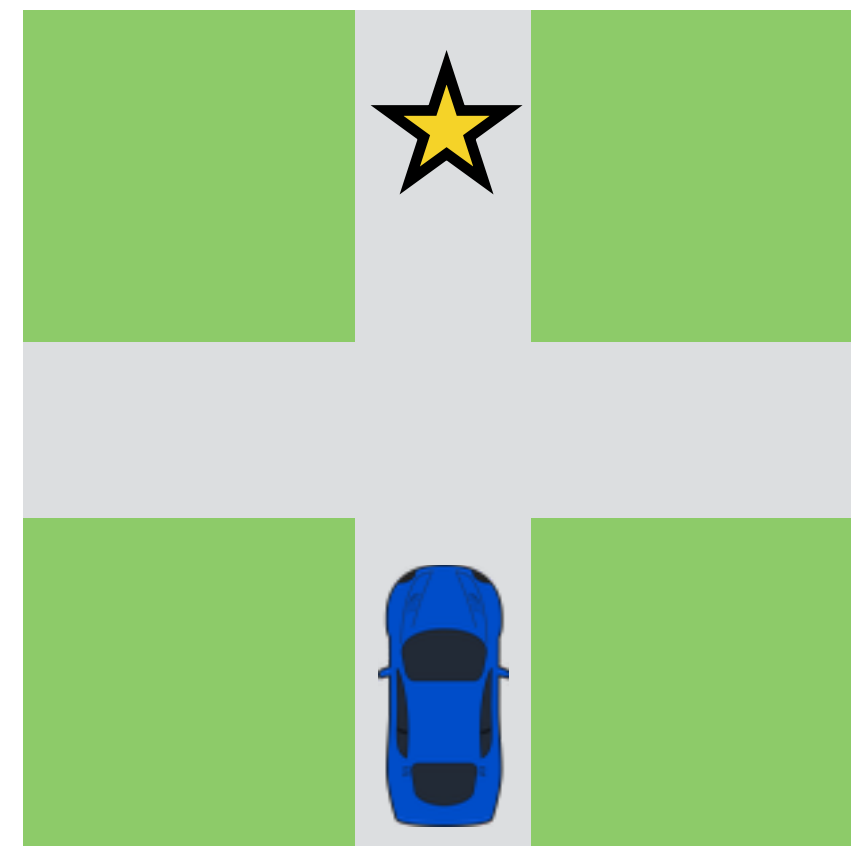
Evaluation



A "semantic MT" problem

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

I'm going north

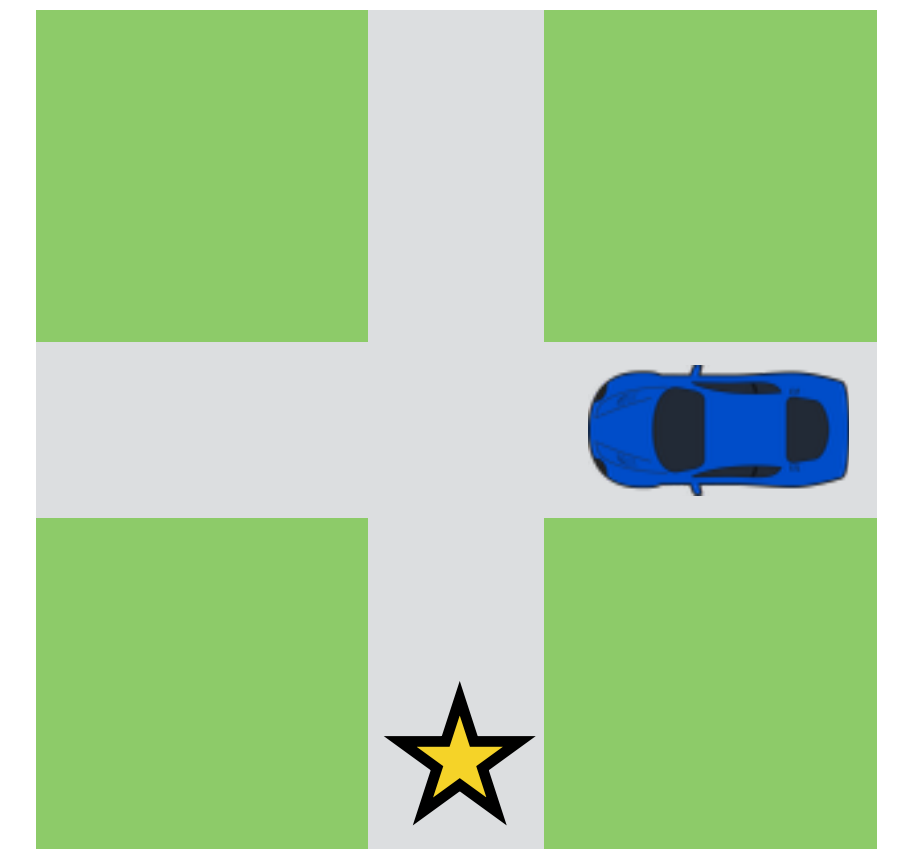
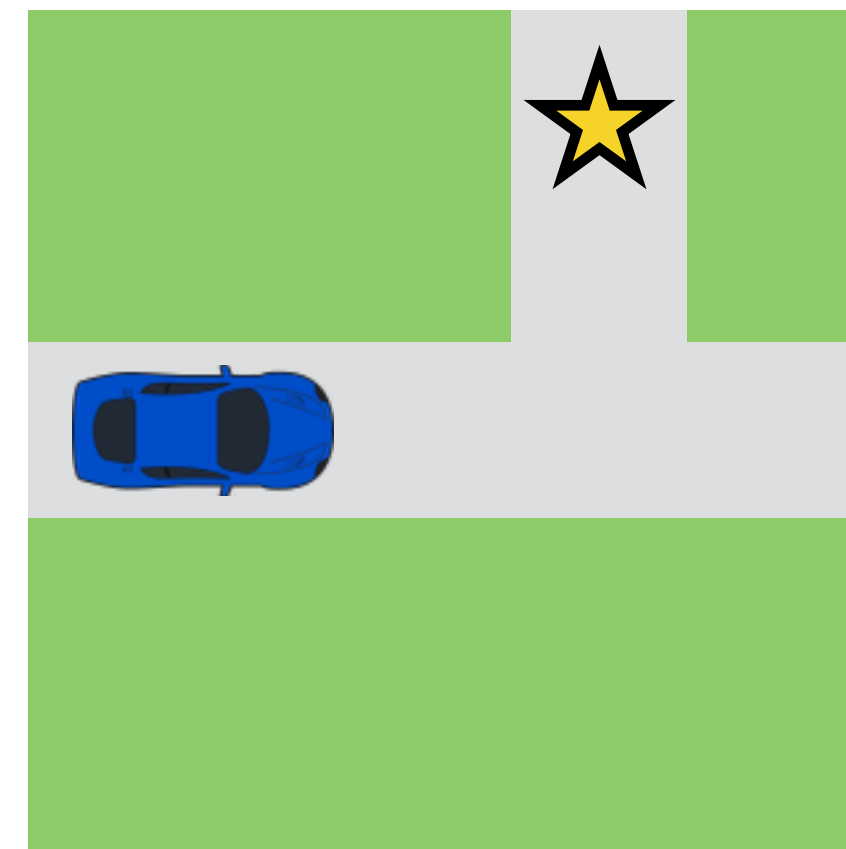
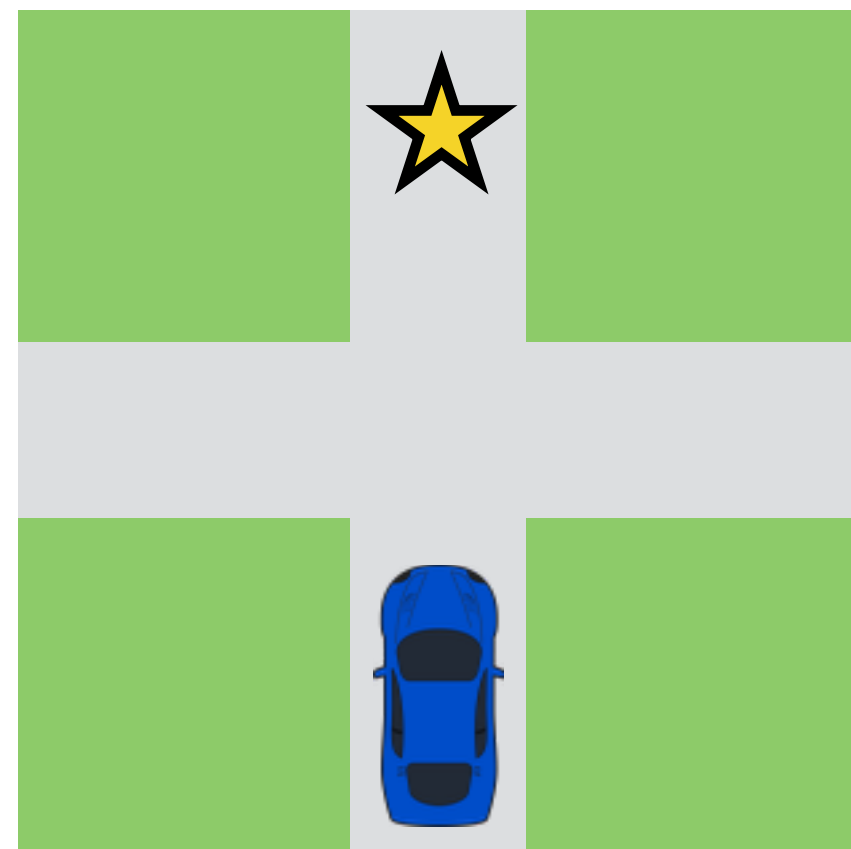




A "semantic MT" problem

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

I'm going north

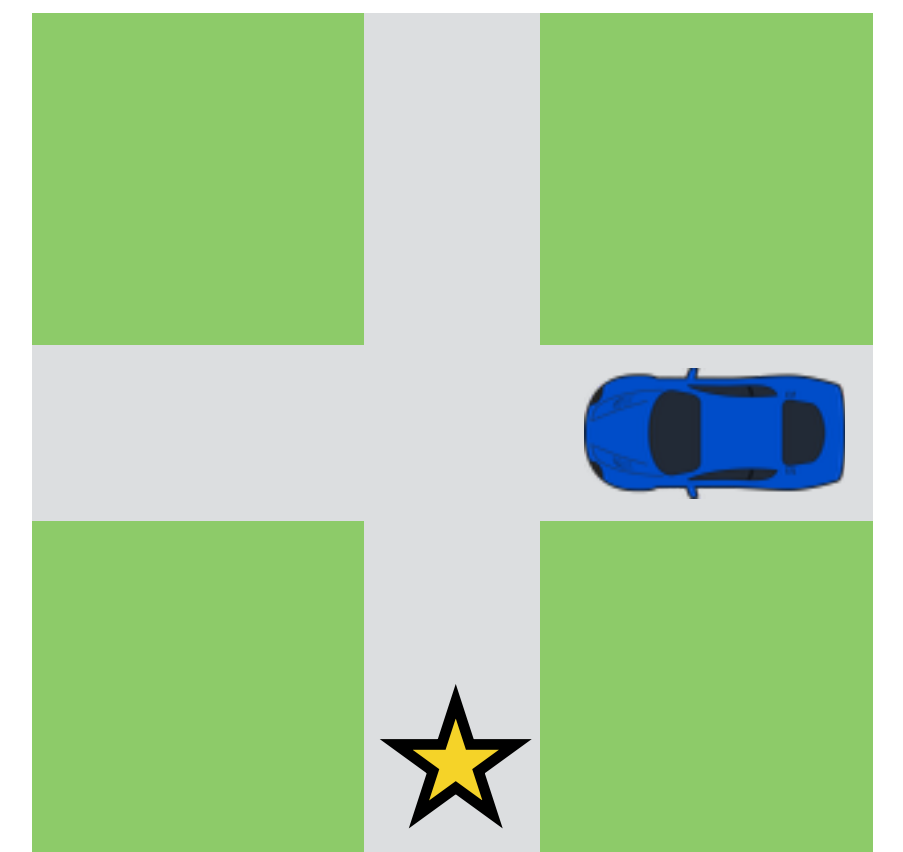
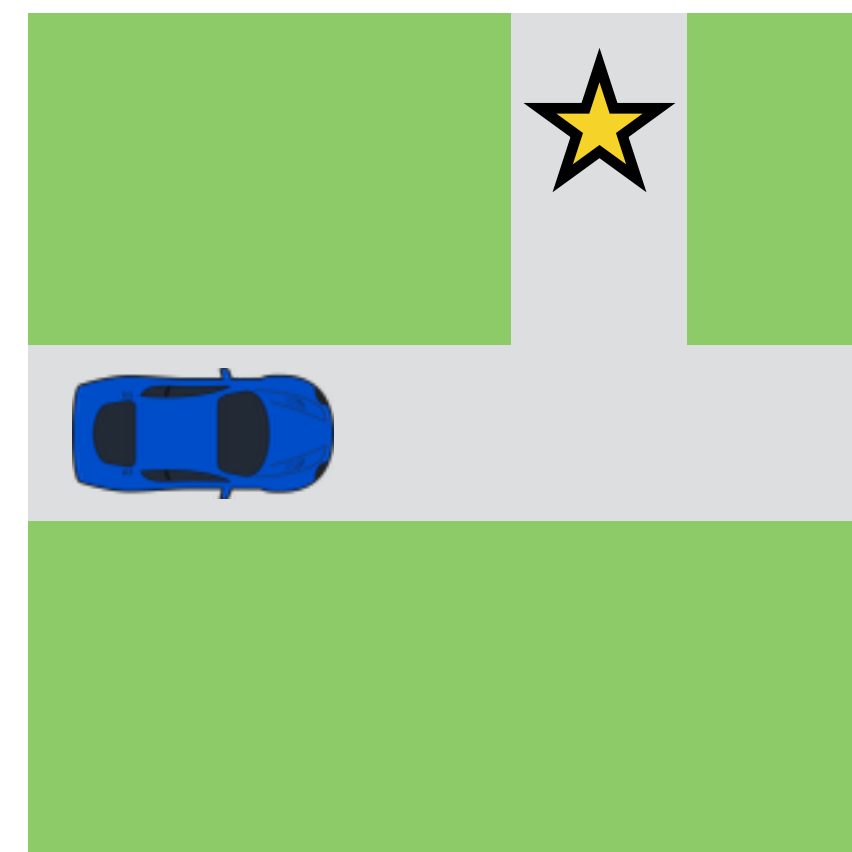
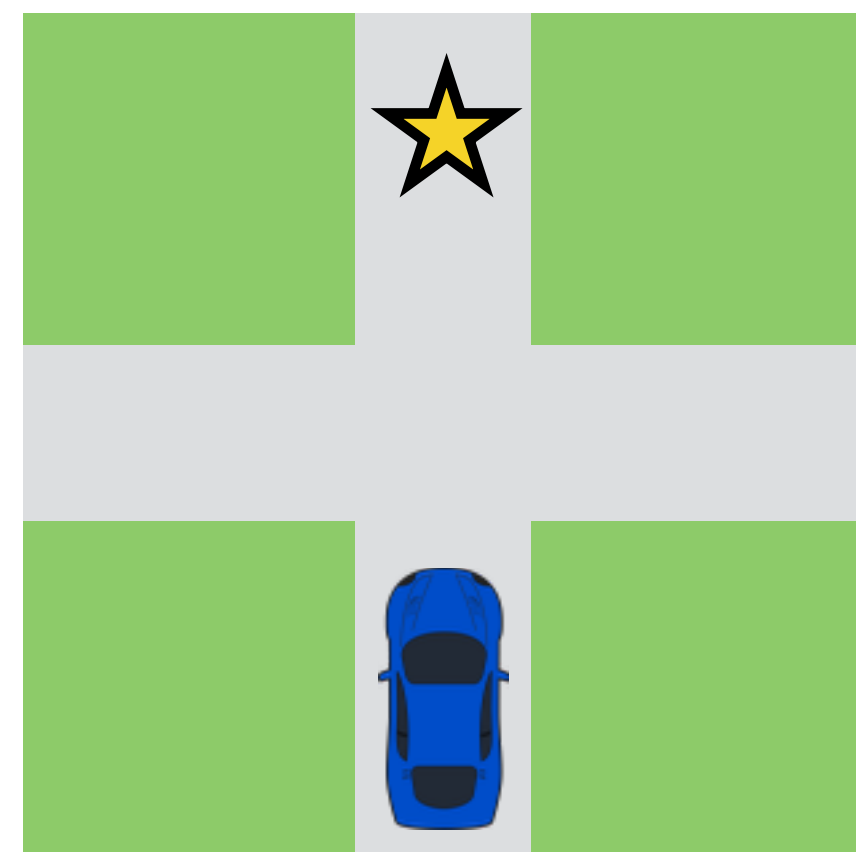




