

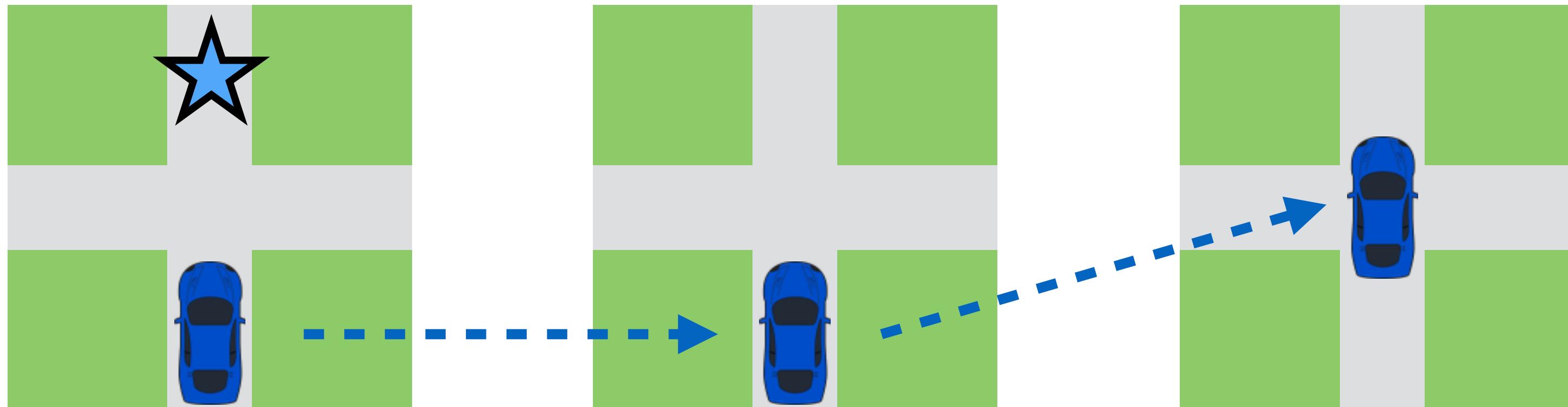
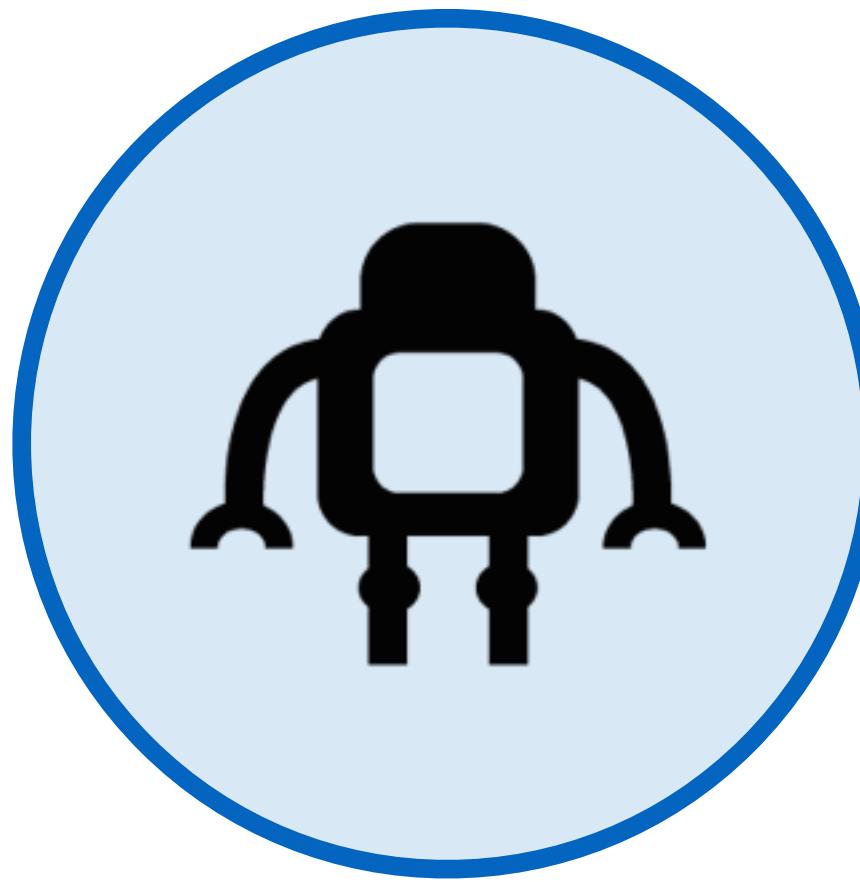
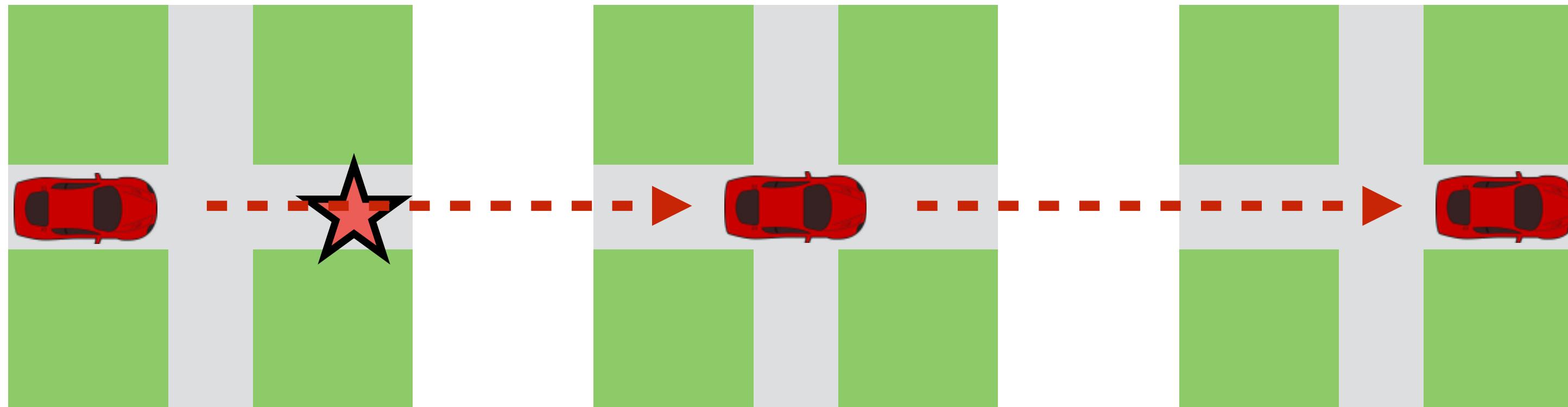
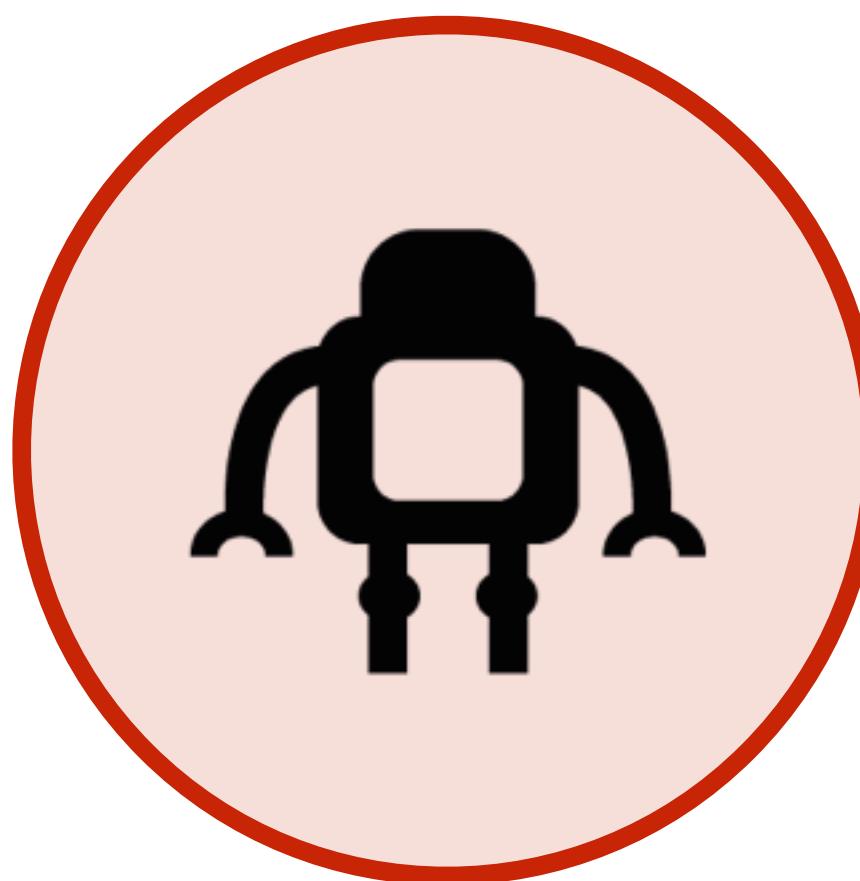
Translating Neuralese



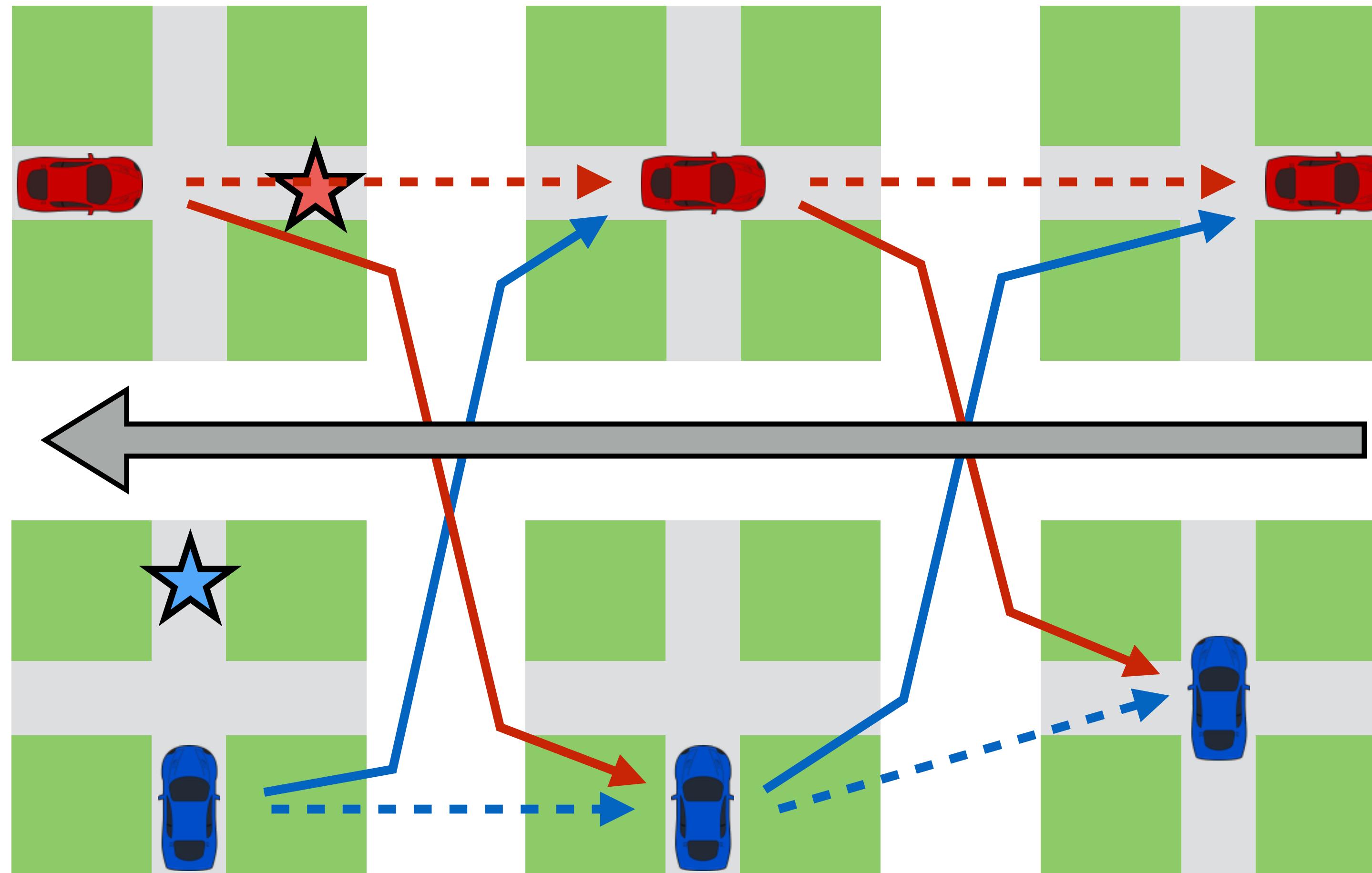
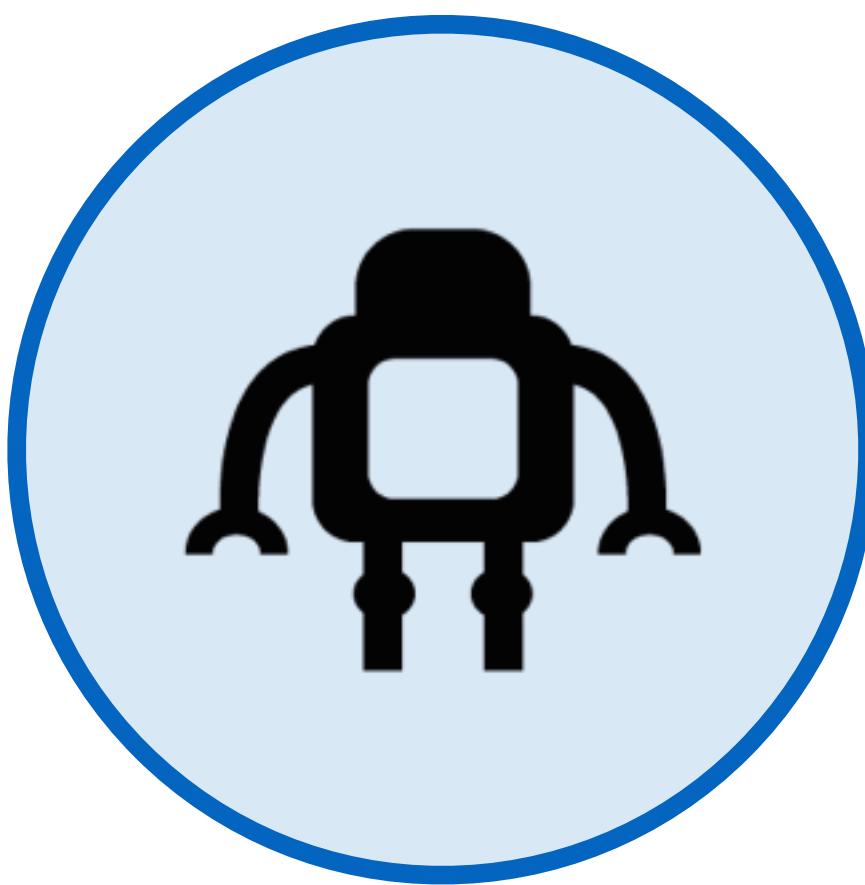
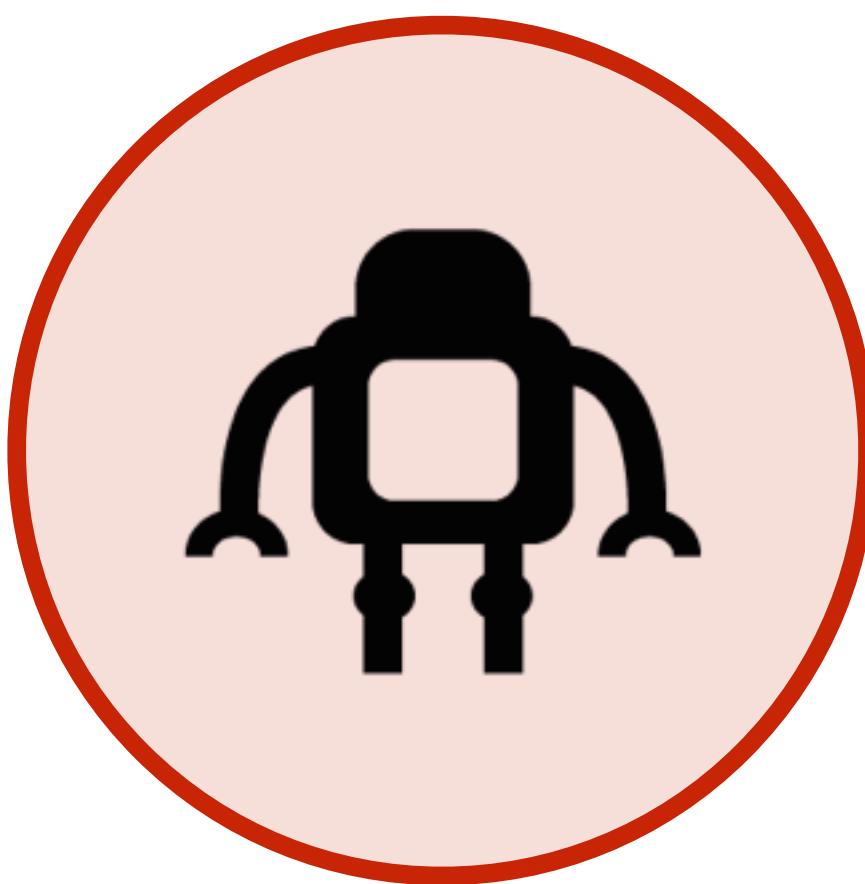
Jacob Andreas, Anca Dragan, and Dan Klein