A "semantic MT" problem

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

I'm going north



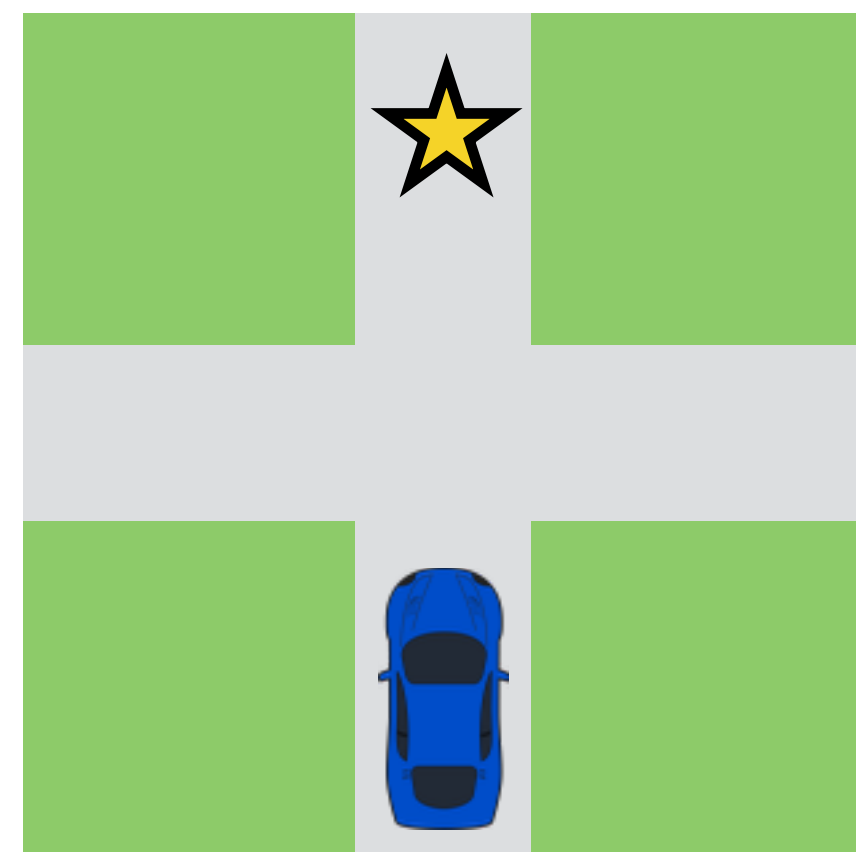
(loc (goal blue) north)



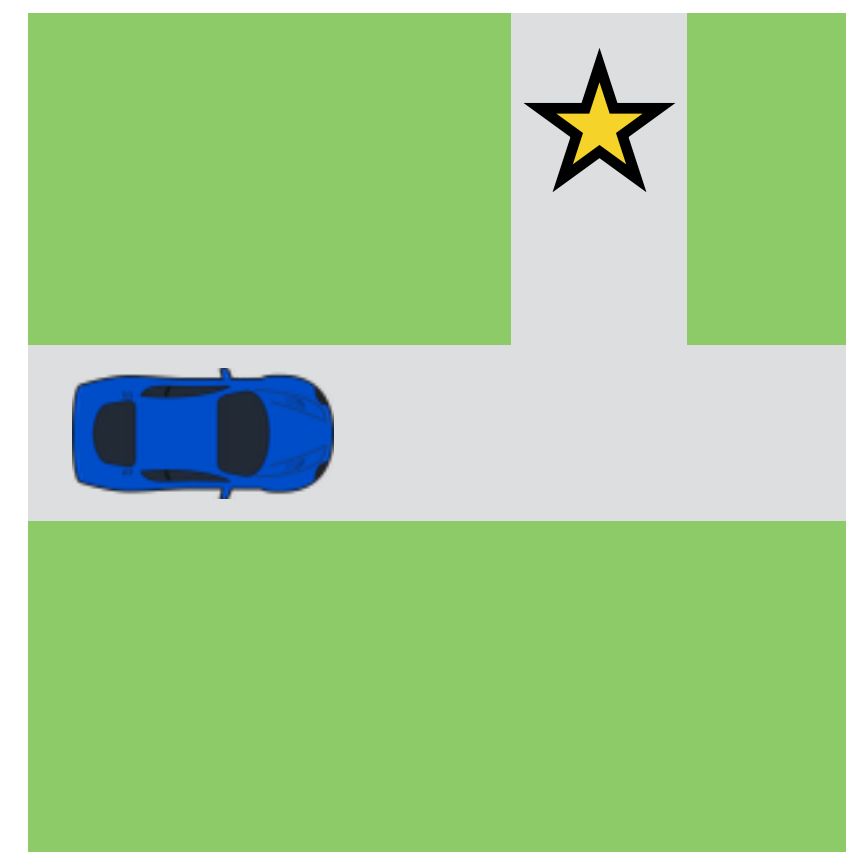
A "semantic MT" problem

The meaning of an utterance is given by its truth conditions
the **distribution over states** in which it is uttered

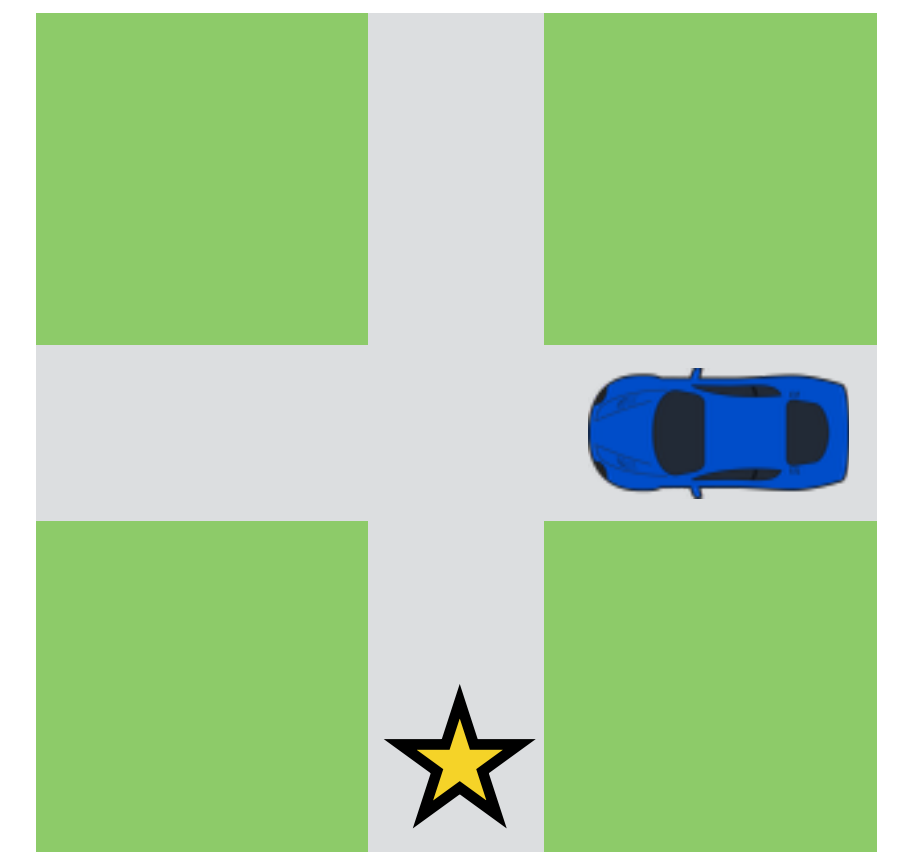
I'm going north



0.4



0.2



0.001



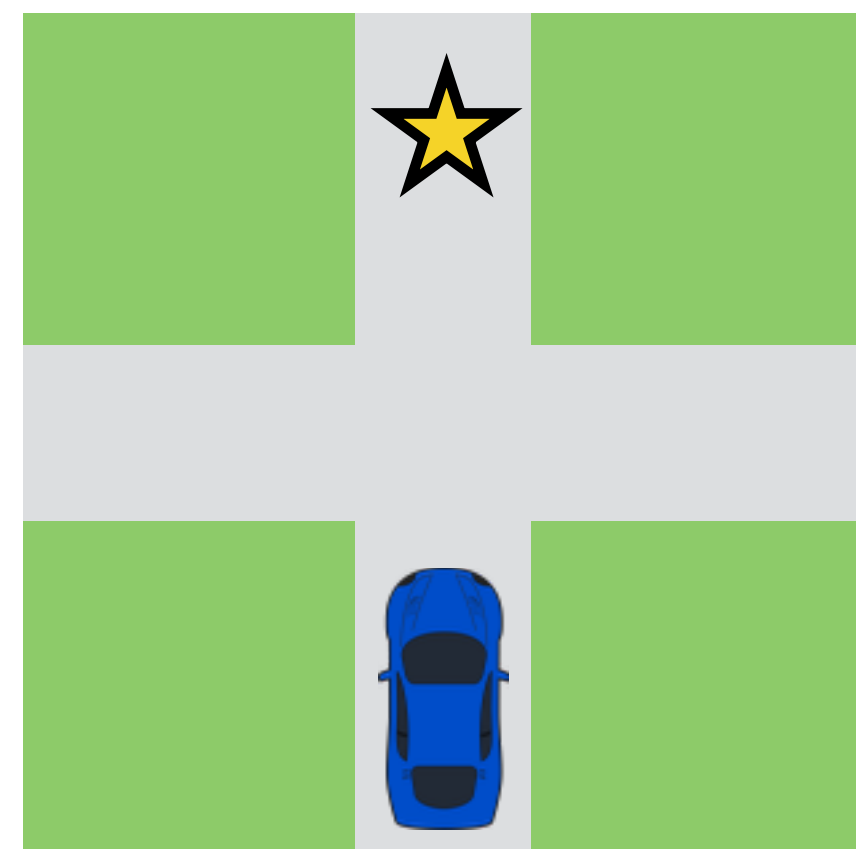
A "semantic MT" problem

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

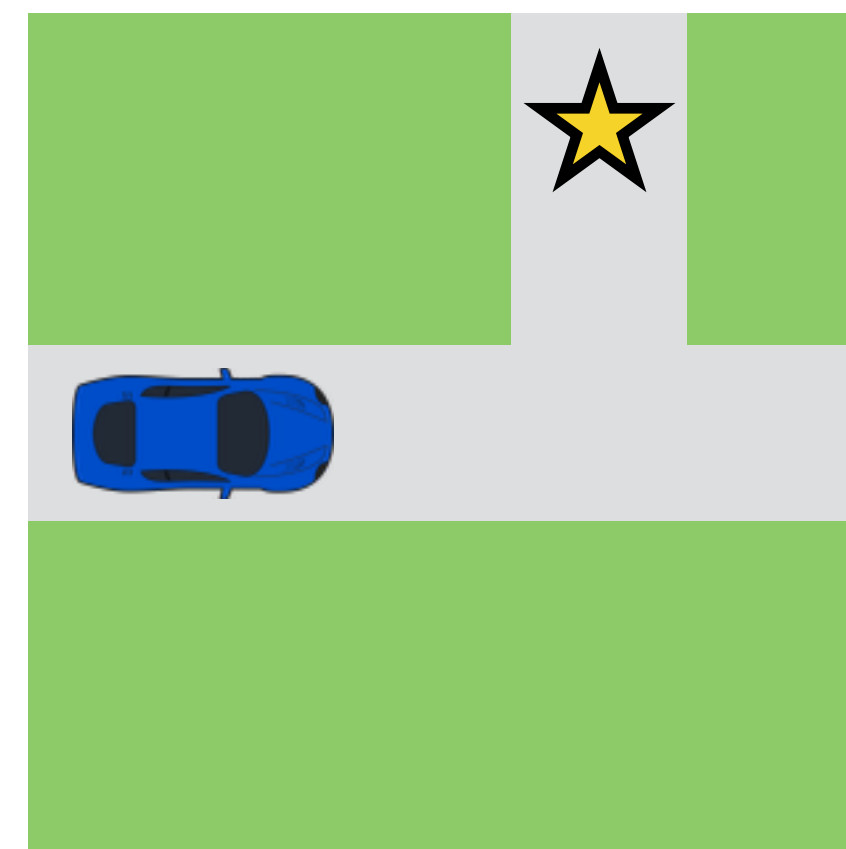
the **distribution over states** in which it is uttered

the **belief** it induces in listeners

I'm going north



0.4



0.2



0.001



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

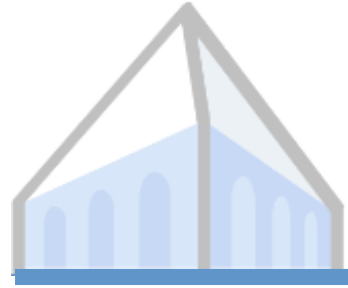


Representing meaning

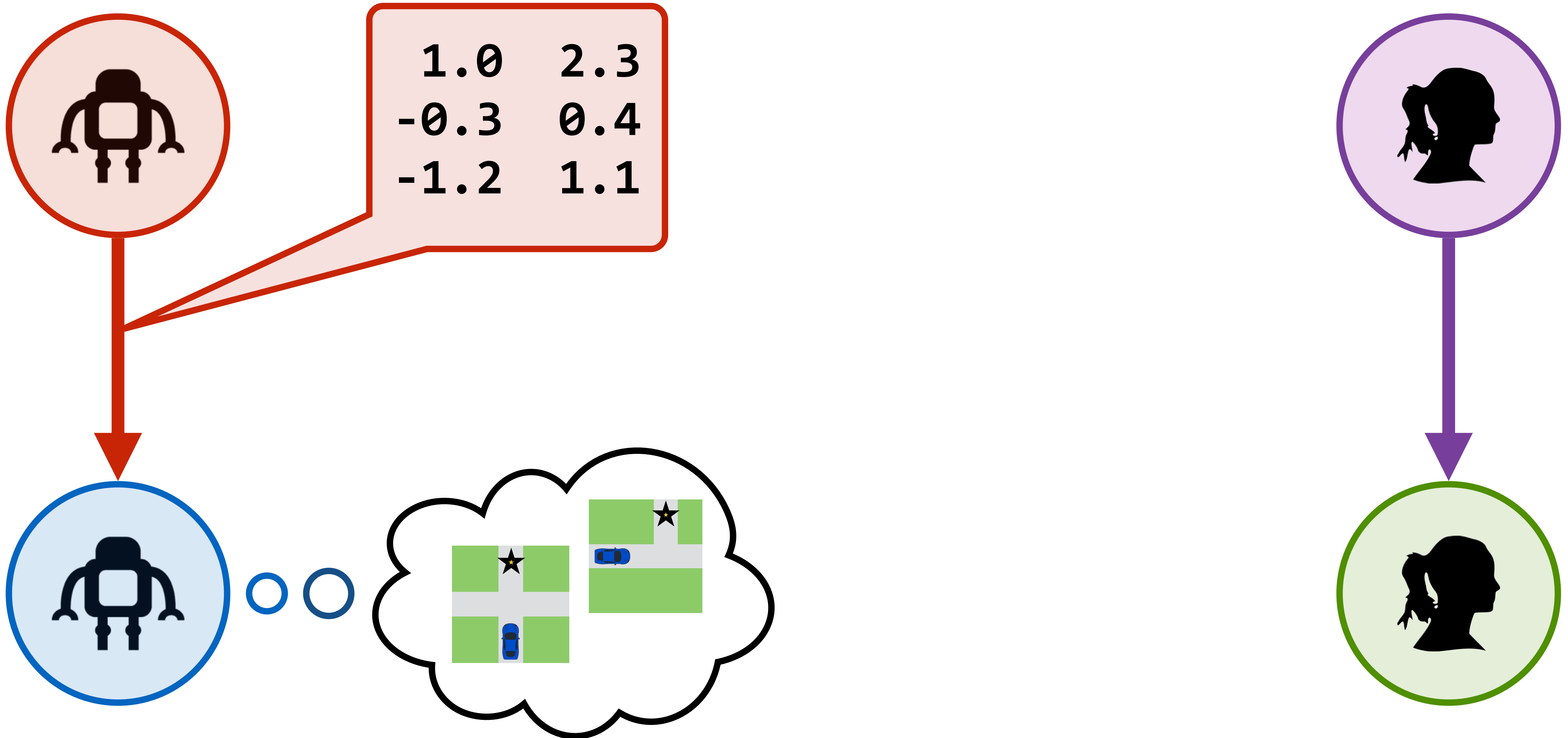
The meaning of an utterance is given by

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.