Learning to Communicate

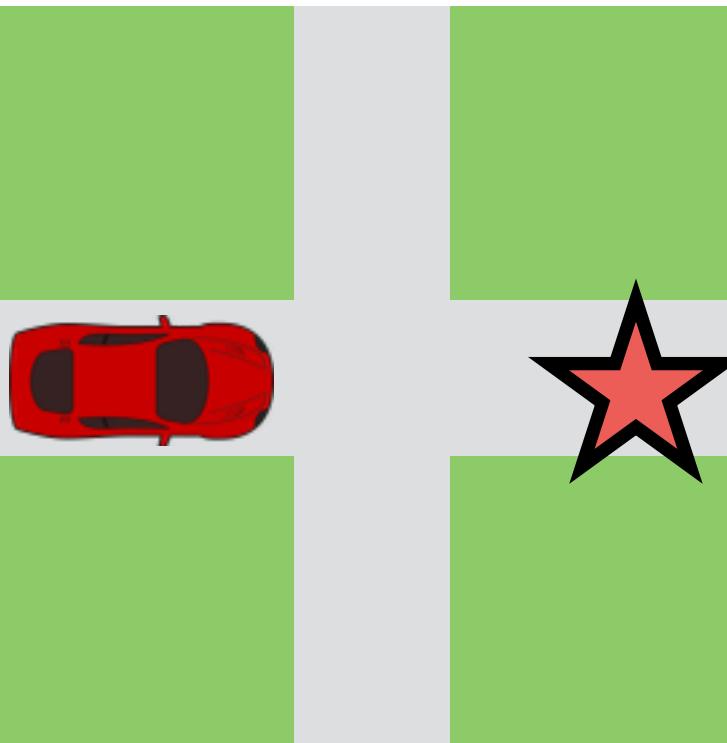
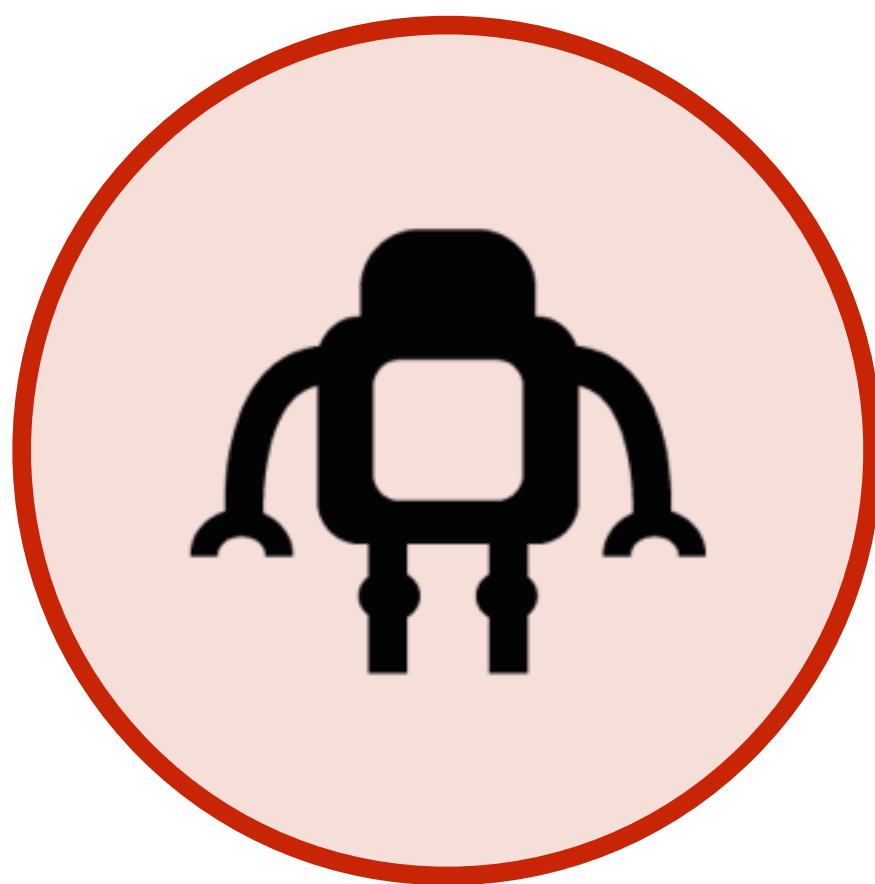


Learning to Communicate

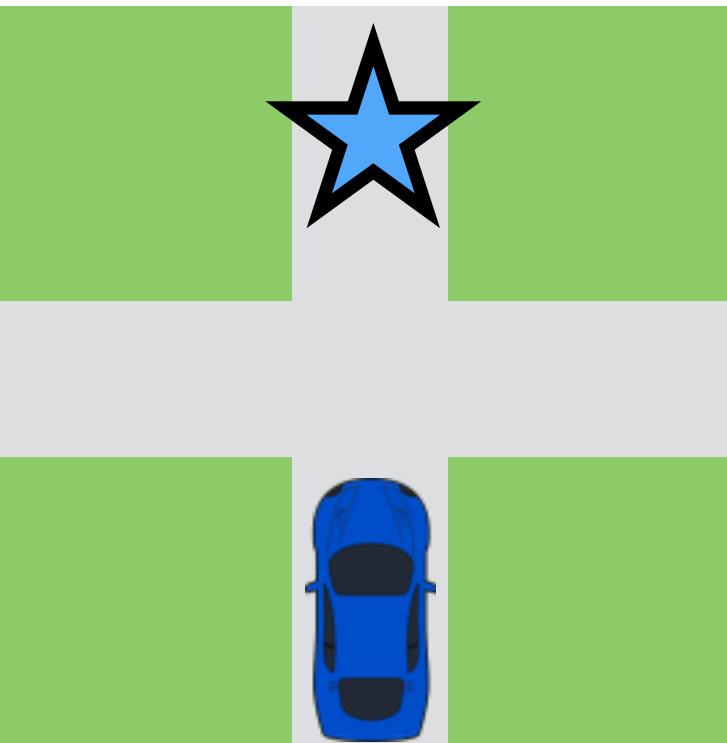
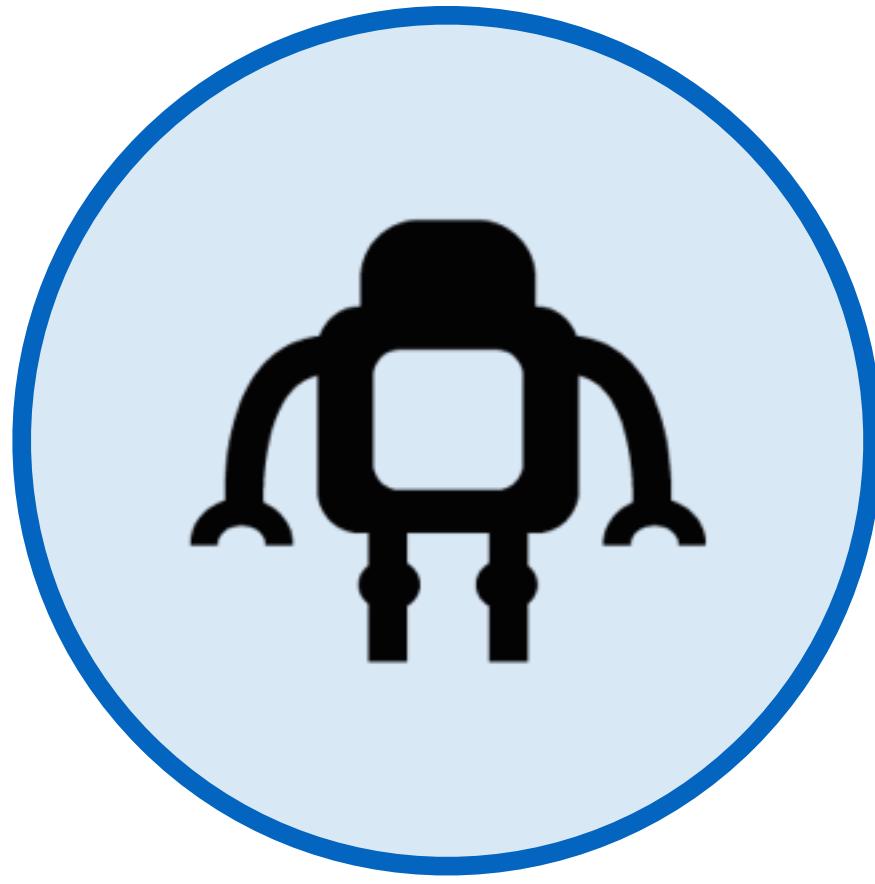
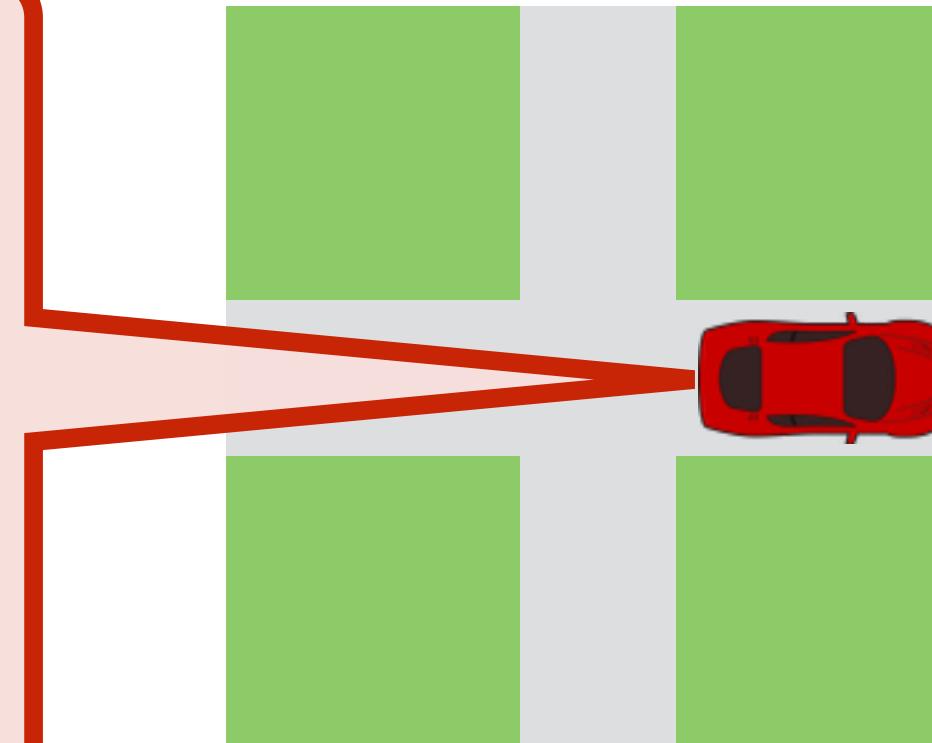