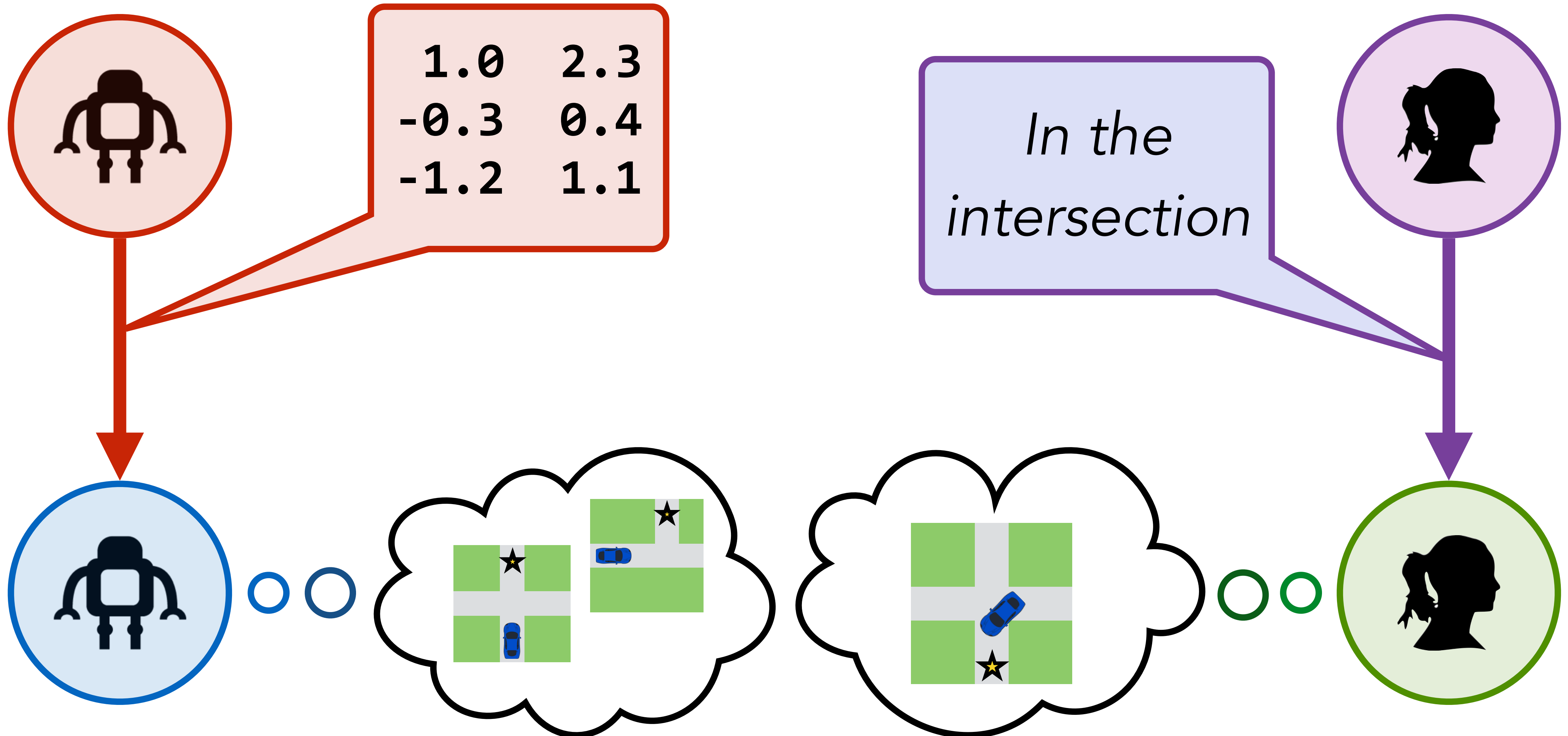


Translating with meaning



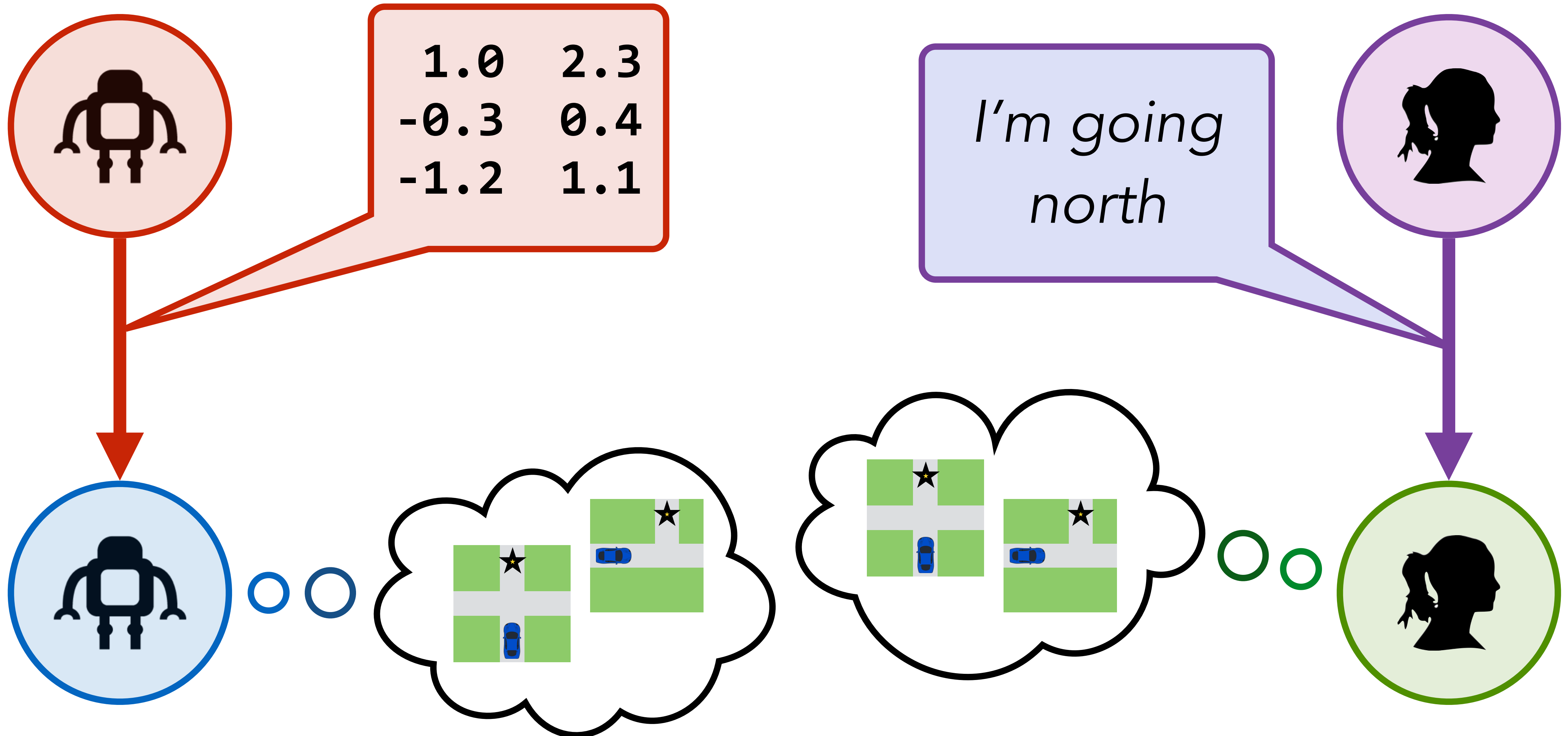


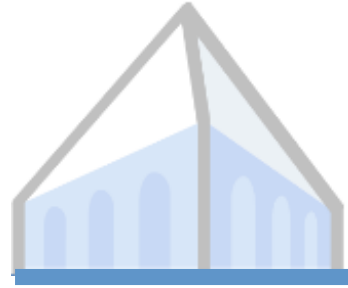
Translating with meaning



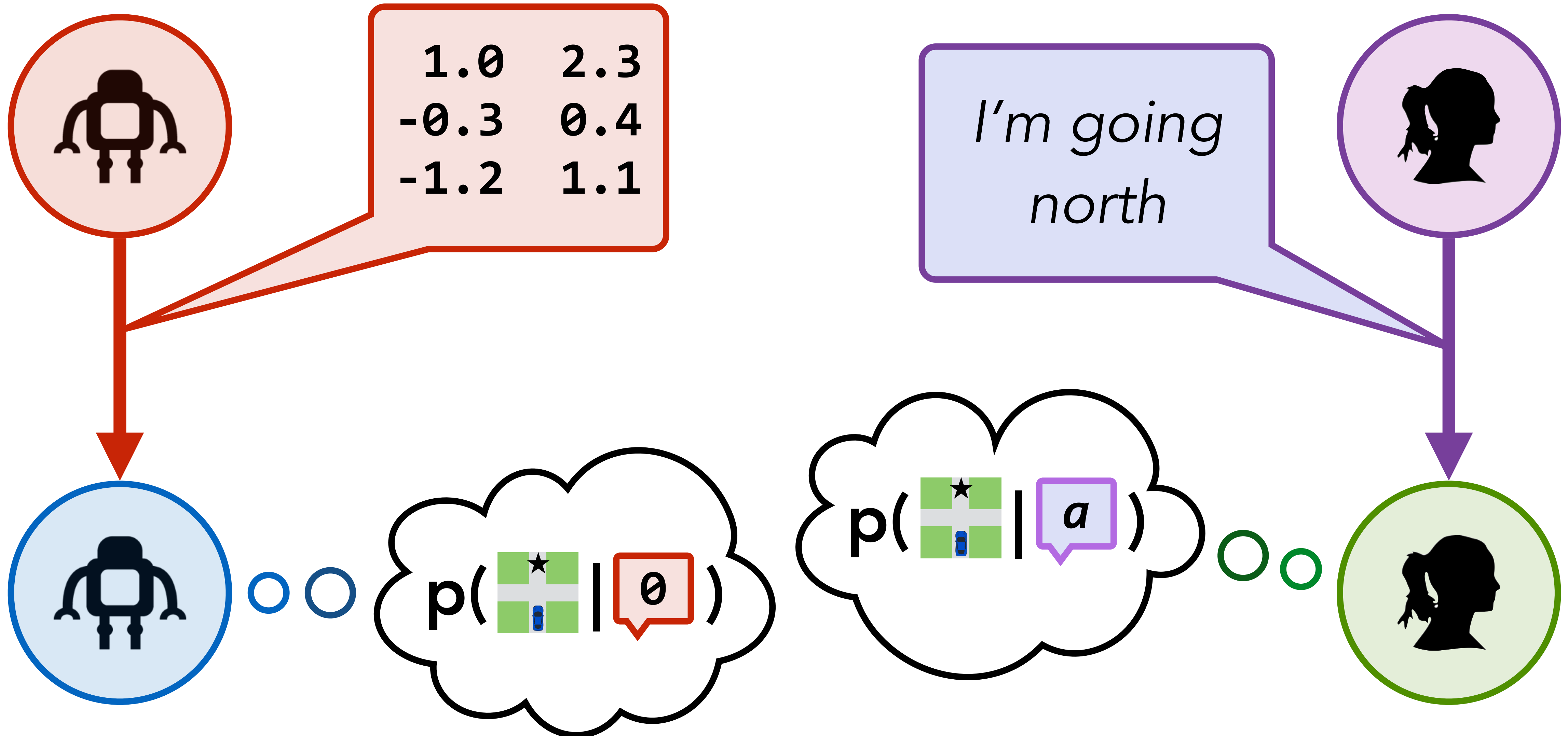


Translating with meaning



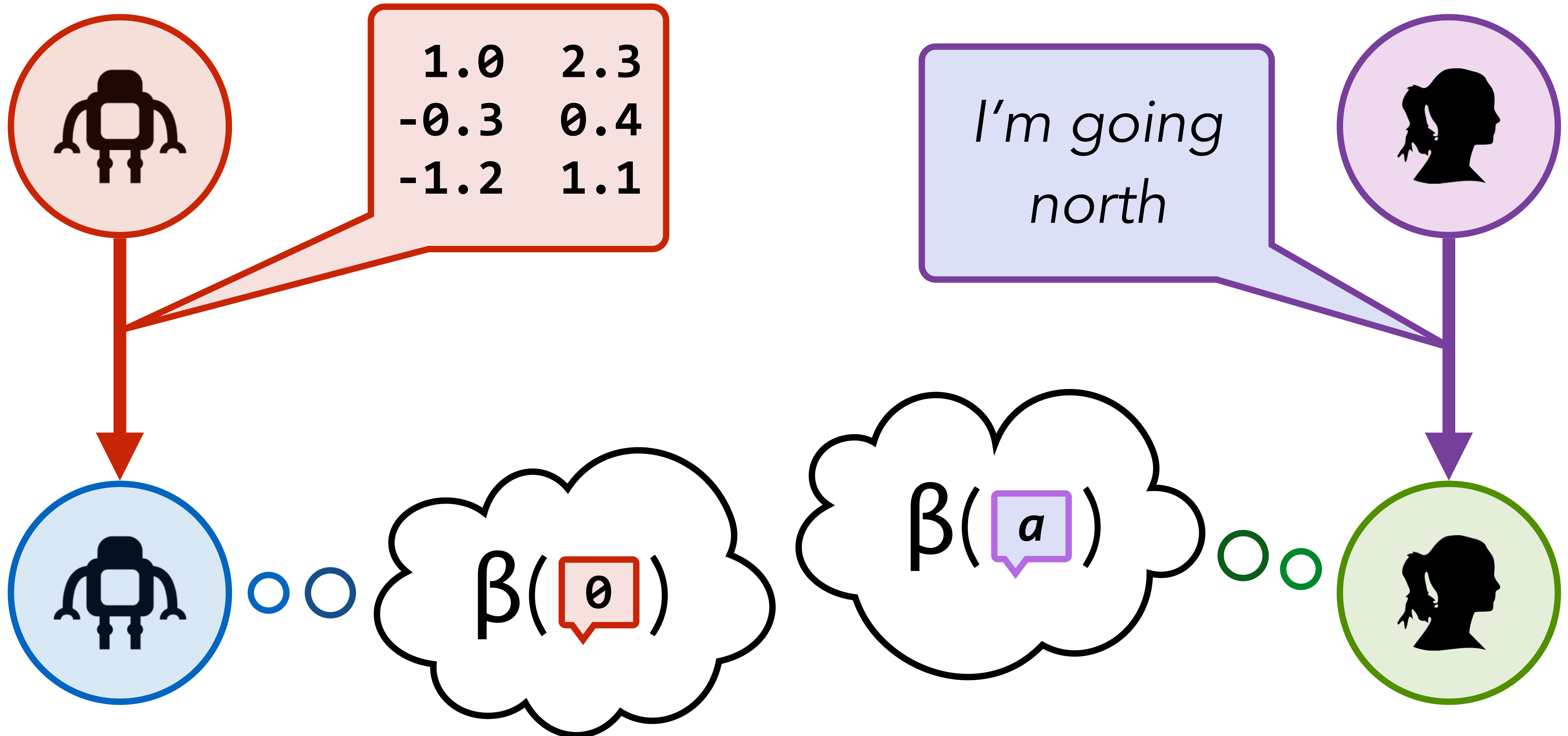


Translating with meaning



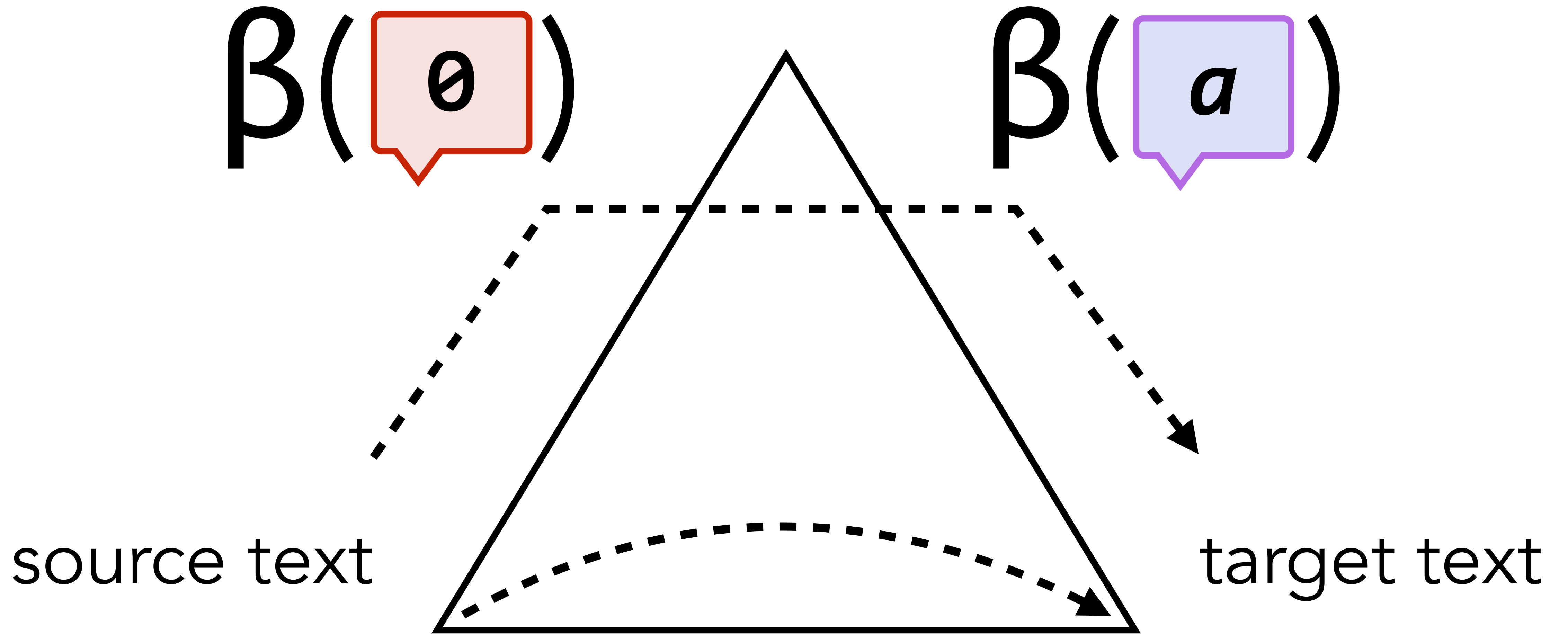


Translating with meaning





Interlingua!





Translation criterion

argmin

a

$$\text{KL}(\beta(\theta) \parallel \beta(a))$$



Translation criterion

argmin

a

$$KL(\beta(\theta) \parallel \beta(a))$$



Translation criterion

argmin

a

$$\text{KL}(\beta(\theta) \parallel \beta(a))$$



Translation criterion

argmin

a

$$\text{KL}(\beta(\theta) \parallel \beta(a))$$



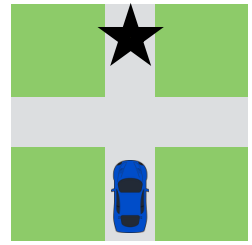
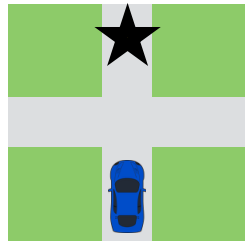
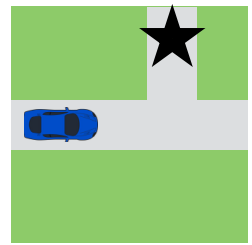
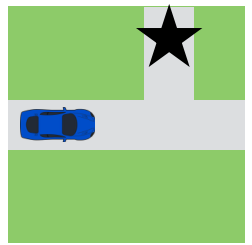
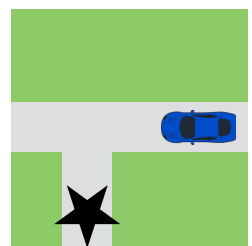
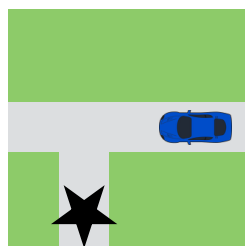
Computing representations

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$



Computing representations: sparsity

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$

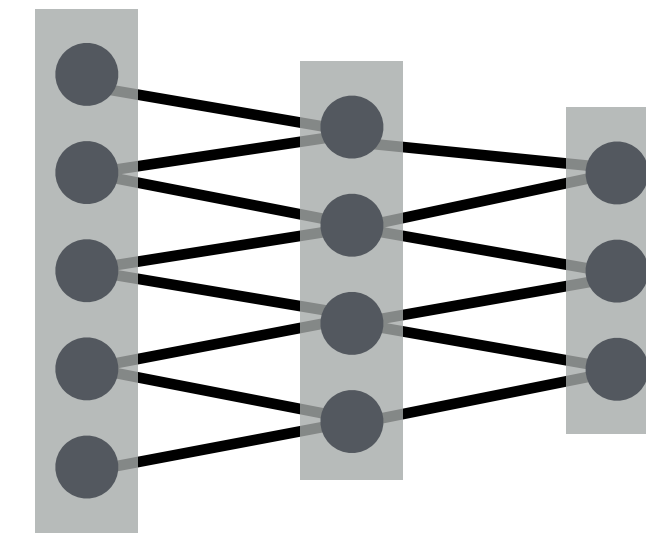
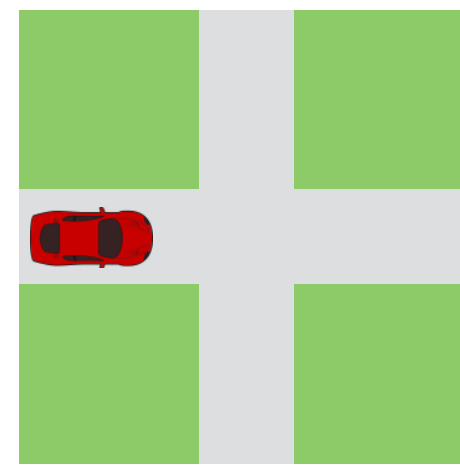
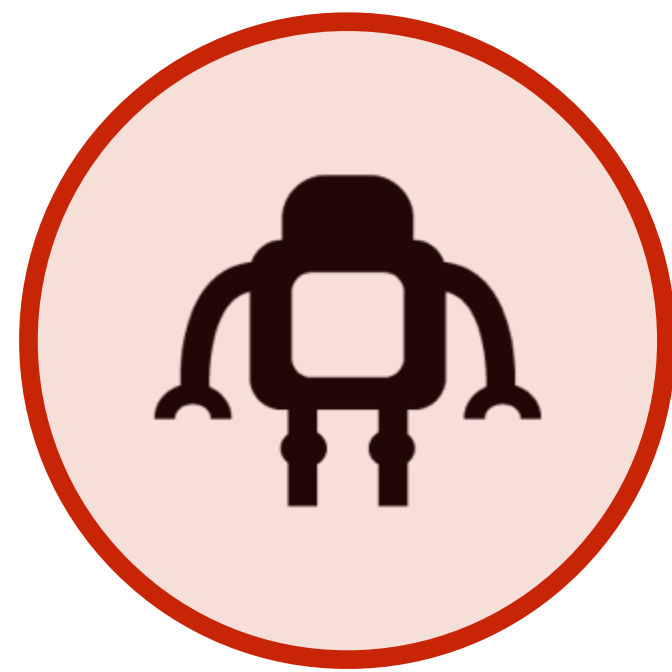
$p(\text{star, car} \mid \theta)$	$p(\text{star, car} \mid a)$
 	
	
	



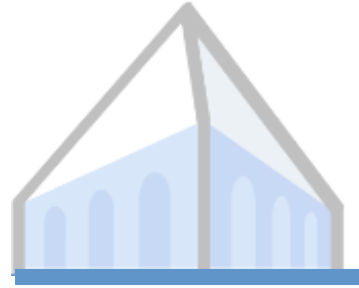
Computing representations: smoothing

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$

agent
policy



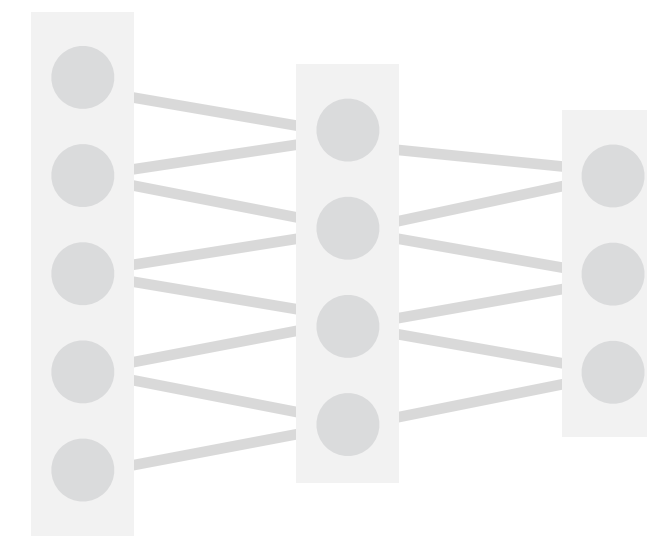
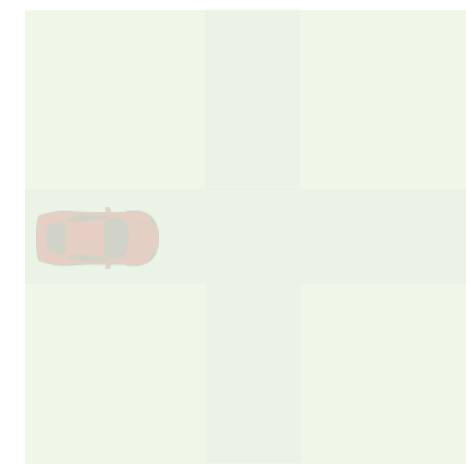
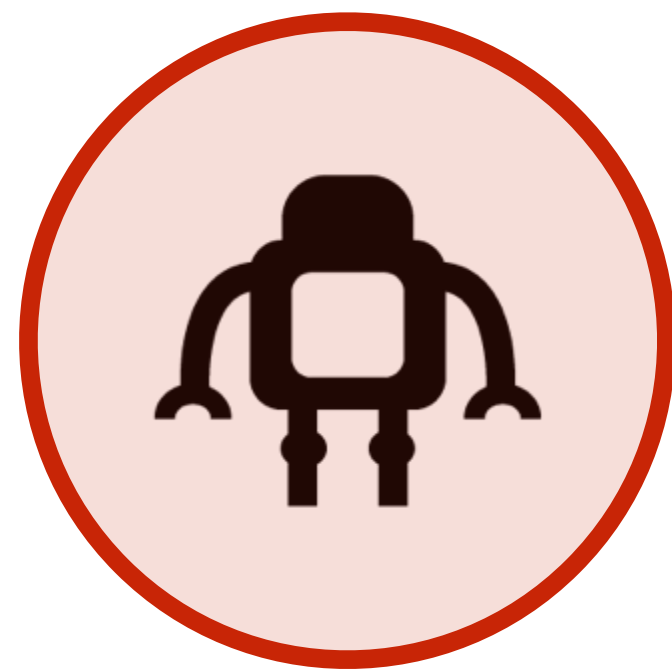
actions &
messages



Computing representations: smoothing

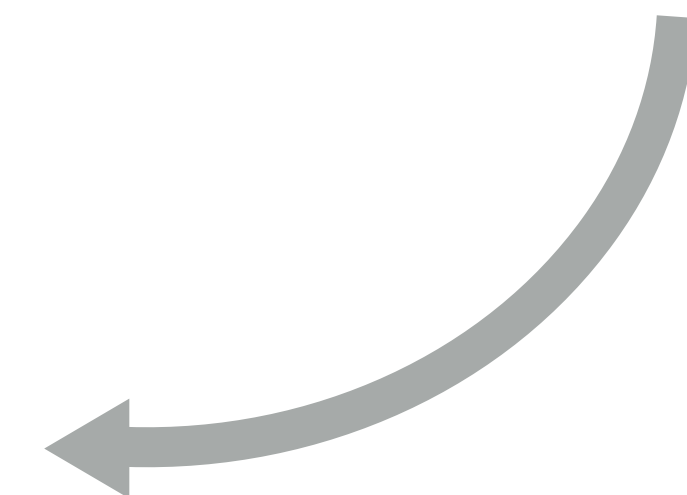
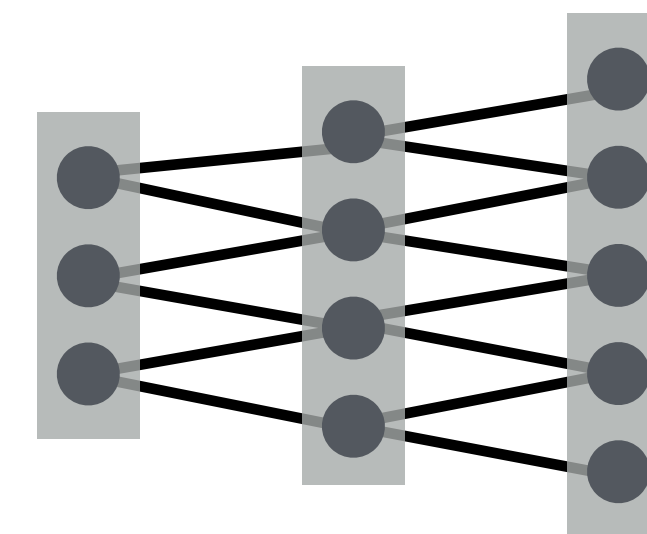
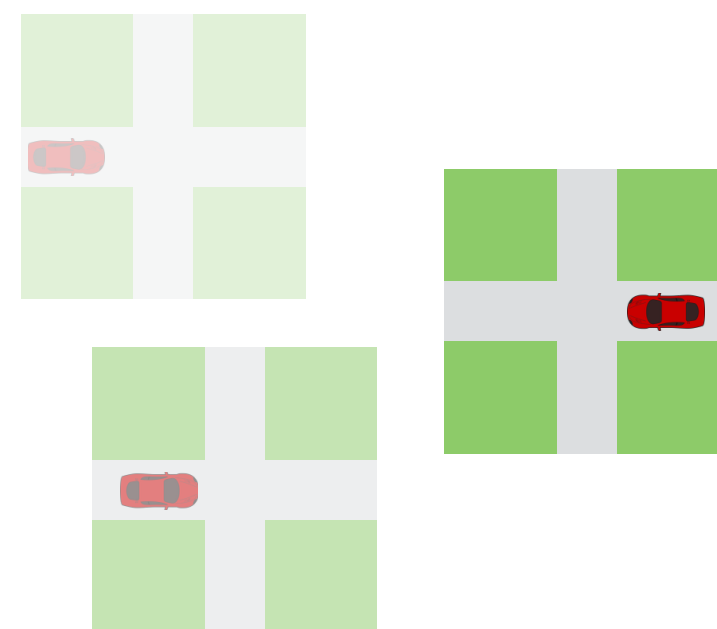
argmin _{a} $KL(\beta(\theta) \parallel \beta(a))$

agent
policy



actions &
messages

agent
model

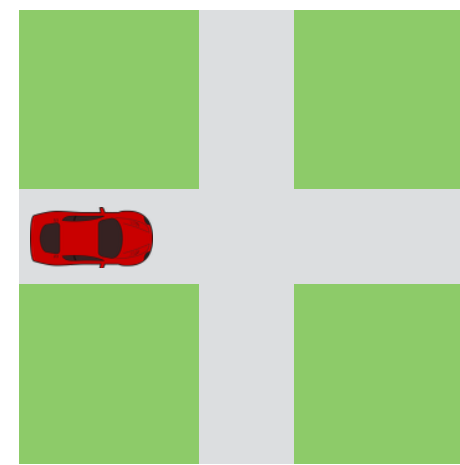




Computing representations: smoothing

argmin _{a} $KL(\beta(\theta) \parallel \beta(a))$

human



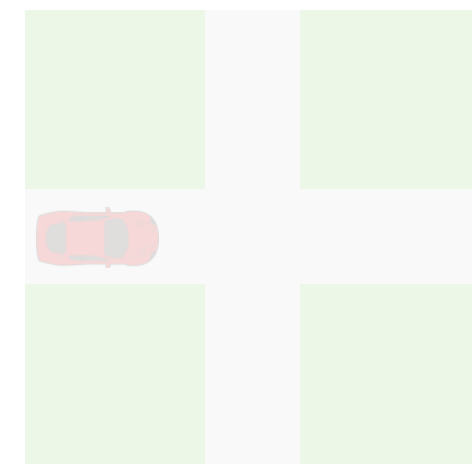
actions & messages



Computing representations: smoothing

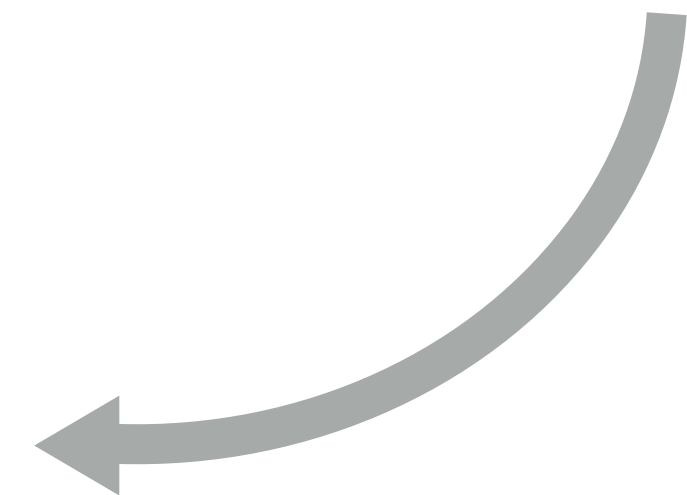
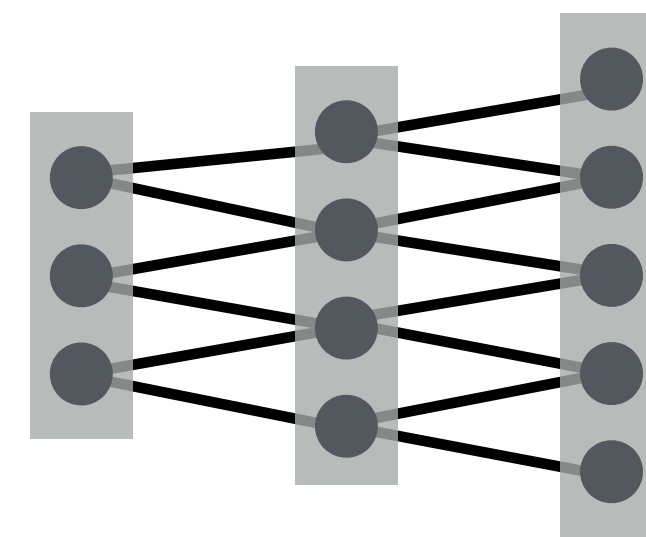
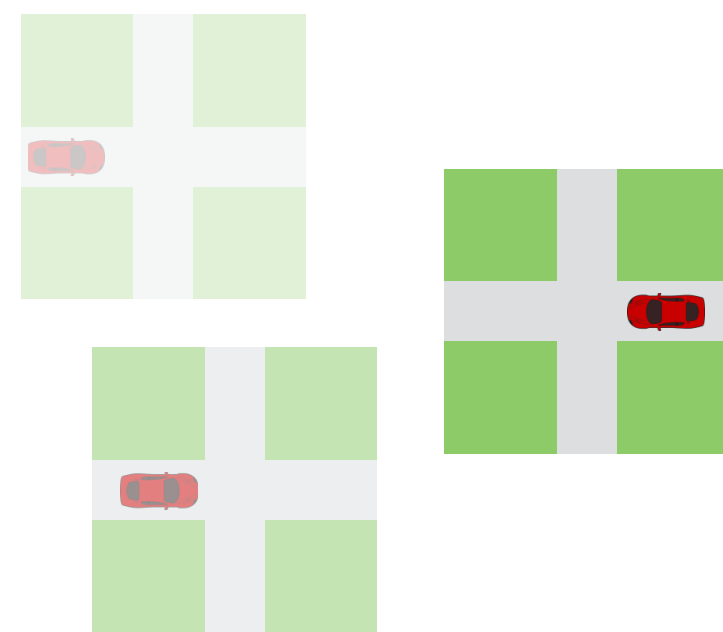
argmin _{a} $KL(\beta(\theta) \parallel \beta(a))$

human
policy



actions &
messages

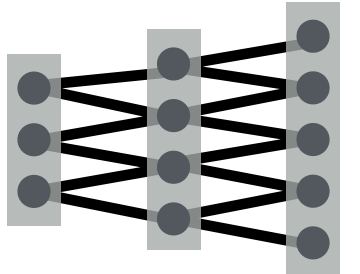
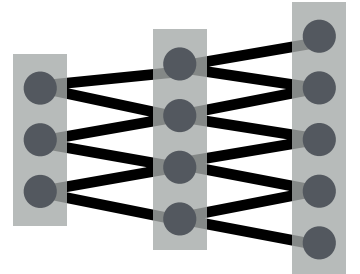
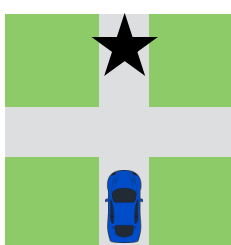
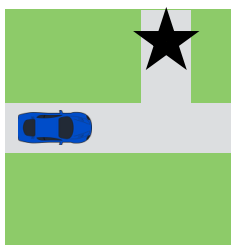
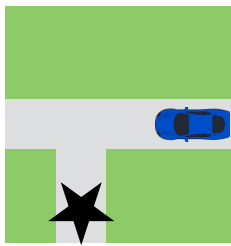
human
model





Computing representations: smoothing

$$\operatorname{argmin}_a \operatorname{KL}(\beta(\theta) \parallel \beta(a))$$

	 θ	 a
	0.10	0.08
	0.05	0.01
	0.13	0.22



Computing KL

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$



Computing KL

$$\operatorname{argmin}_a \operatorname{KL}(\beta(\theta) \parallel \beta(a))$$

$$\operatorname{KL}(p \parallel q) = \mathbf{E}_{\star}^p \frac{p(\star)}{q(\star)}$$



Computing KL: sampling

$$\operatorname{argmin}_a \operatorname{KL}(\beta(\theta) \parallel \beta(a))$$

$$\operatorname{KL}(p \parallel q) = \sum_{i} p(\text{star on } i) \log \frac{p(\text{star on } i)}{q(\text{star on } i)}$$