Neuralese

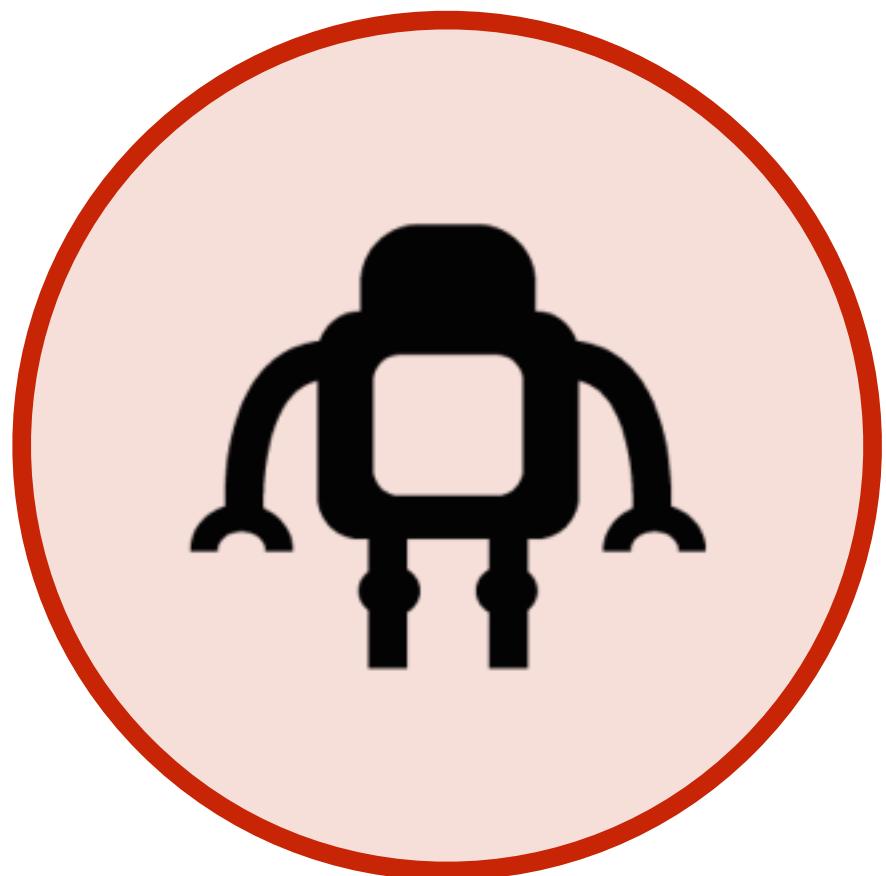


| | |
|------|-----|
| 1.0 | 2.3 |
| -0.3 | 0.4 |
| -1.2 | 1.1 |

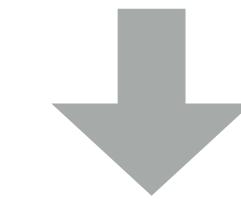




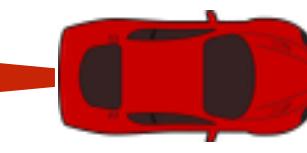
Translating neuralese

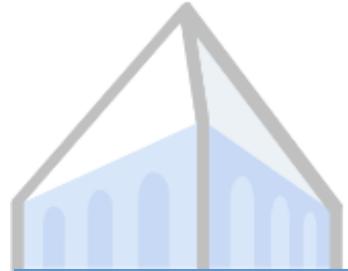


| | |
|------|-----|
| 1.0 | 2.3 |
| -0.3 | 0.4 |
| -1.2 | 1.1 |



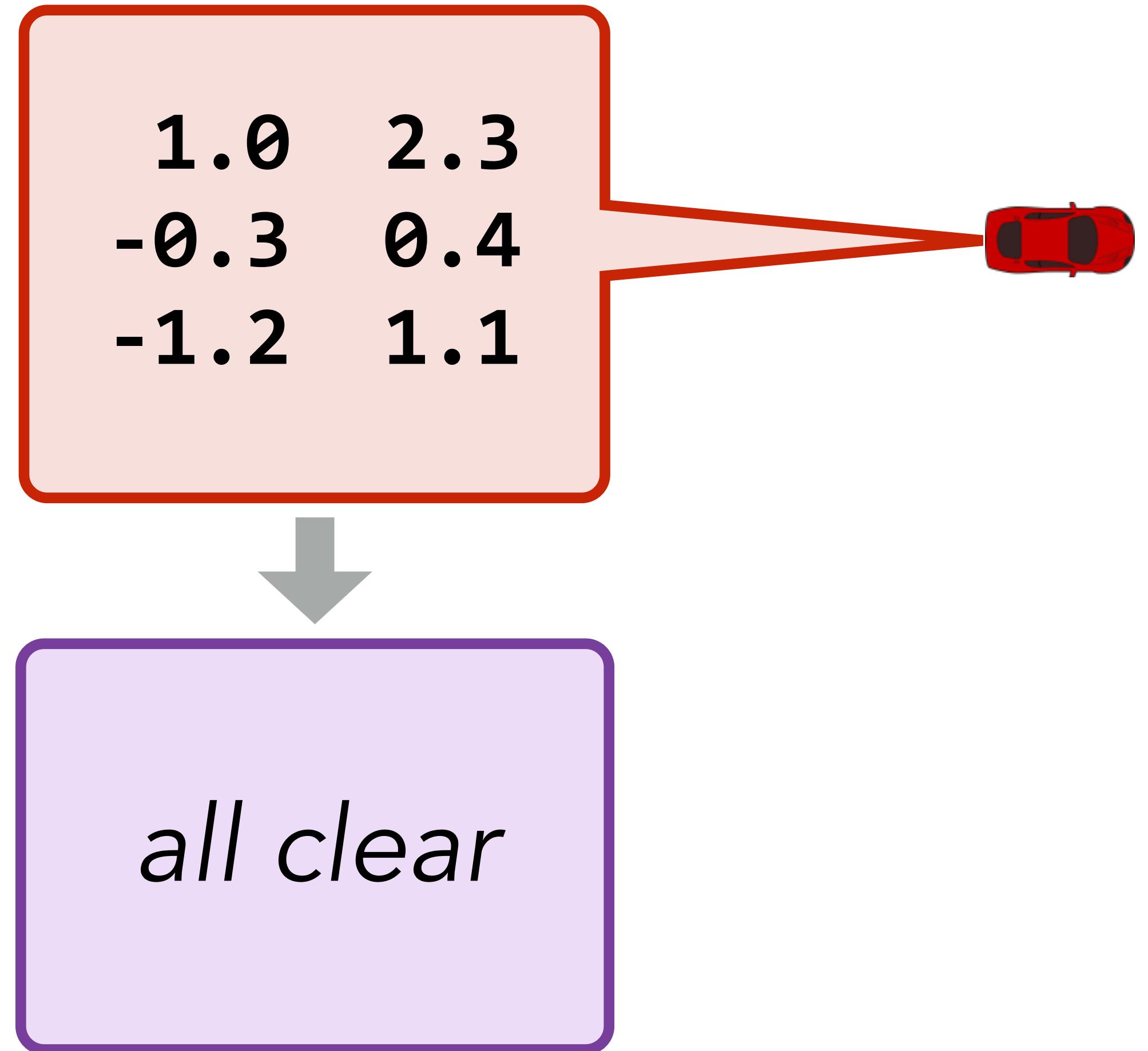
all clear





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

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

a



a

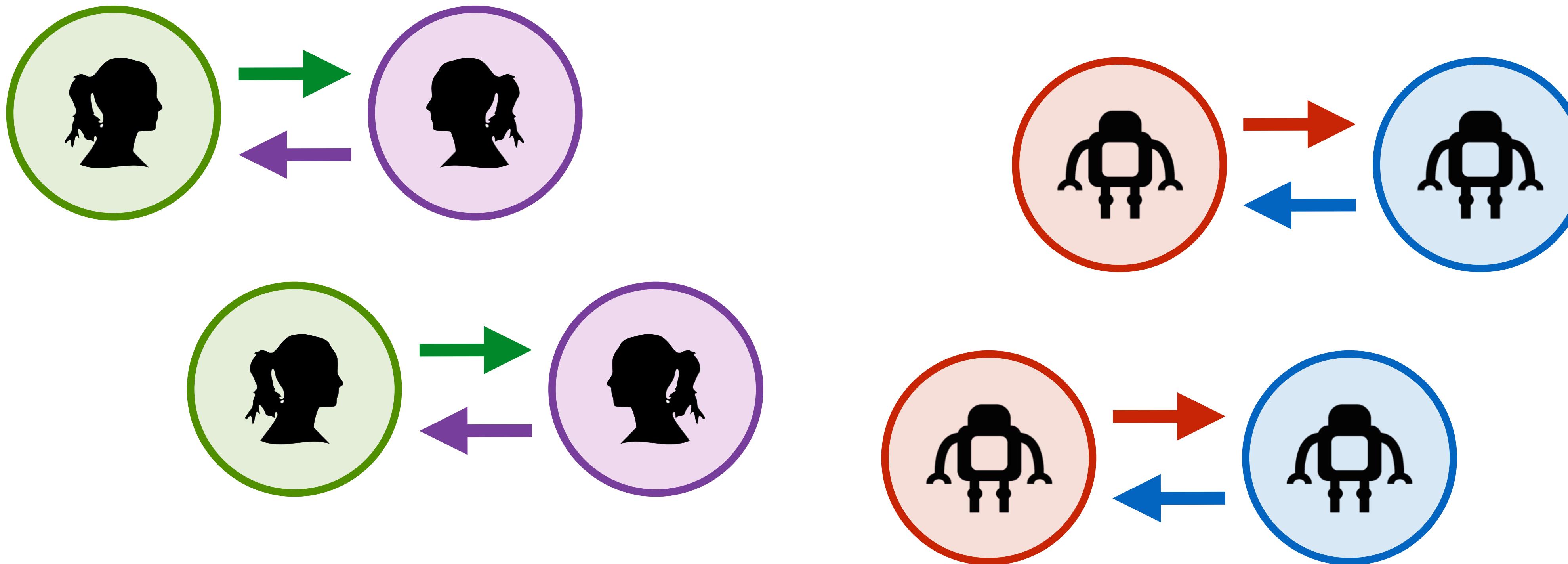


| | |
|------|-----|
| 1.0 | 2.3 |
| -0.3 | 0.4 |
| -1.2 | 1.1 |

all clear



A statistical MT problem



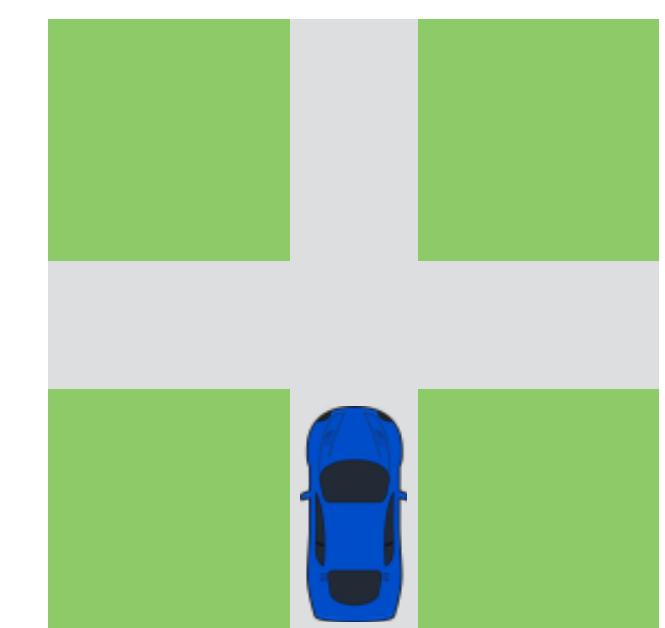
How do we induce a translation model?

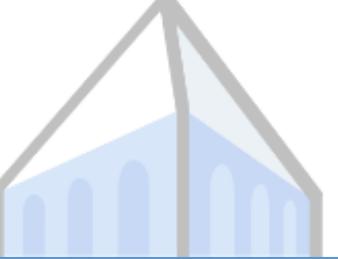


A statistical MT problem

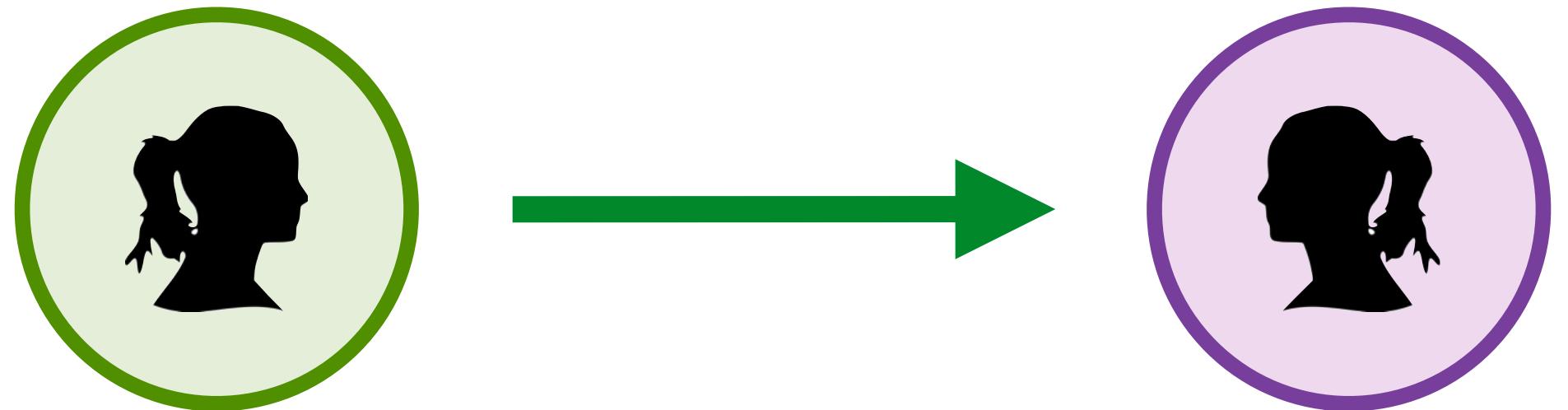
$$\max_a p(\theta | a) p(a)$$

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

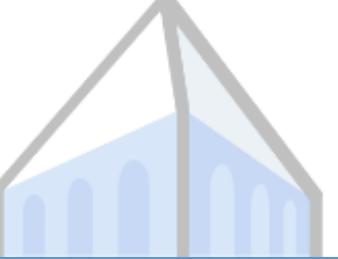




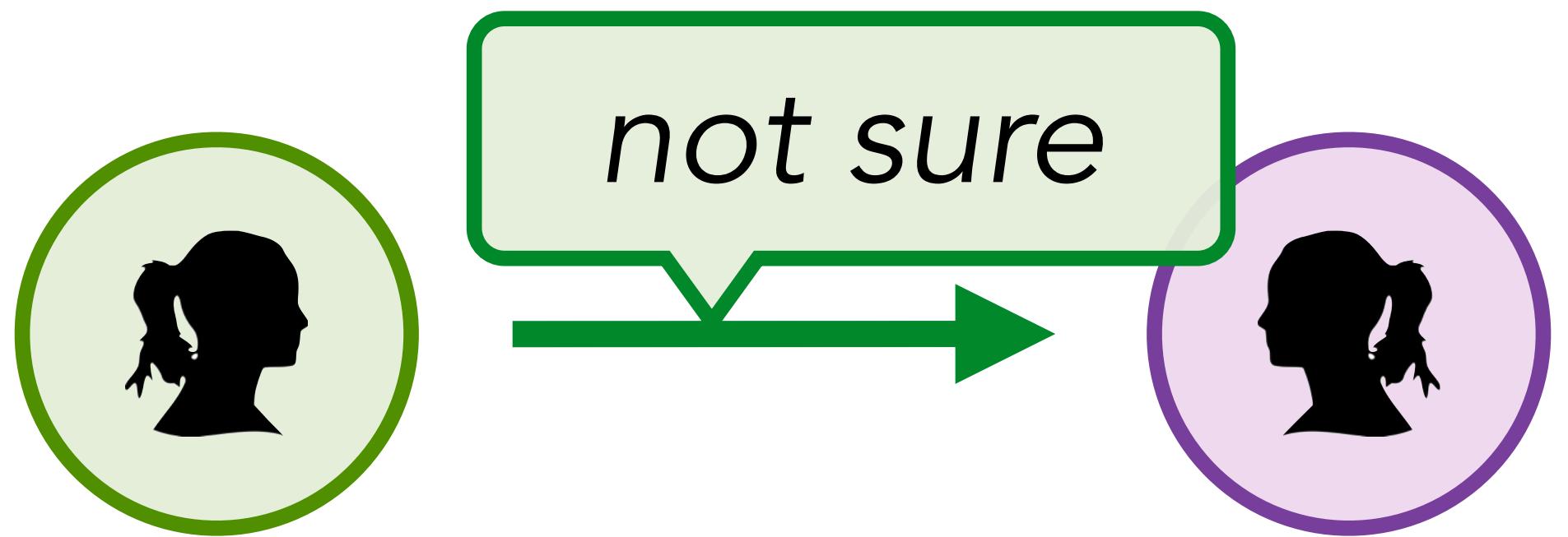
Strategy mismatch



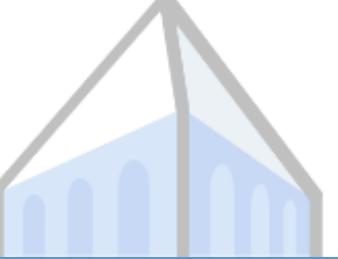
$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{1}{e^x - 1} x^s \frac{dx}{x}$$



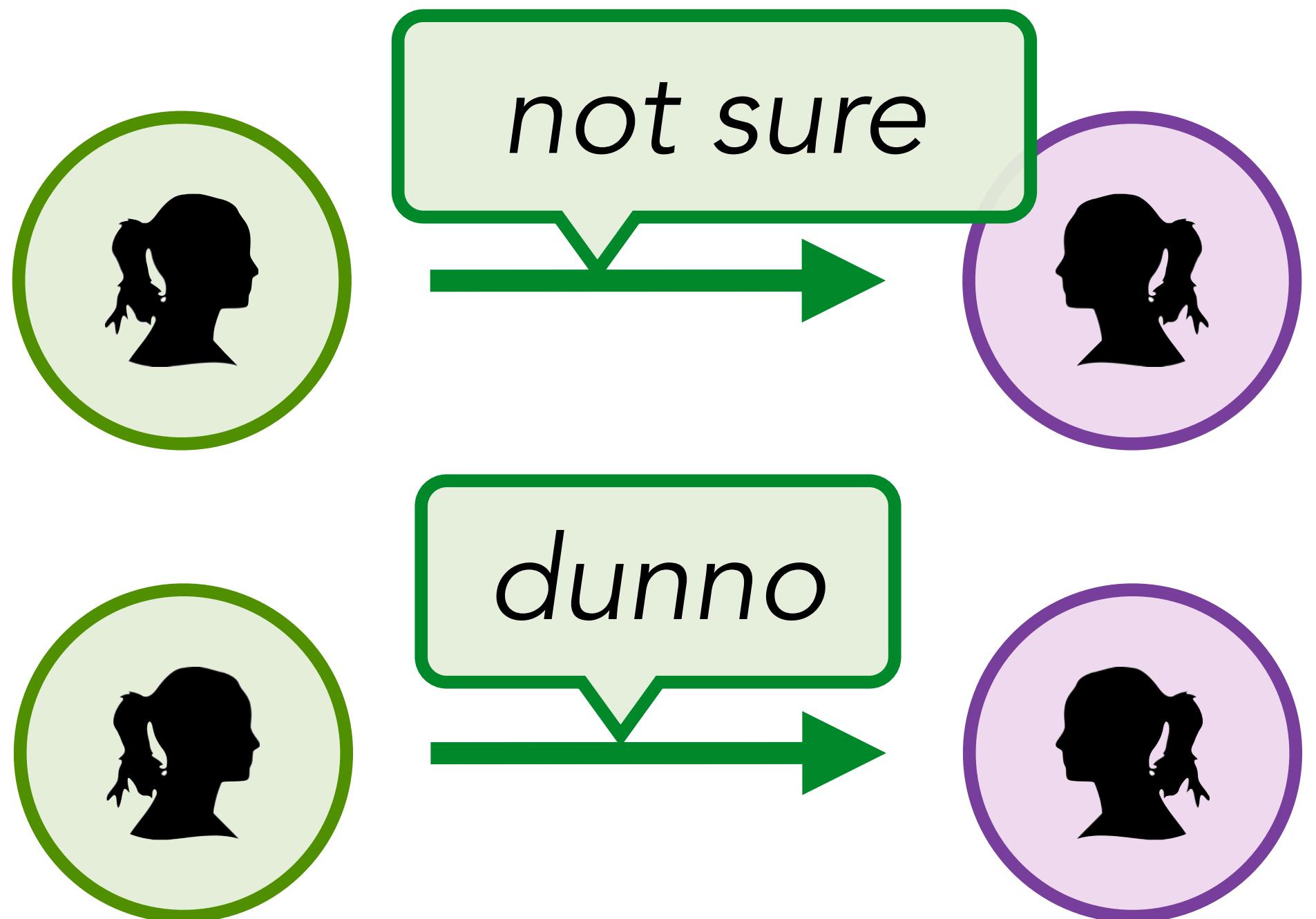
Strategy mismatch

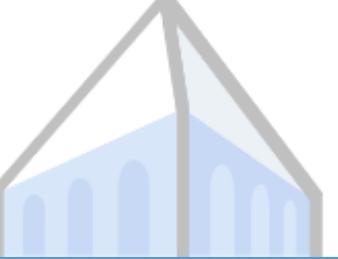


$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{1}{e^x - 1} x^s \frac{dx}{x}$$

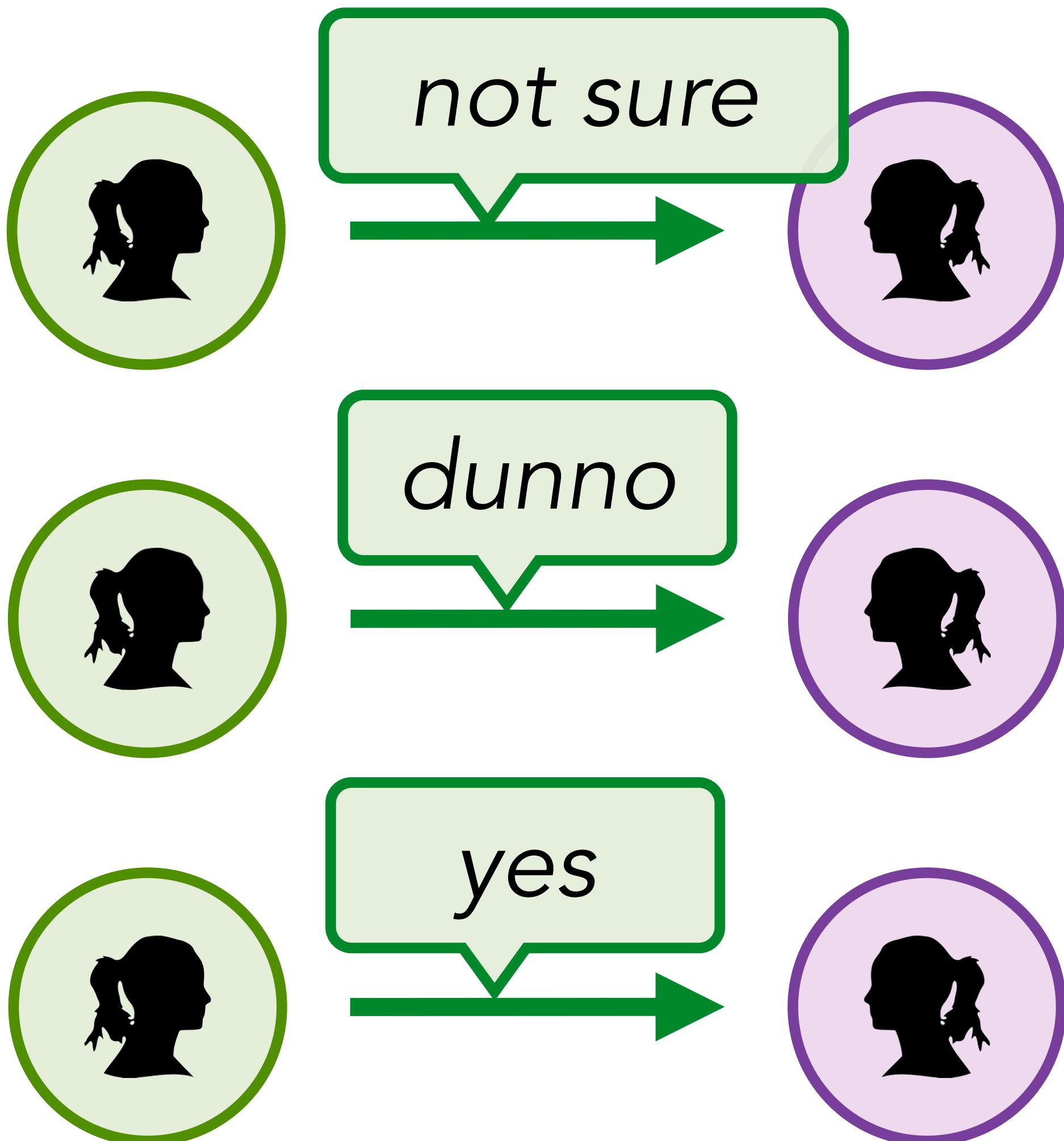


Strategy mismatch



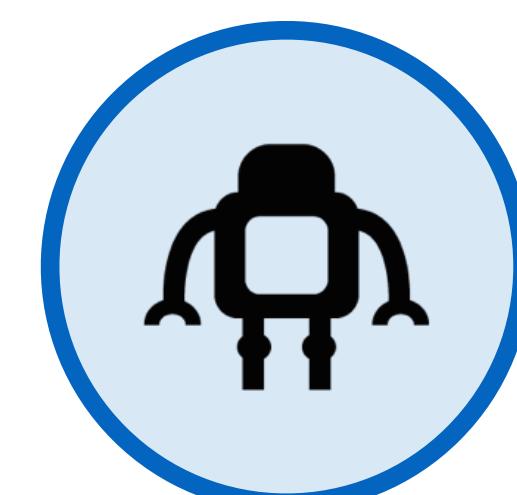
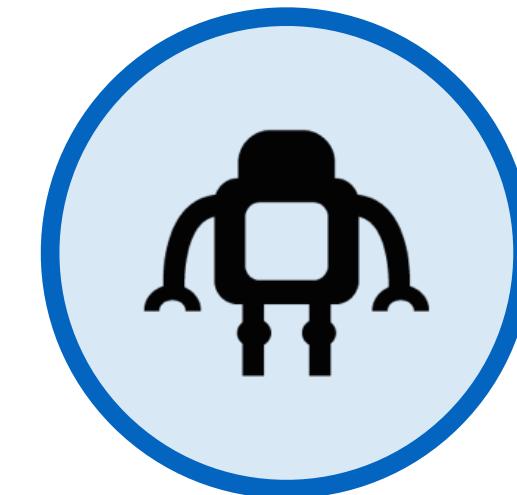
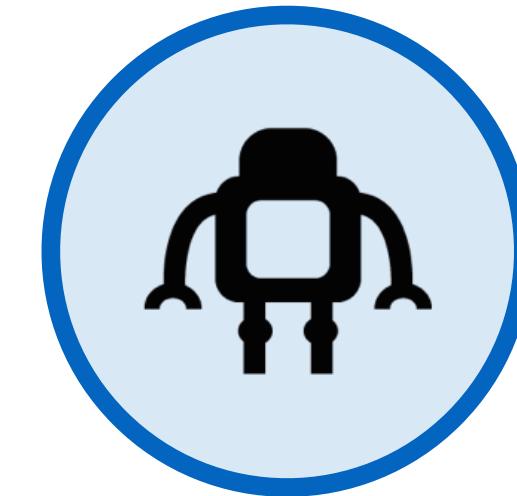
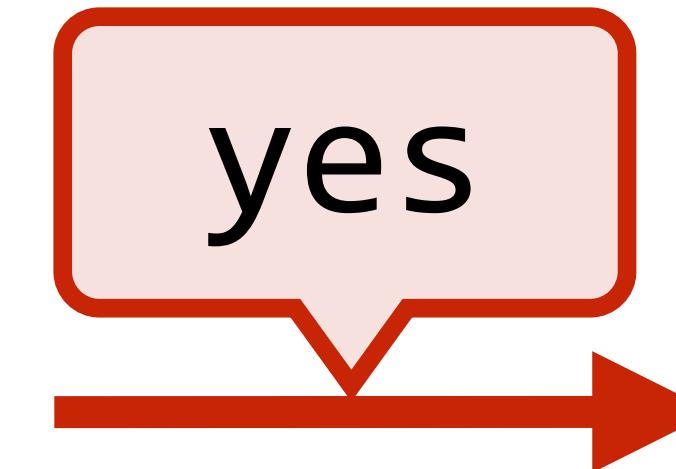
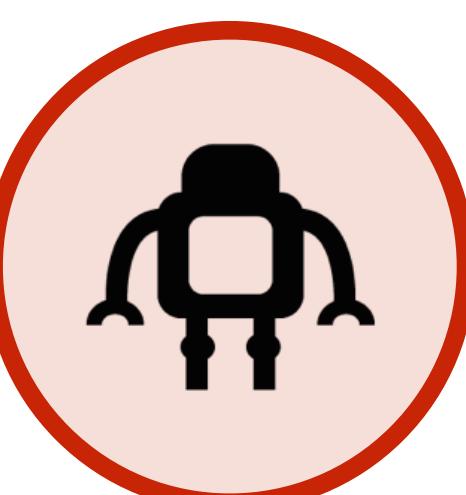
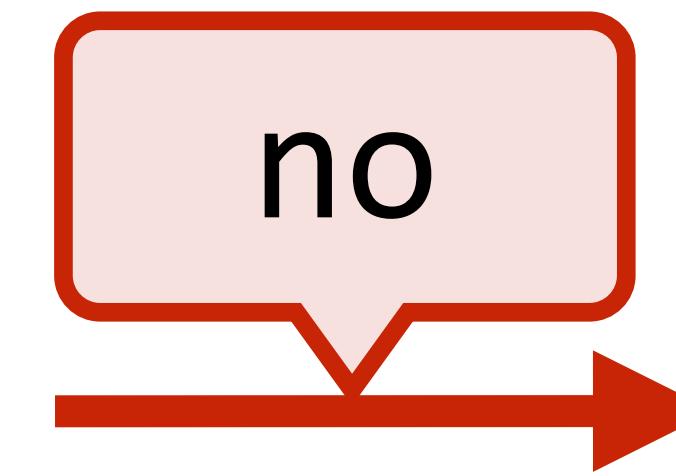
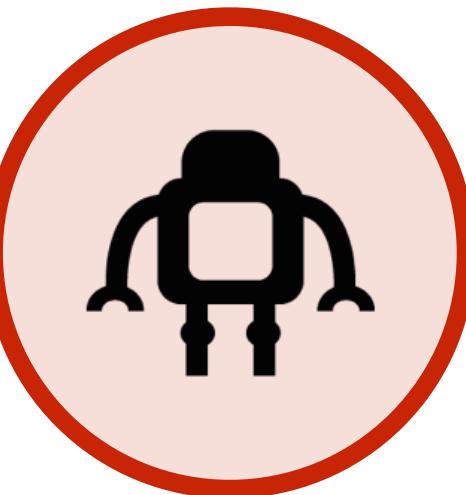
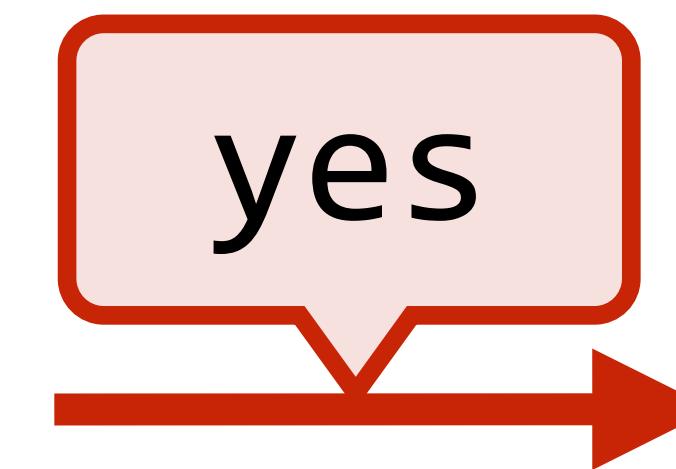
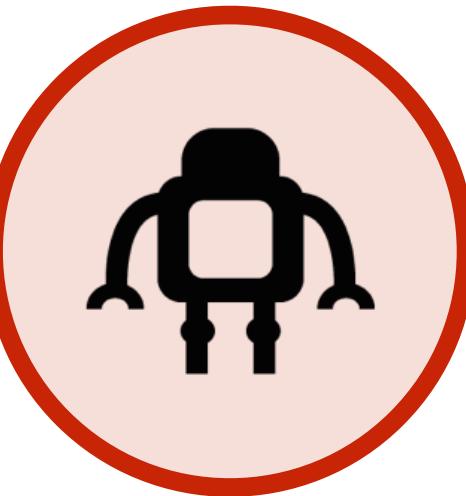
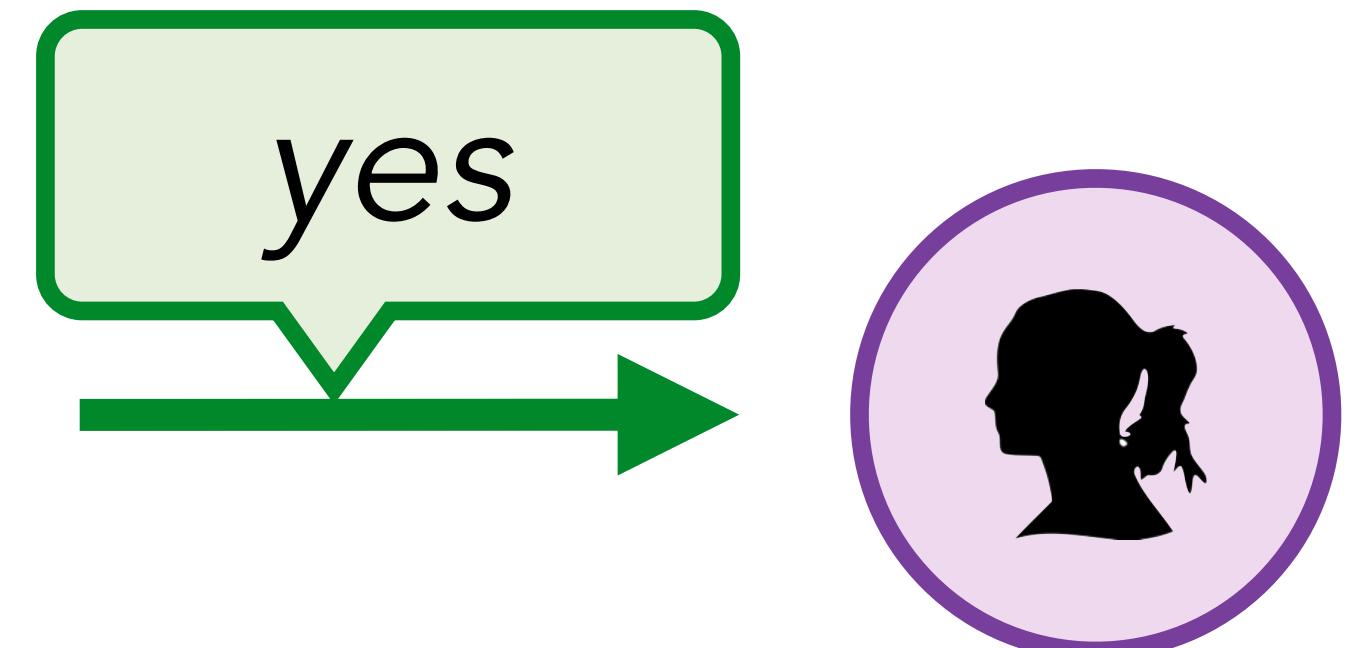
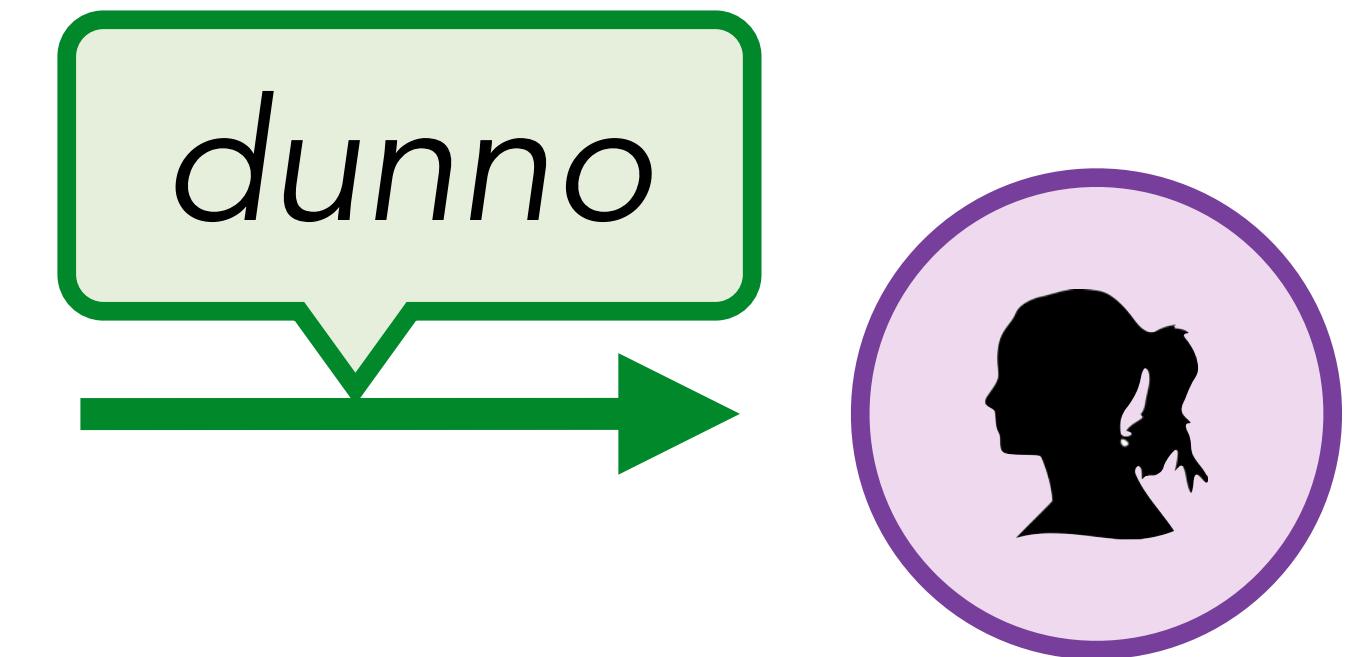
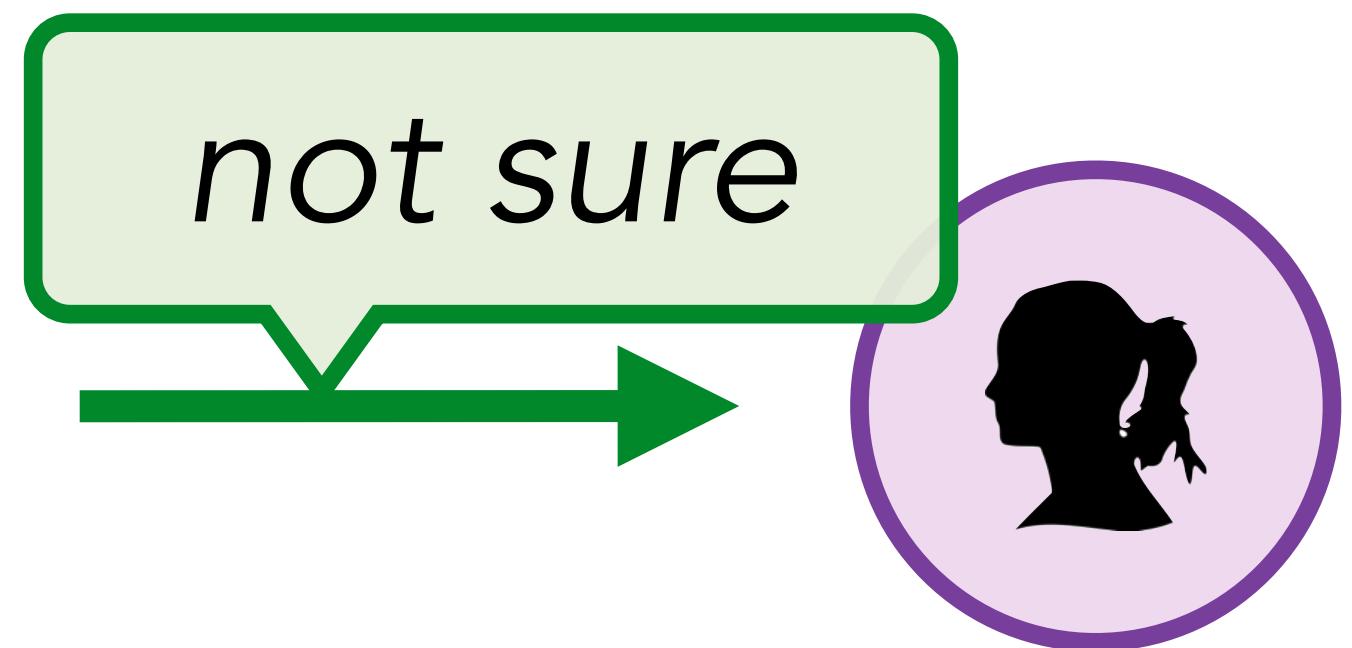