Finding translations

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$



Finding translations: brute force

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$

going north → 0.5

crossing the intersection → 2.3

I'm done → 0.2

after you → 9.7



Finding translations: brute force

$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$

going north → 0.5

crossing the intersection → 2.3

I'm done → **0.2**

after you → 9.7

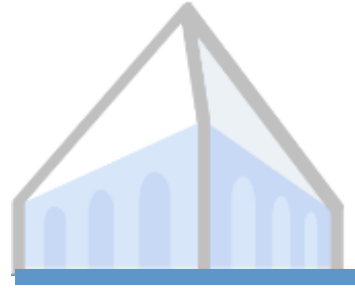


Finding translations

argmin

a

$$\text{KL}(\beta(\theta) \parallel \beta(a))$$



Outline

Natural language & neuralese

Statistical machine translation

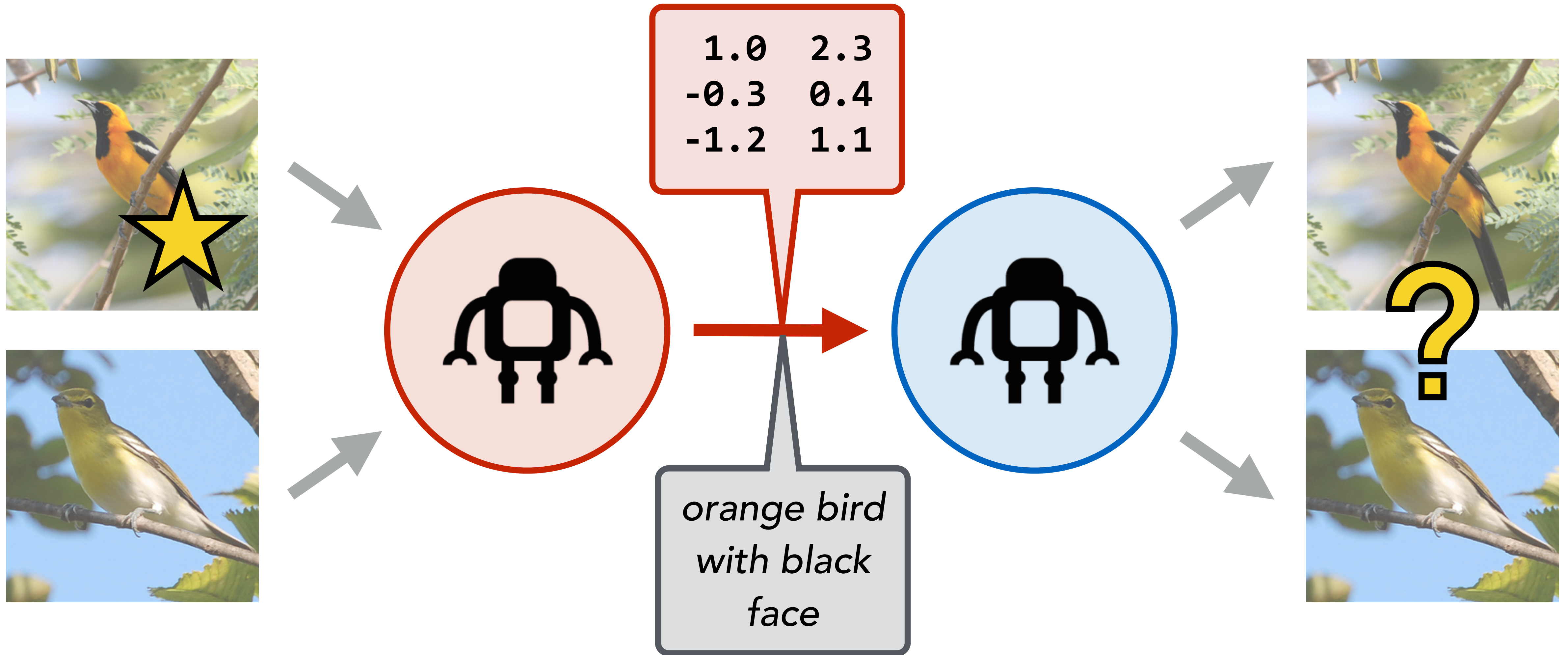
Semantic machine translation

Implementation details

Evaluation

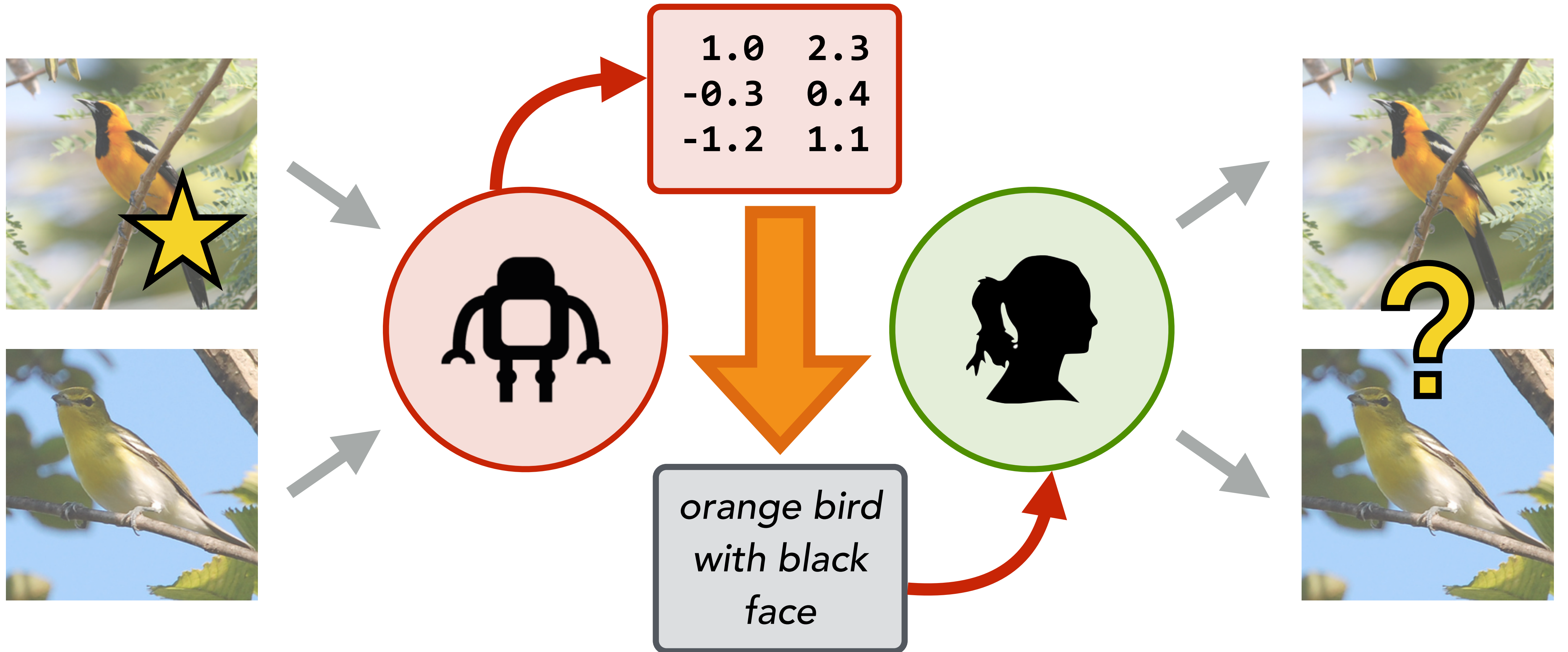


Referring expression games