Strategy mismatch



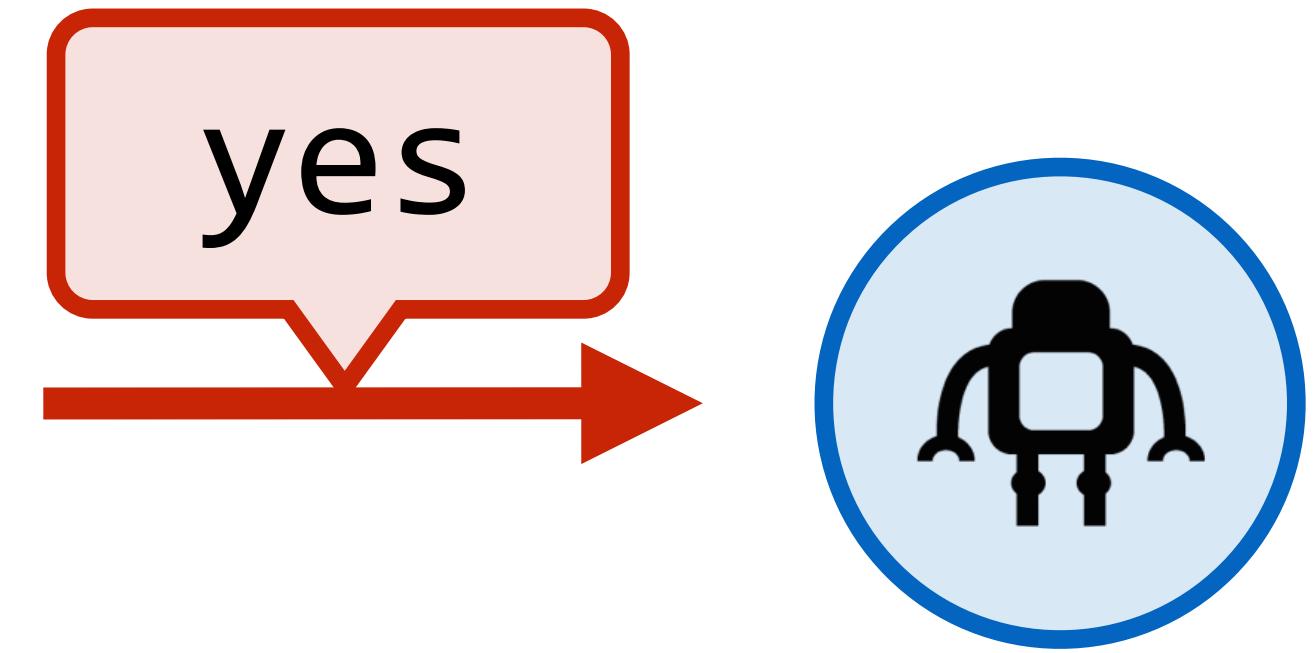
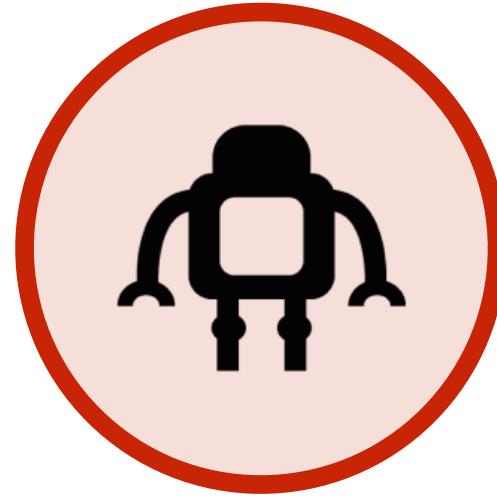
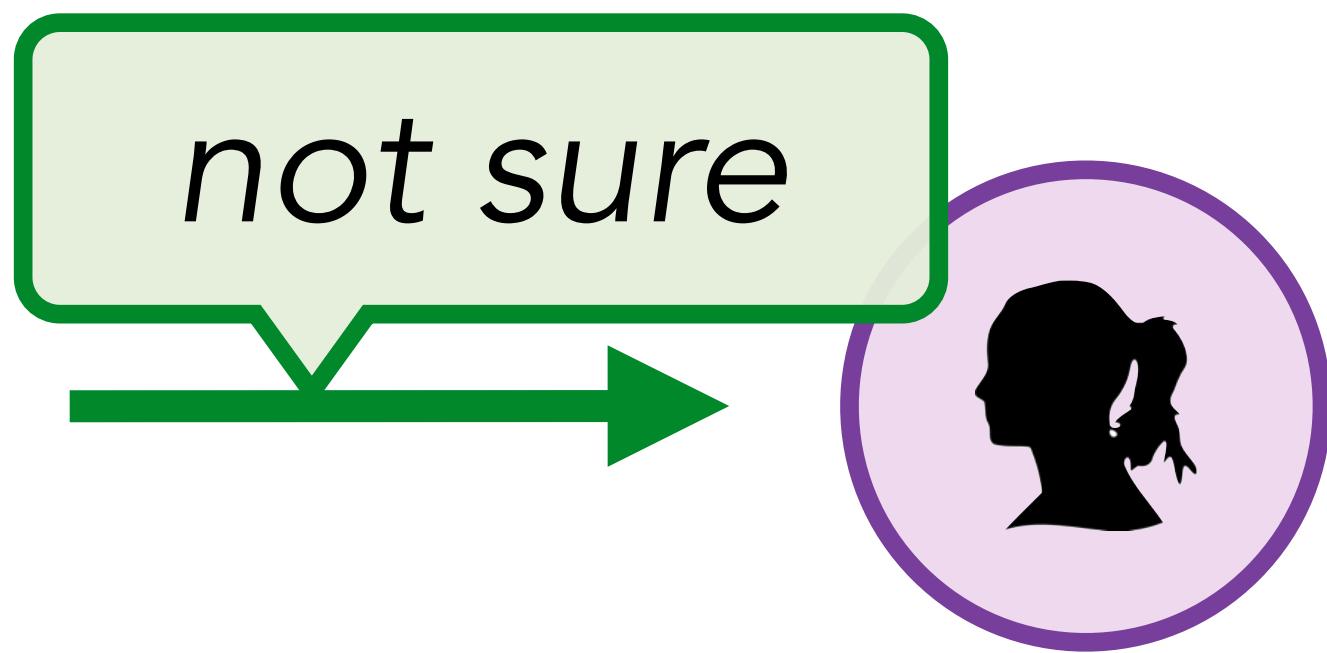