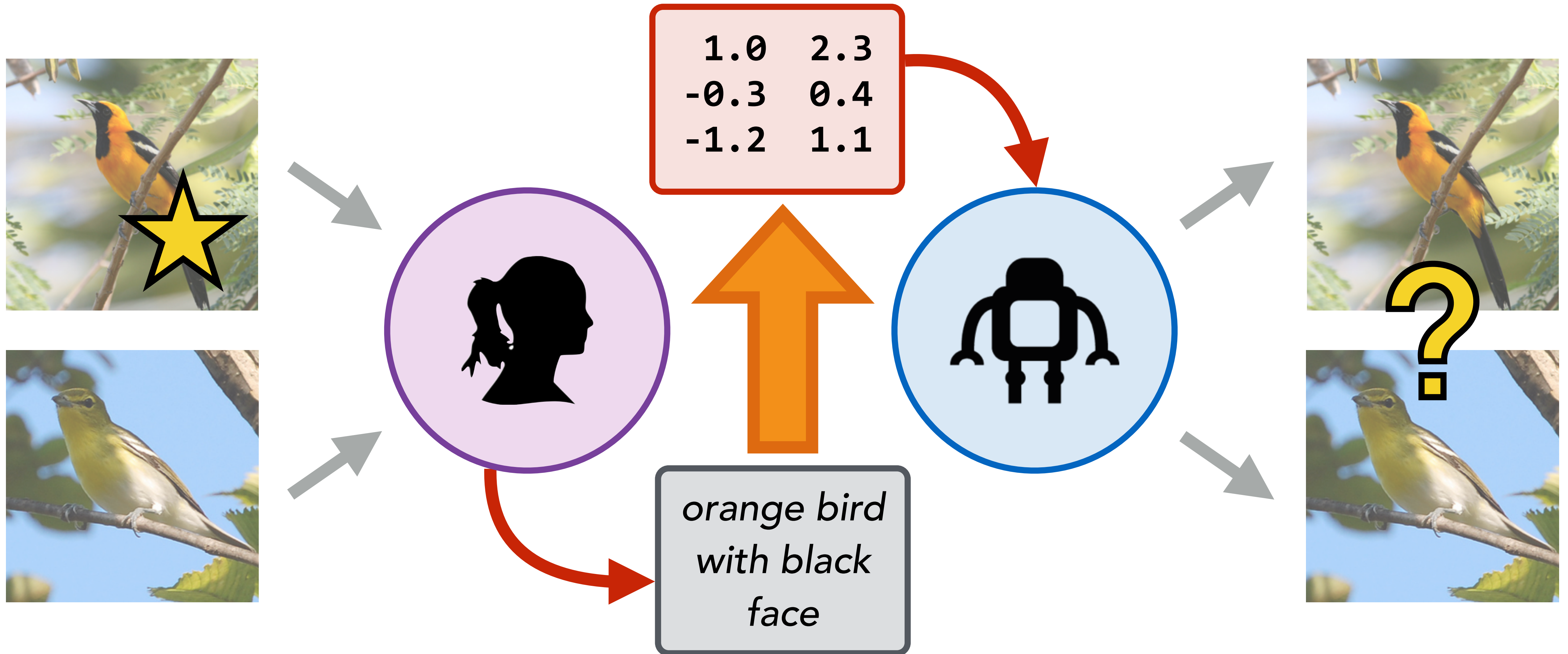


Evaluation: translator-in-the-loop



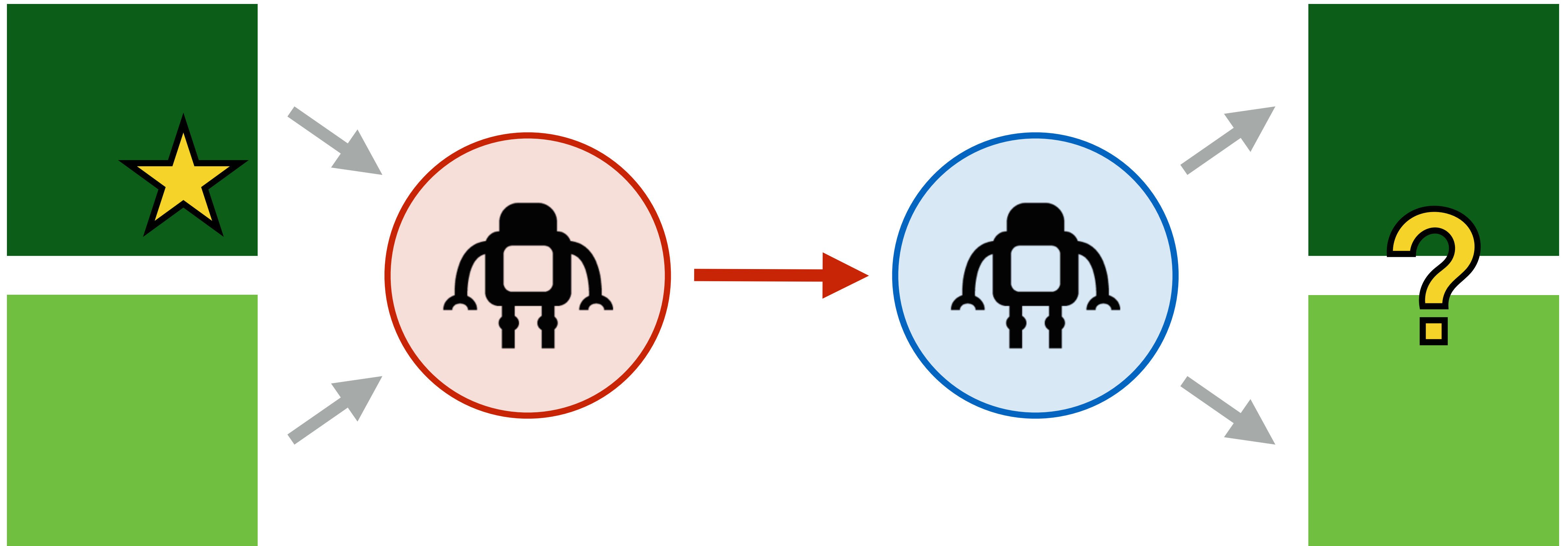


Evaluation: translator-in-the-loop







Experiment: color references





Experiment: color references

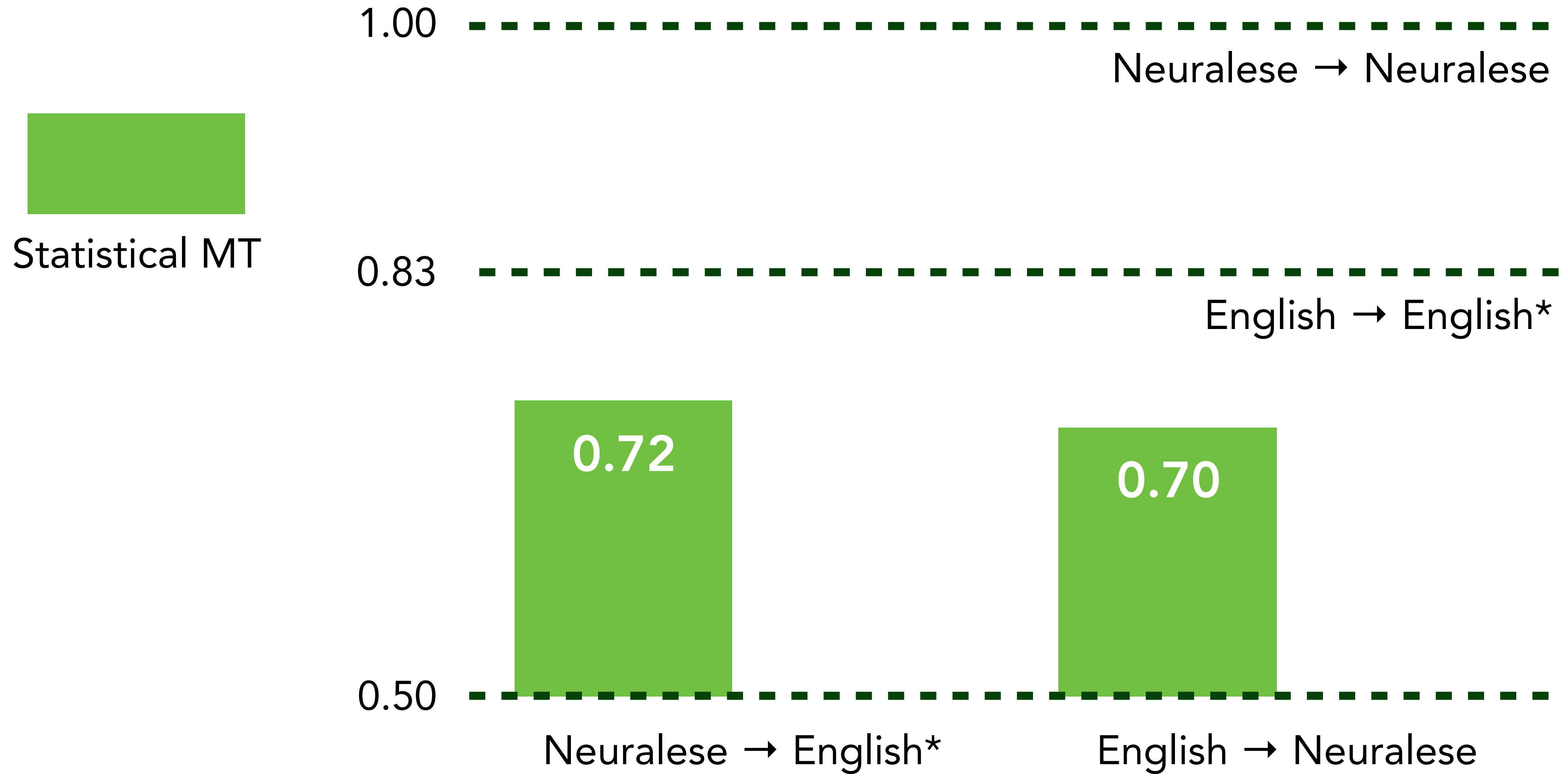
1.00  Neuralese → Neuralese

0.83  English → English*

0.50 

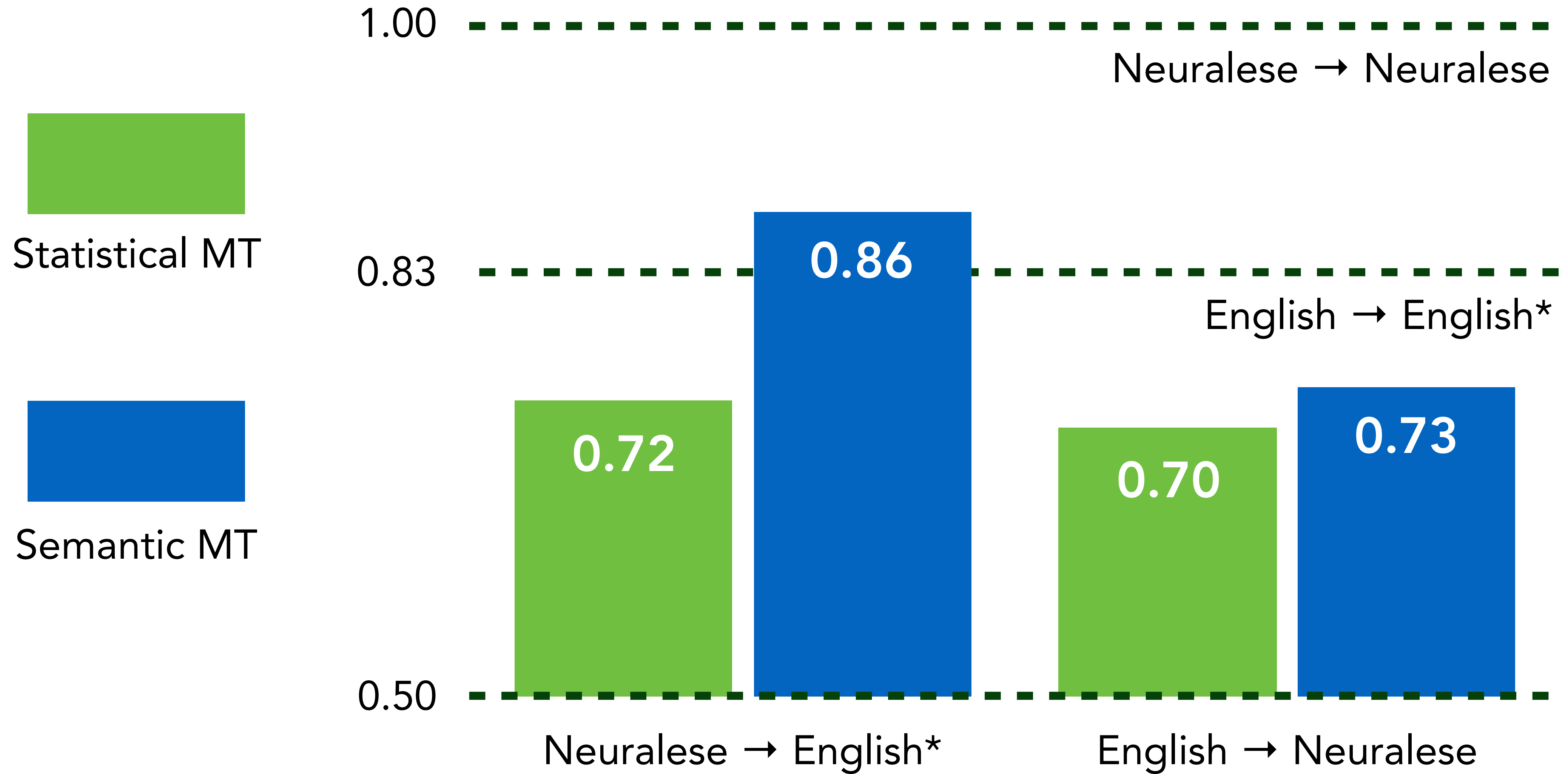


Experiment: color references



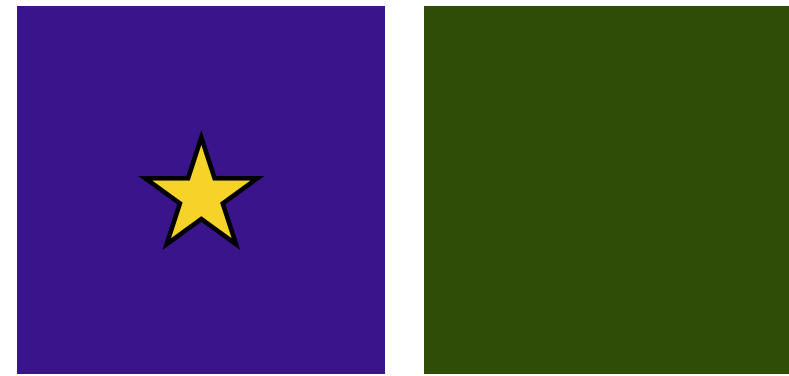


Experiment: color references





Experiment: color references



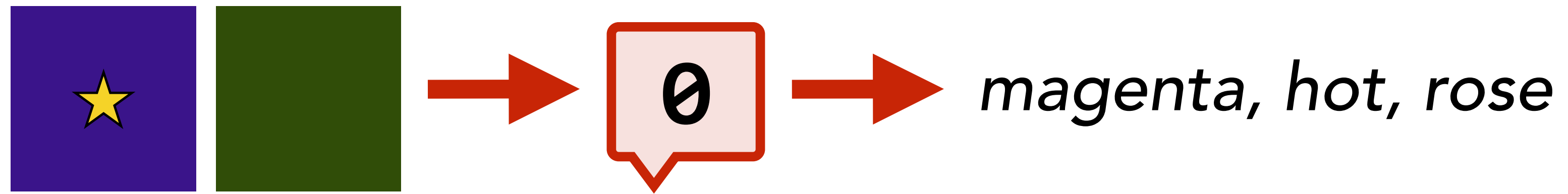


Experiment: color references



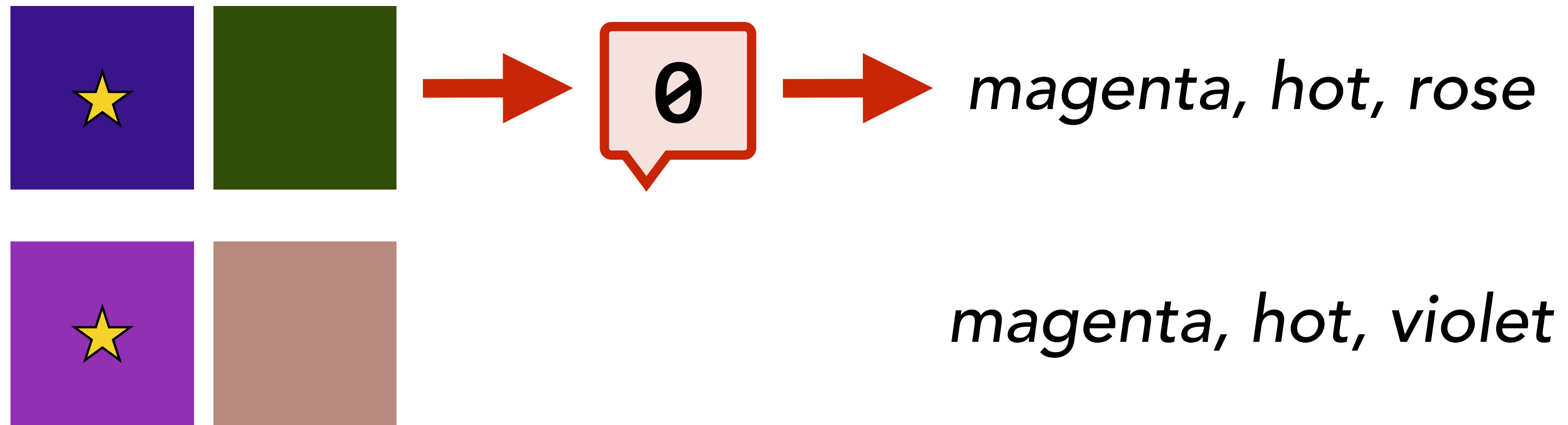


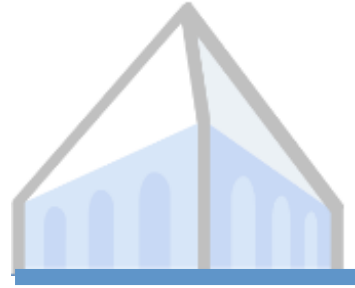
Experiment: color references



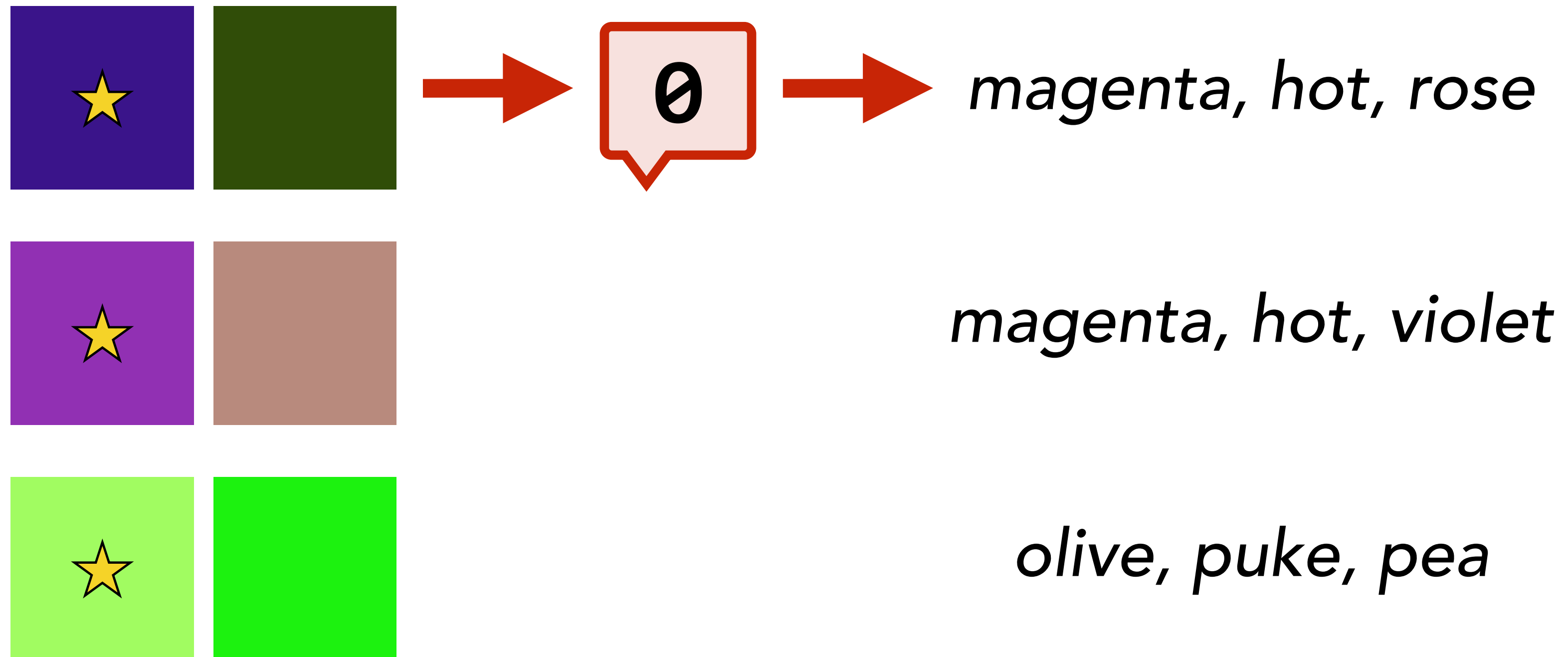


Experiment: color references



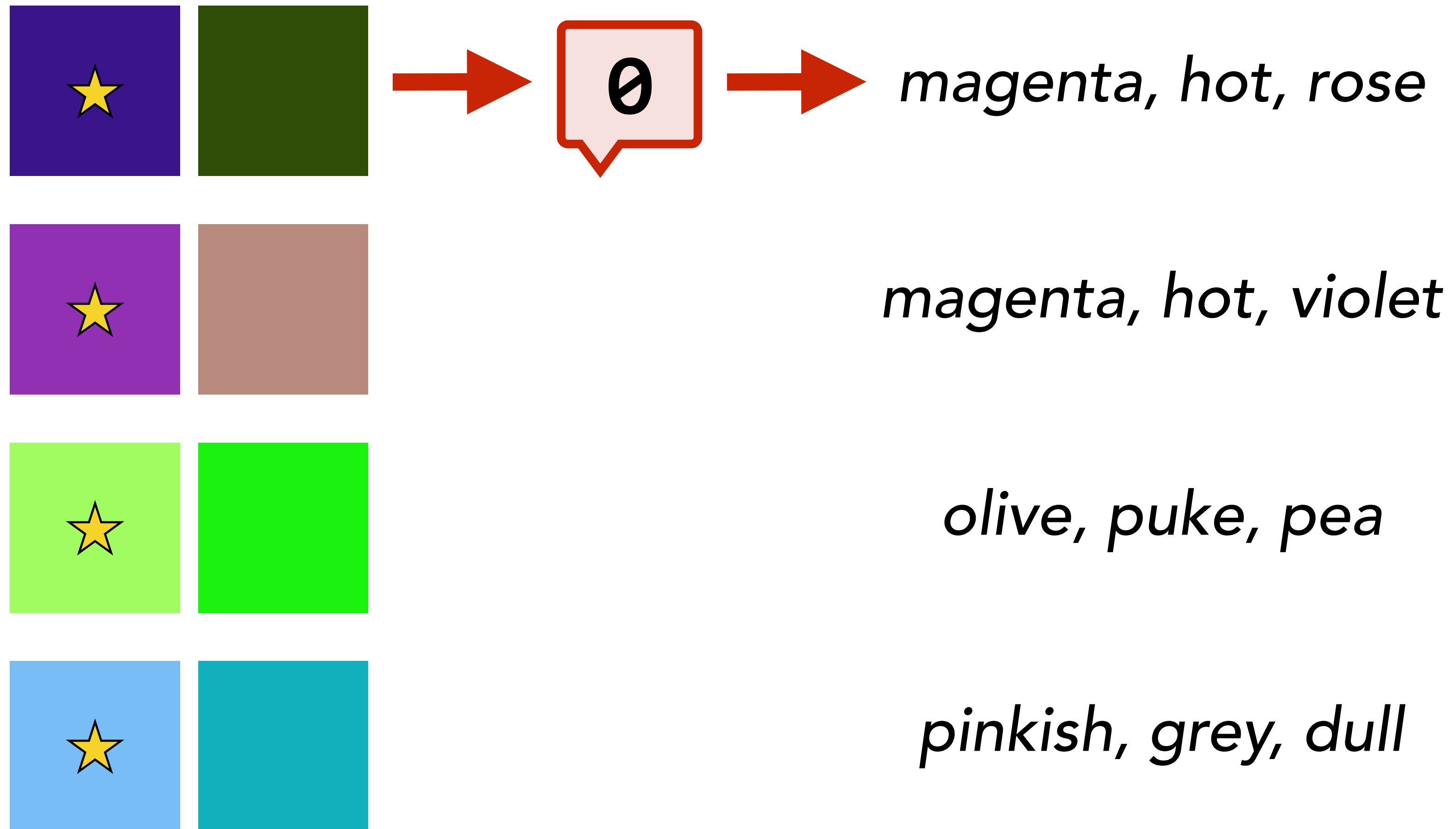


Experiment: color references



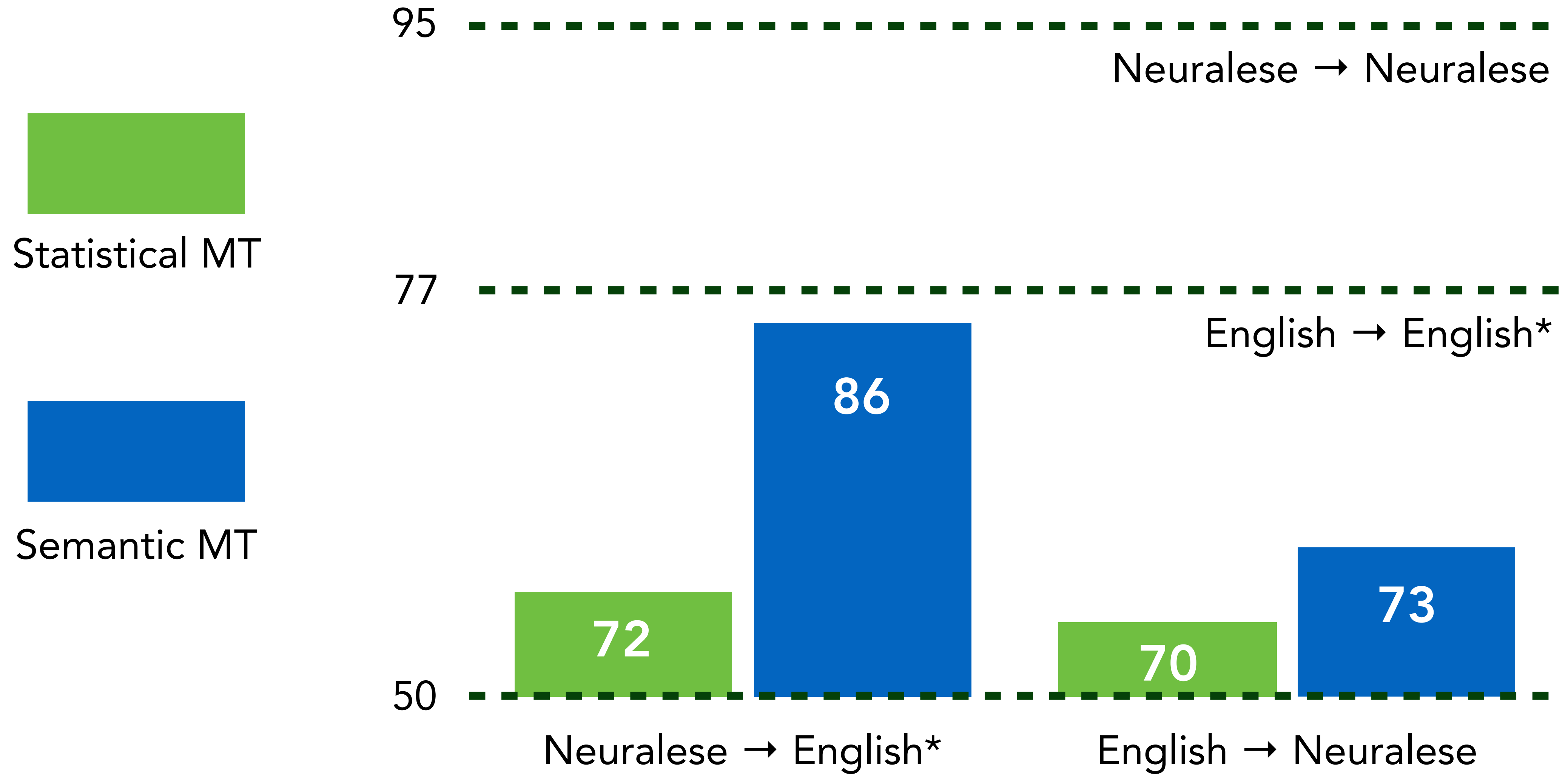


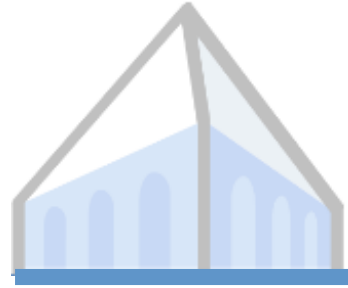
Experiment: color references





Experiment: image references





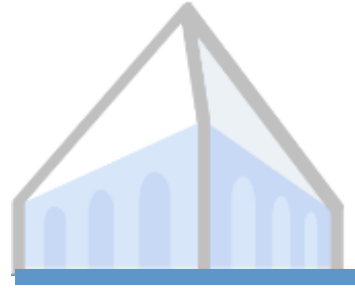
Experiment: image references



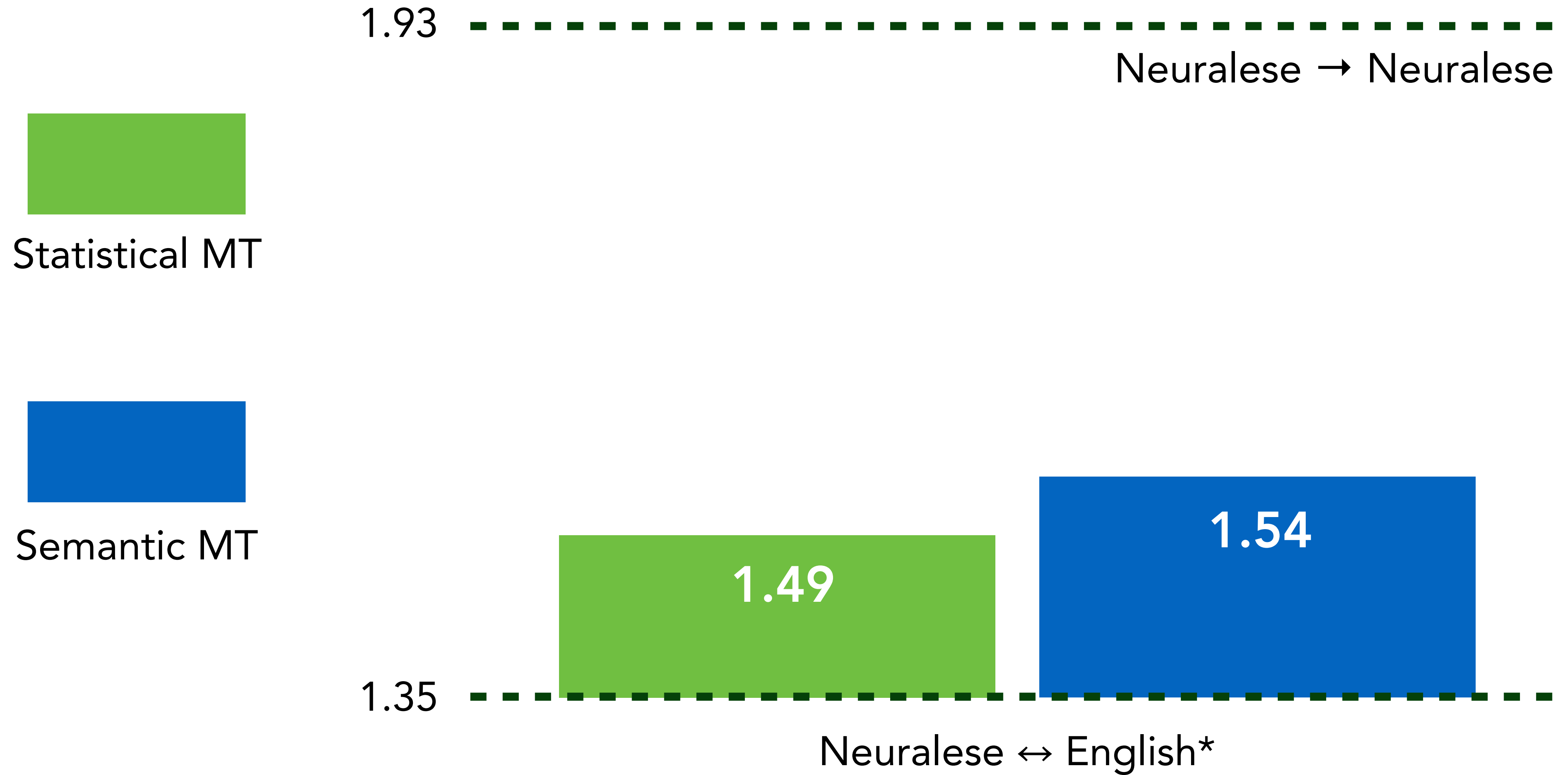
large bird, black wings, black crown



small brown, light brown, dark brown



Experiment: driving game





Conclusions

- 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

- 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

- 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!



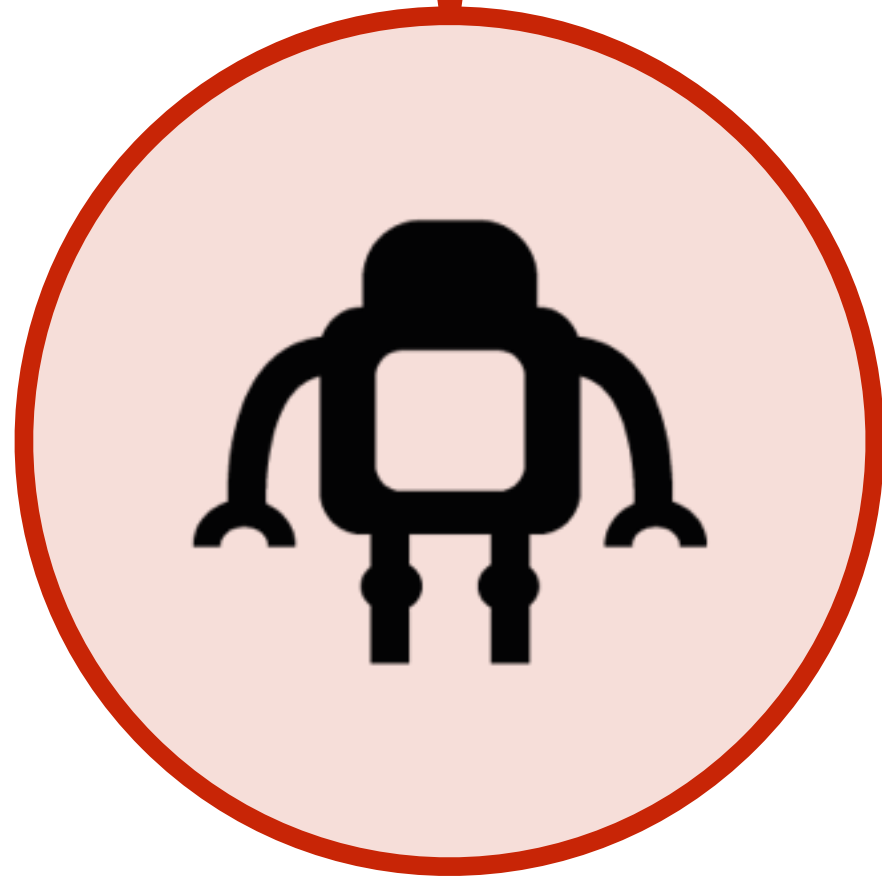
What about compositionality?

Analogs of linguistic structure
in deep representations



Jacob Andreas and Dan Klein

1.0	2.3
-0.3	0.4
-1.2	1.1



Thank you!