Strategy mismatch





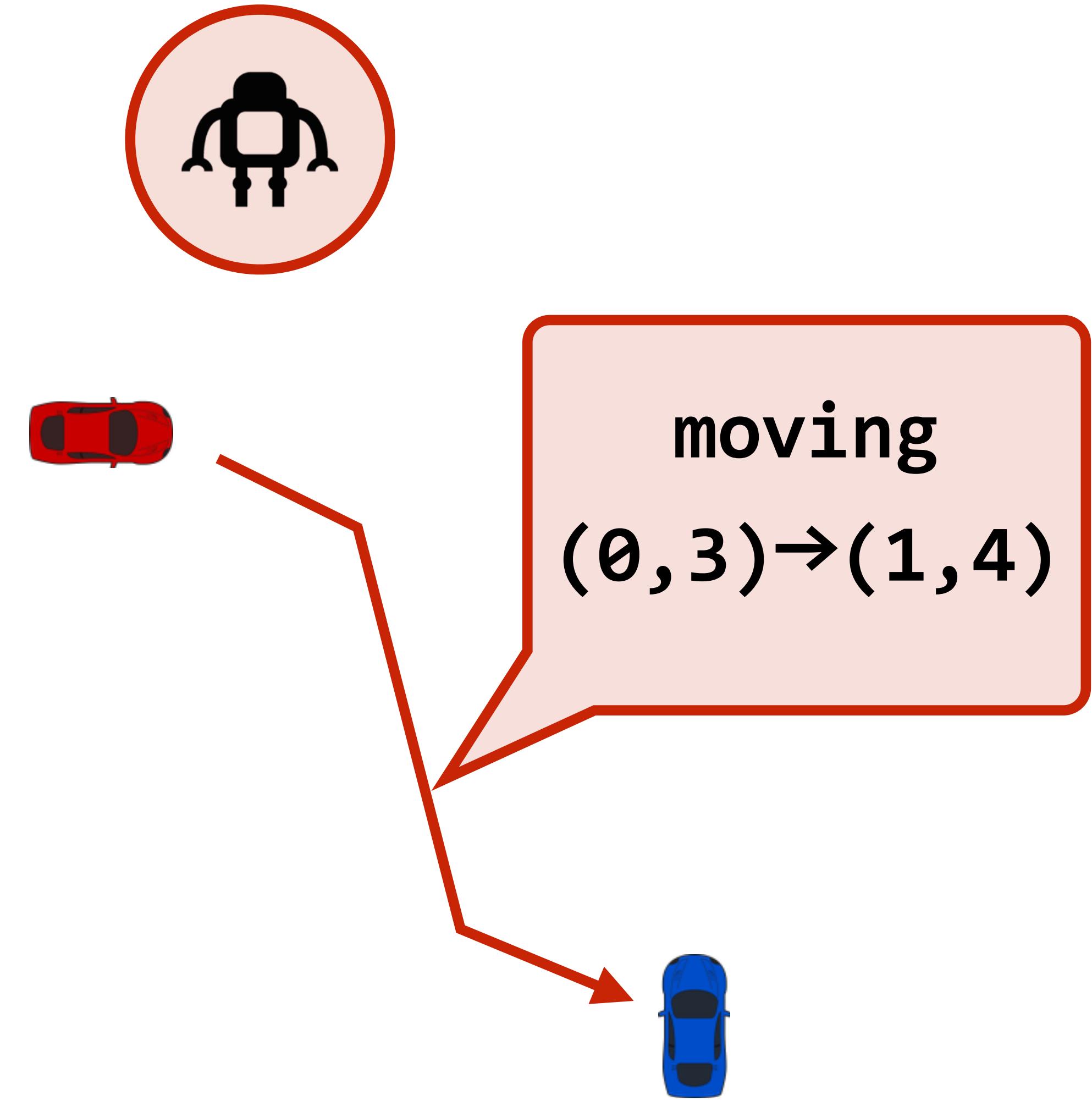
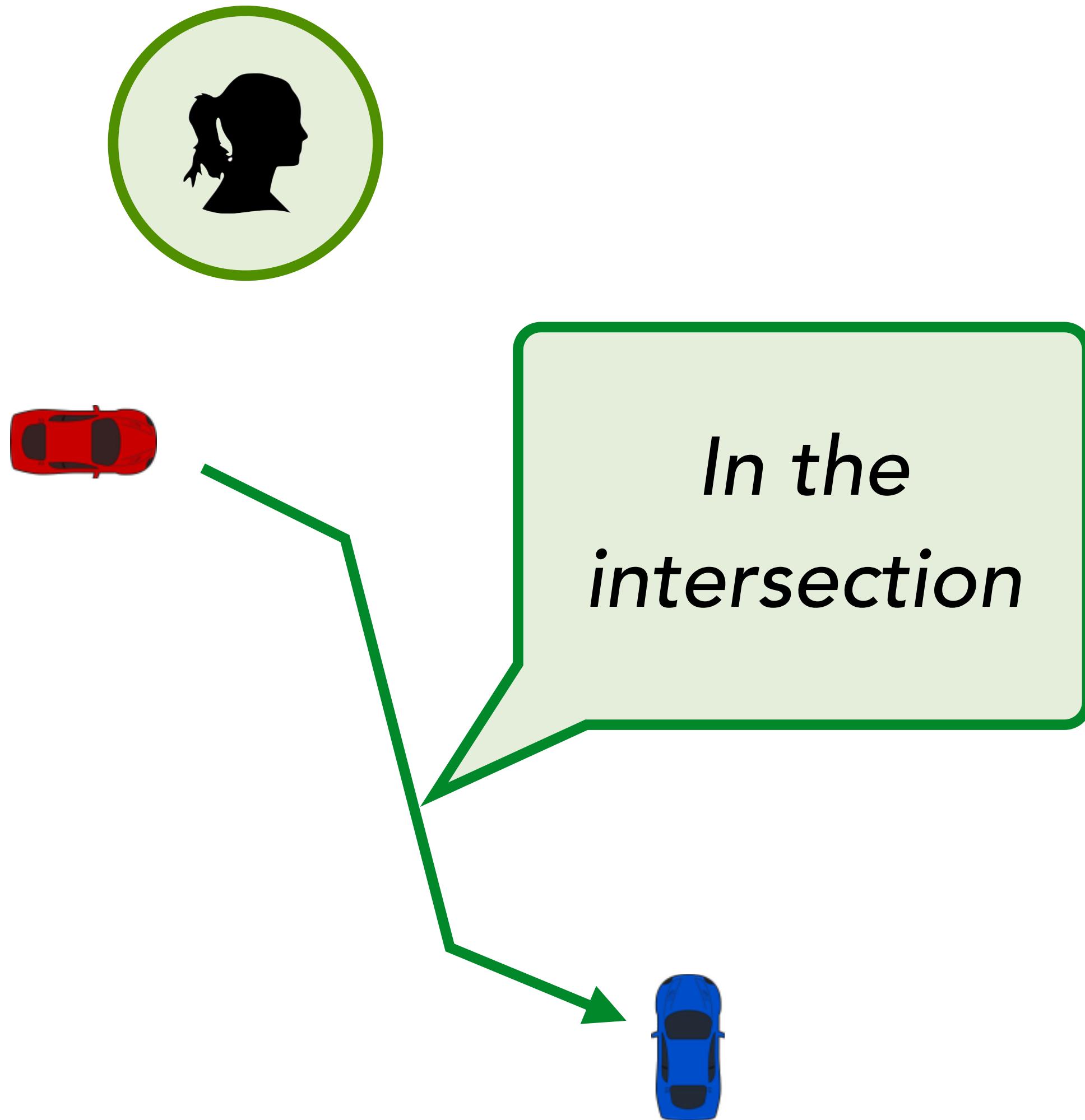
Strategy mismatch



$$\sum p(\theta, \text{[red speech bubble]} | \text{not sure}) p(\text{not sure})$$



Stat MT criterion doesn't capture meaning





Outline

Natural language & neuralese

X Statistical machine translation

Semantic machine translation

Implementation details

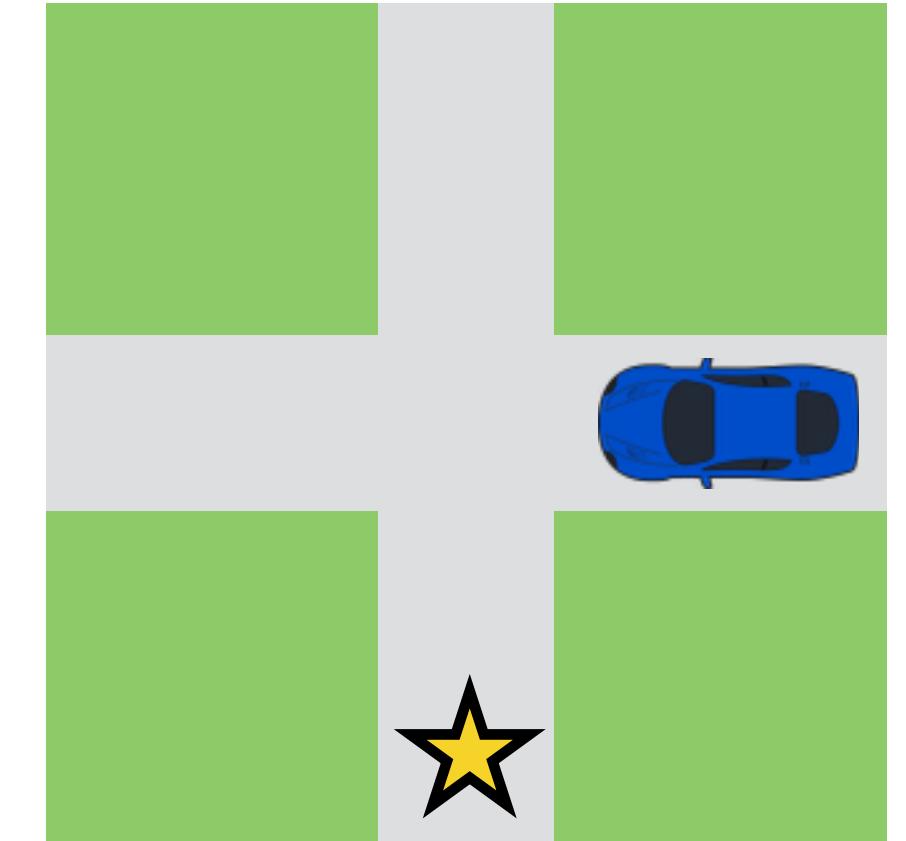
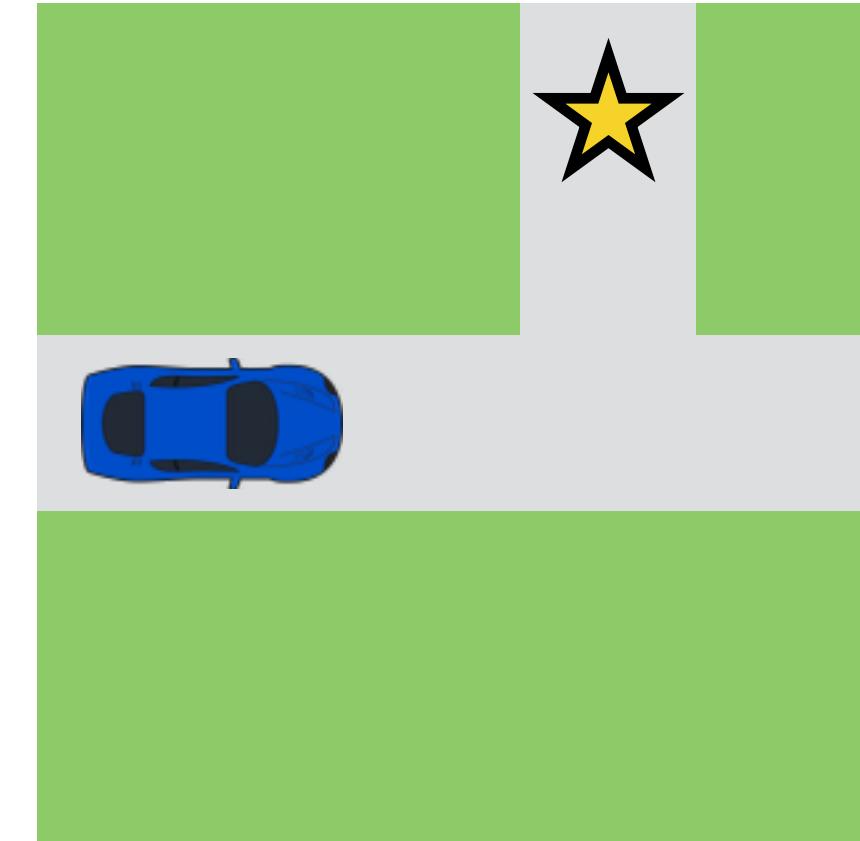
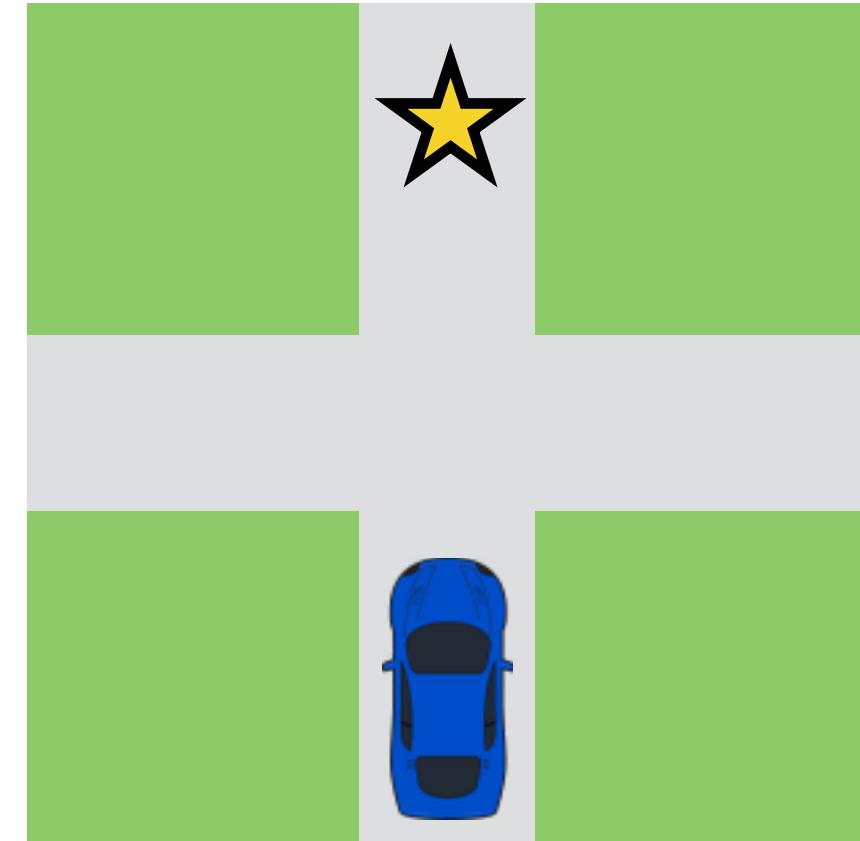
Evaluation



A “semantic MT” problem

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

I'm going
north

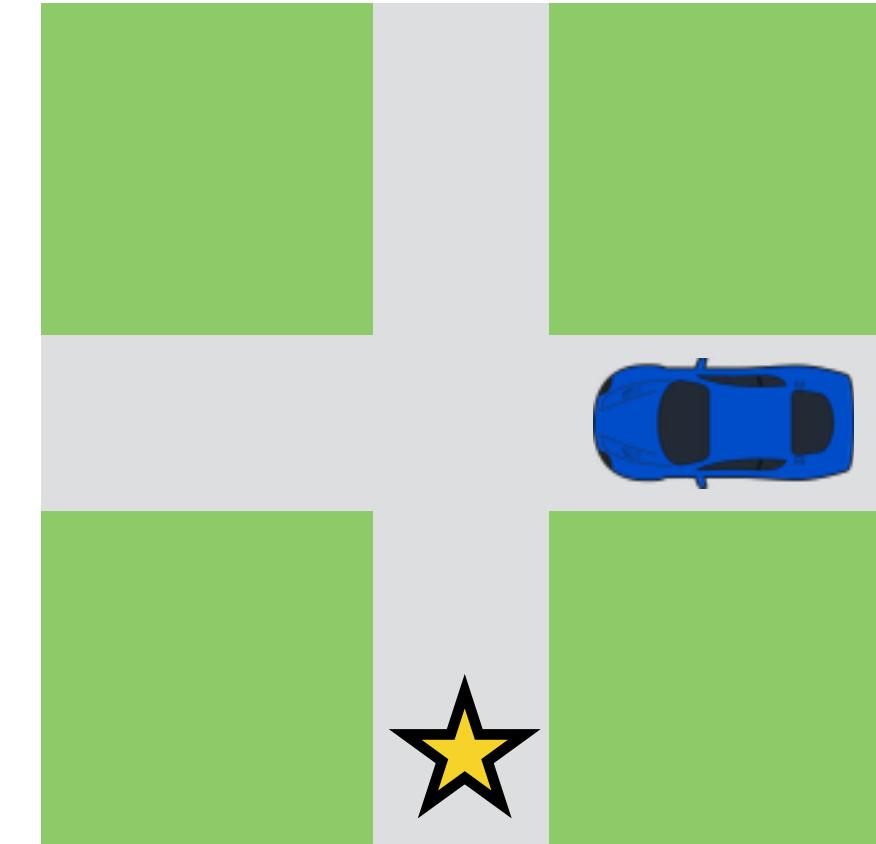
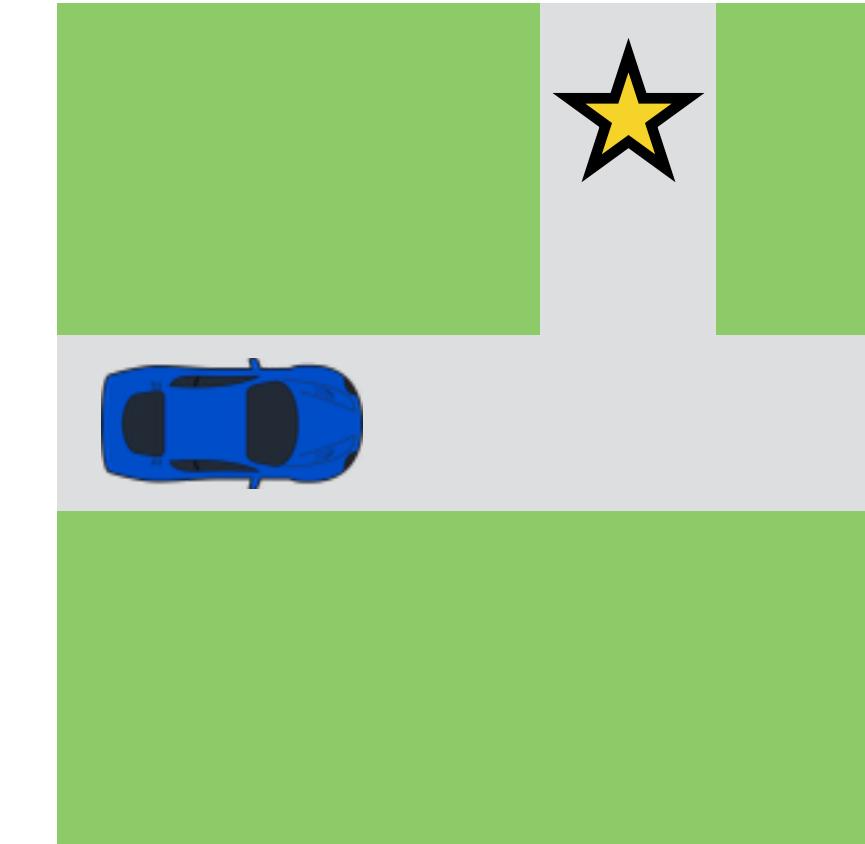
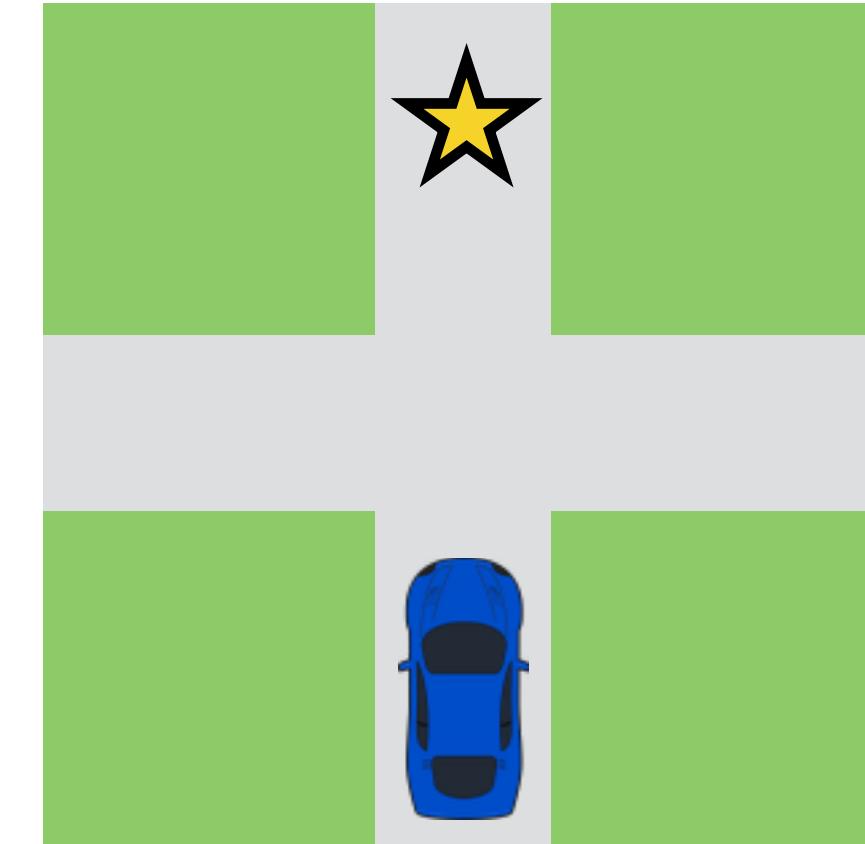




A “semantic MT” problem

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

I'm going
north

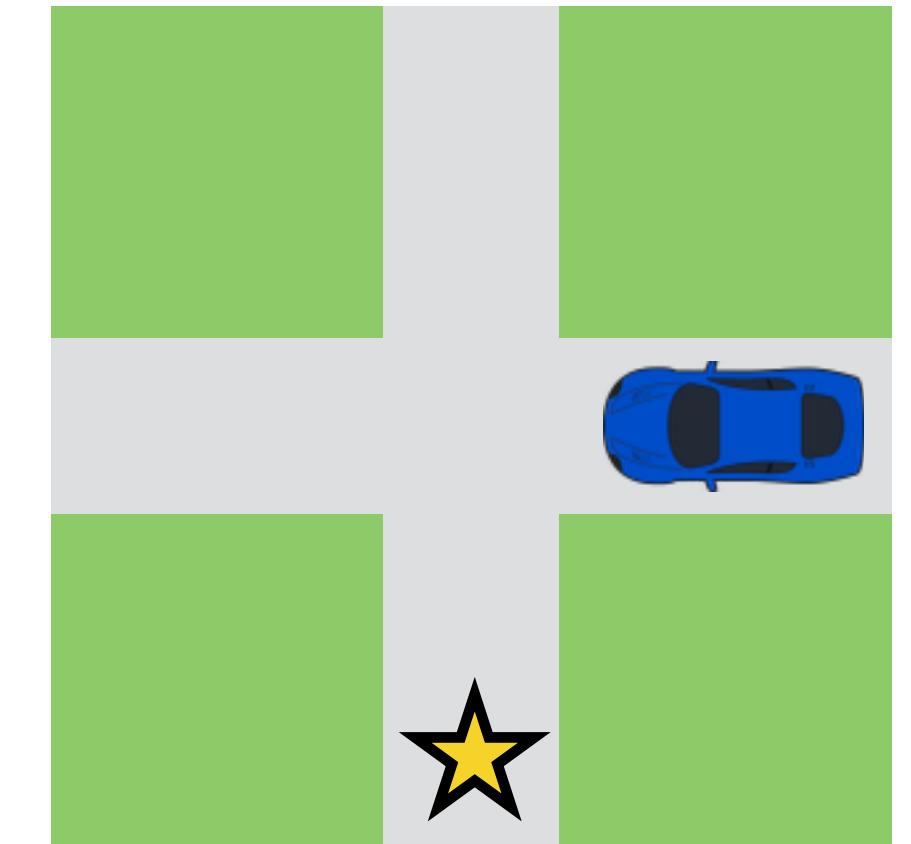
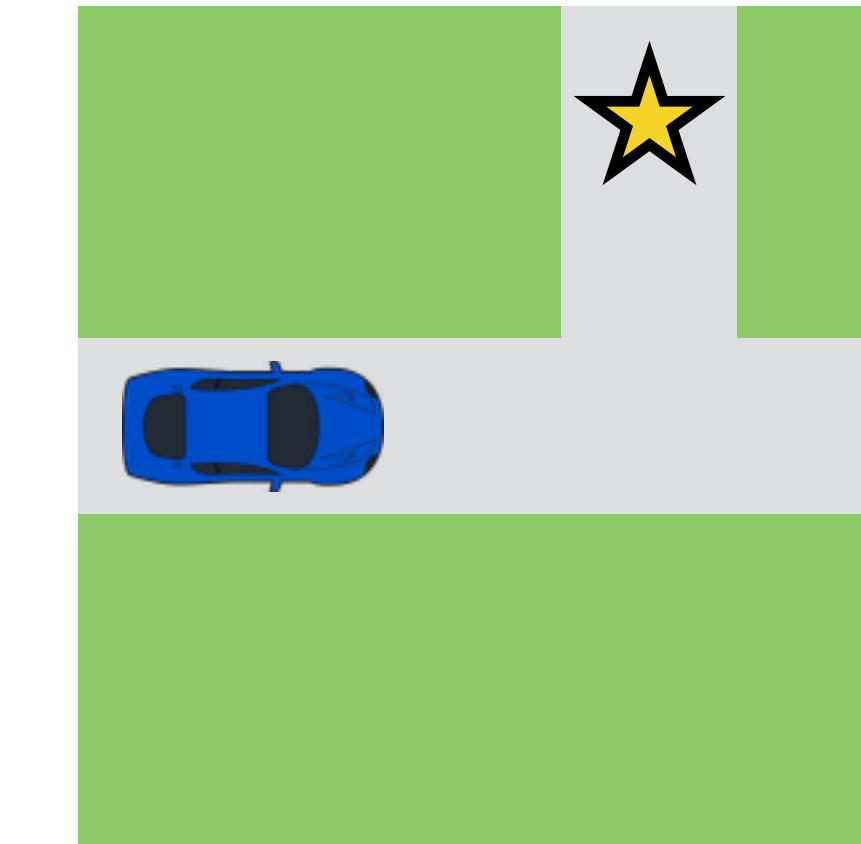
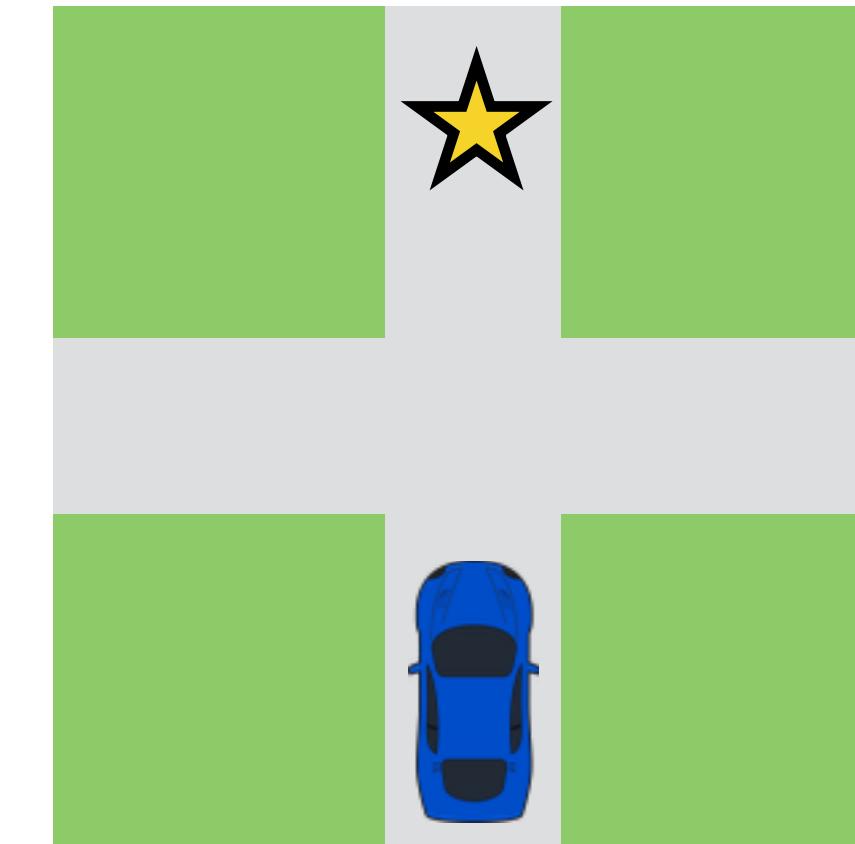




A “semantic MT” problem

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

I'm going
north



(loc (goal blue) north)

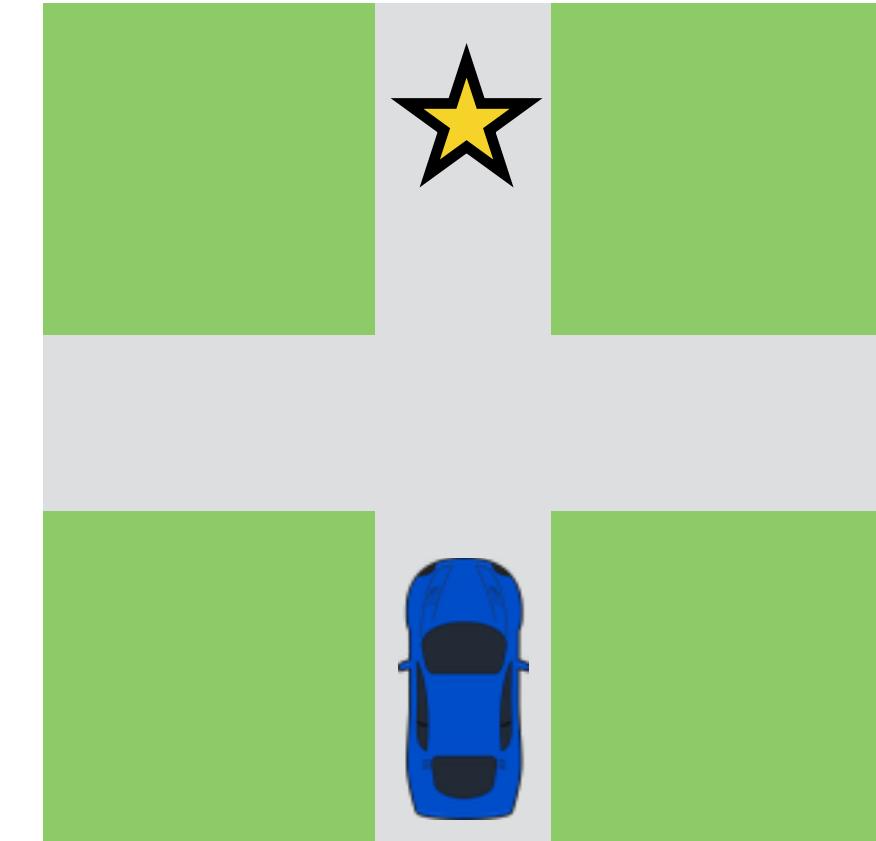


A “semantic MT” problem

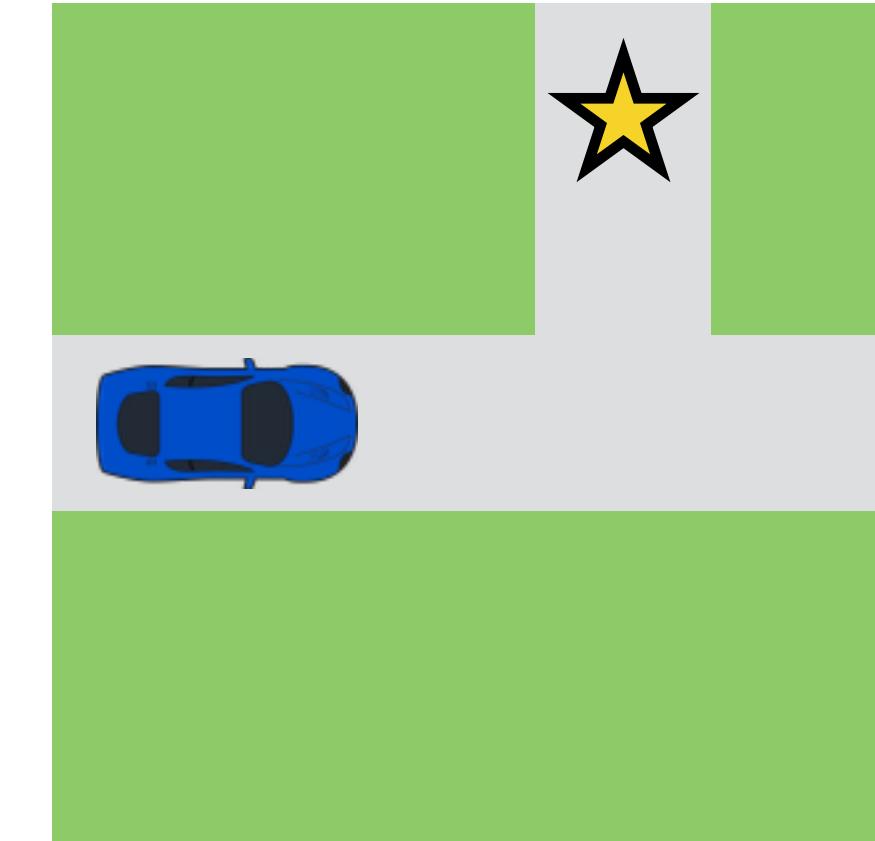
The meaning of an utterance is given by its **truth conditions**

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

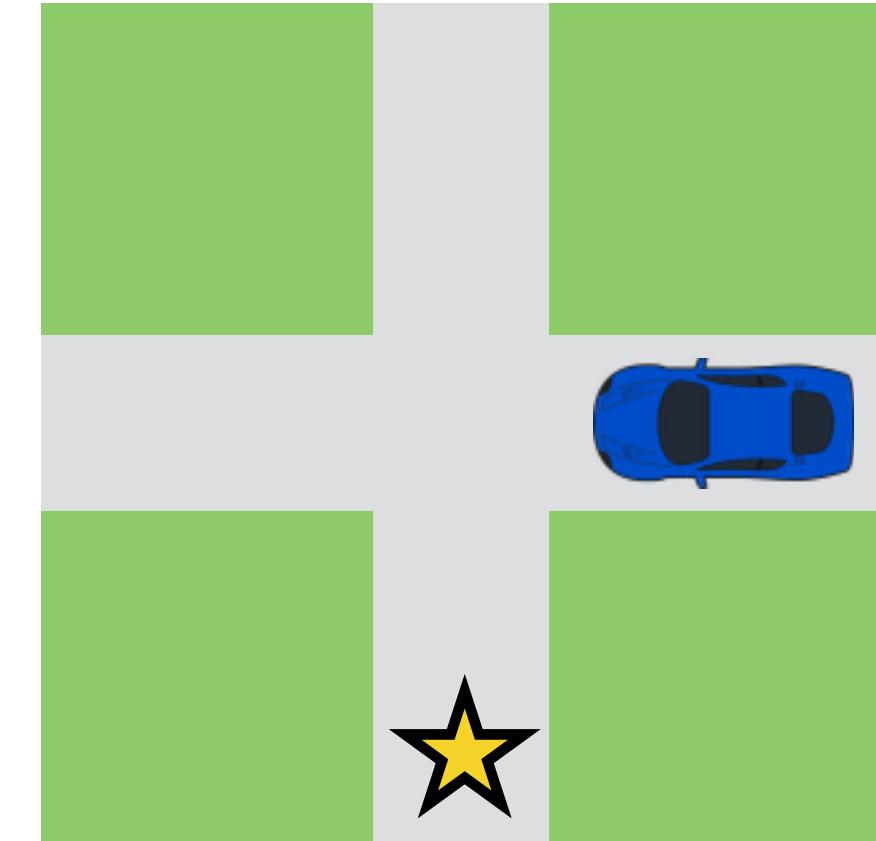
I'm going
north



0.4



0.2



0.001



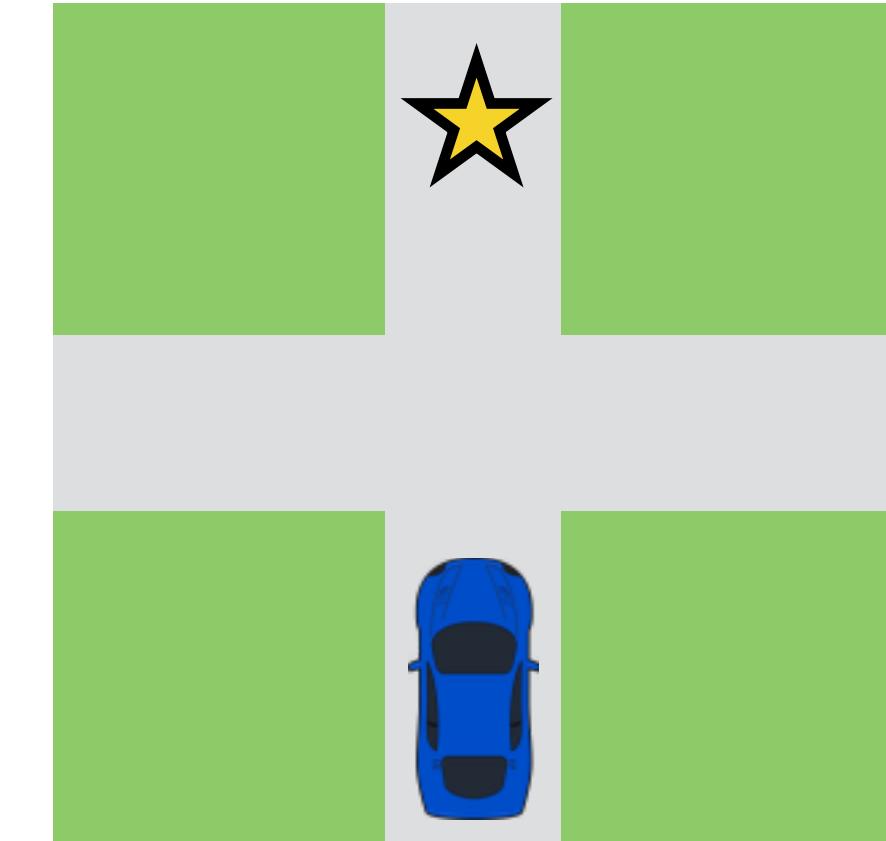
A “semantic MT” problem

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

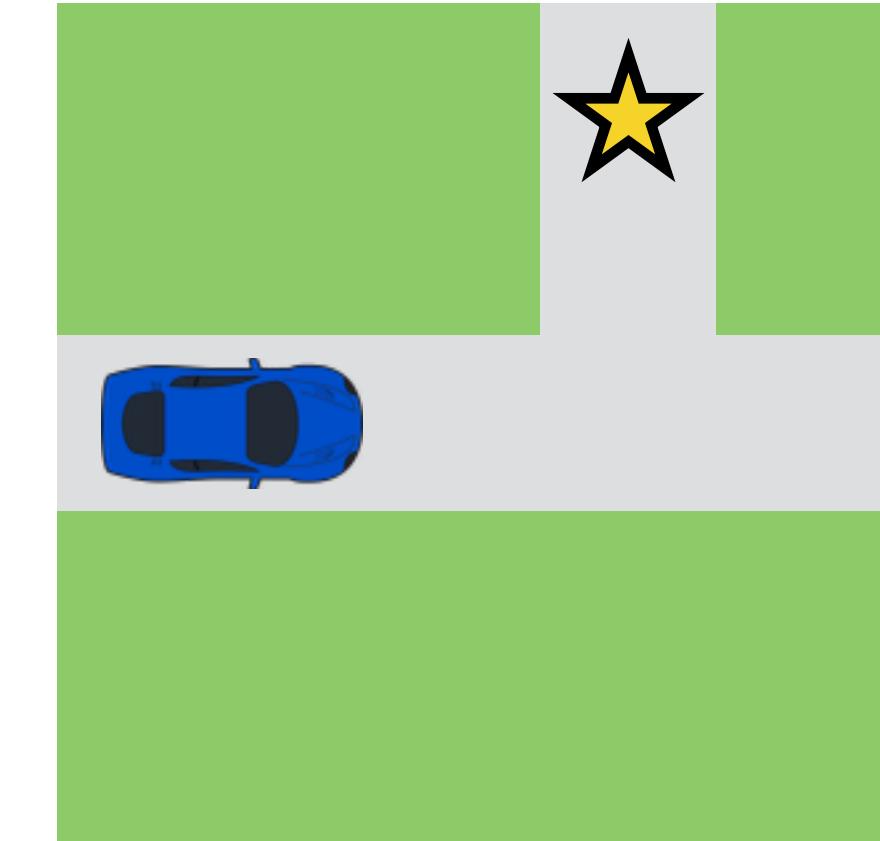
the **distribution over states** in which it is uttered

the **belief** it induces in listeners

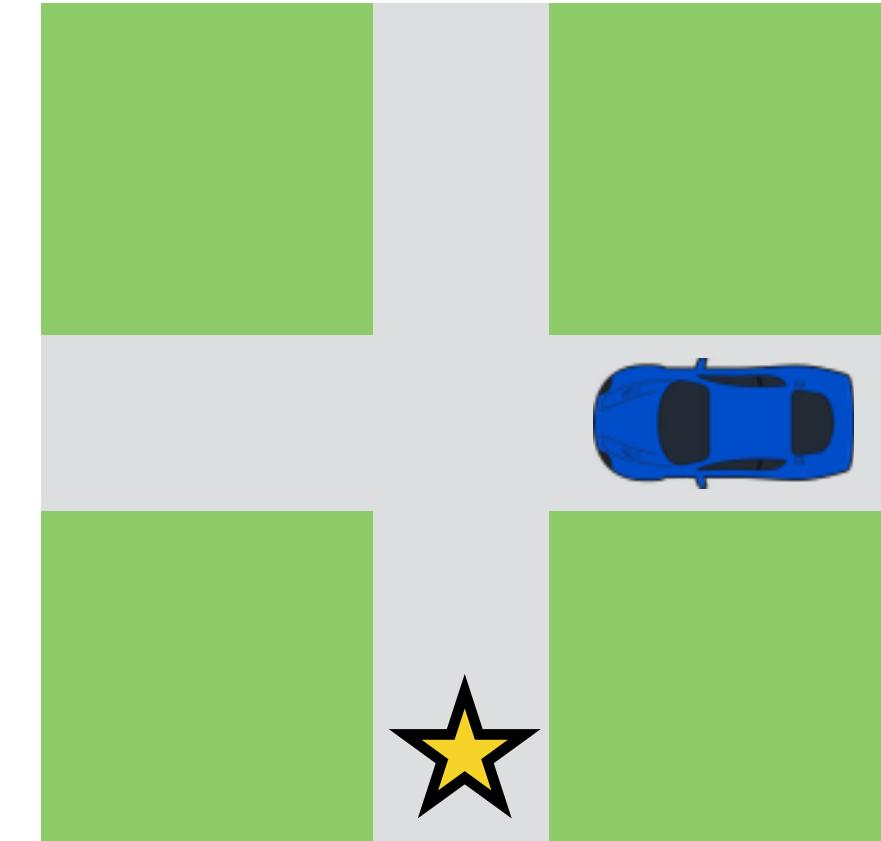
I'm going
north



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

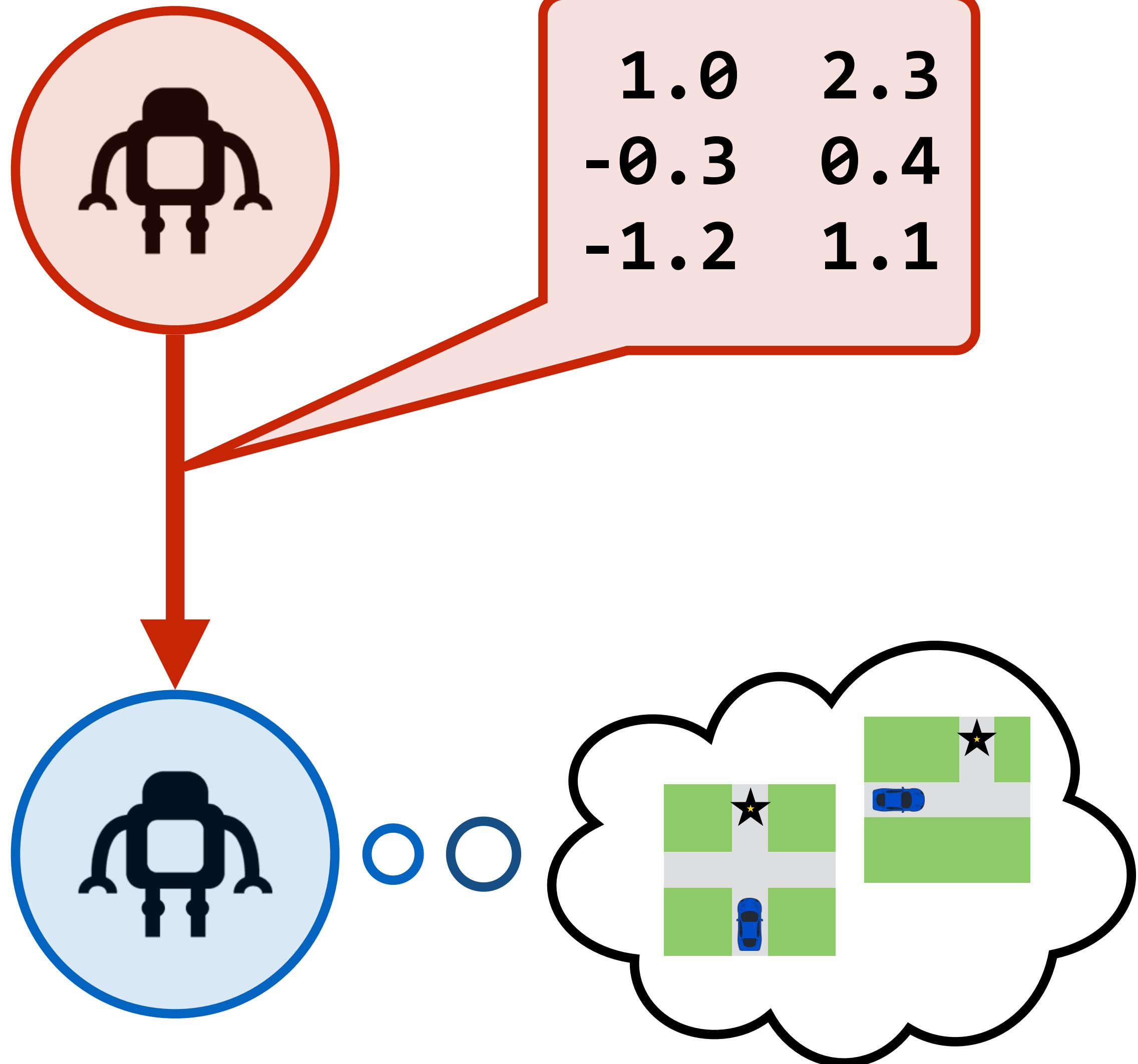
the **distribution over states** in which it is uttered

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.

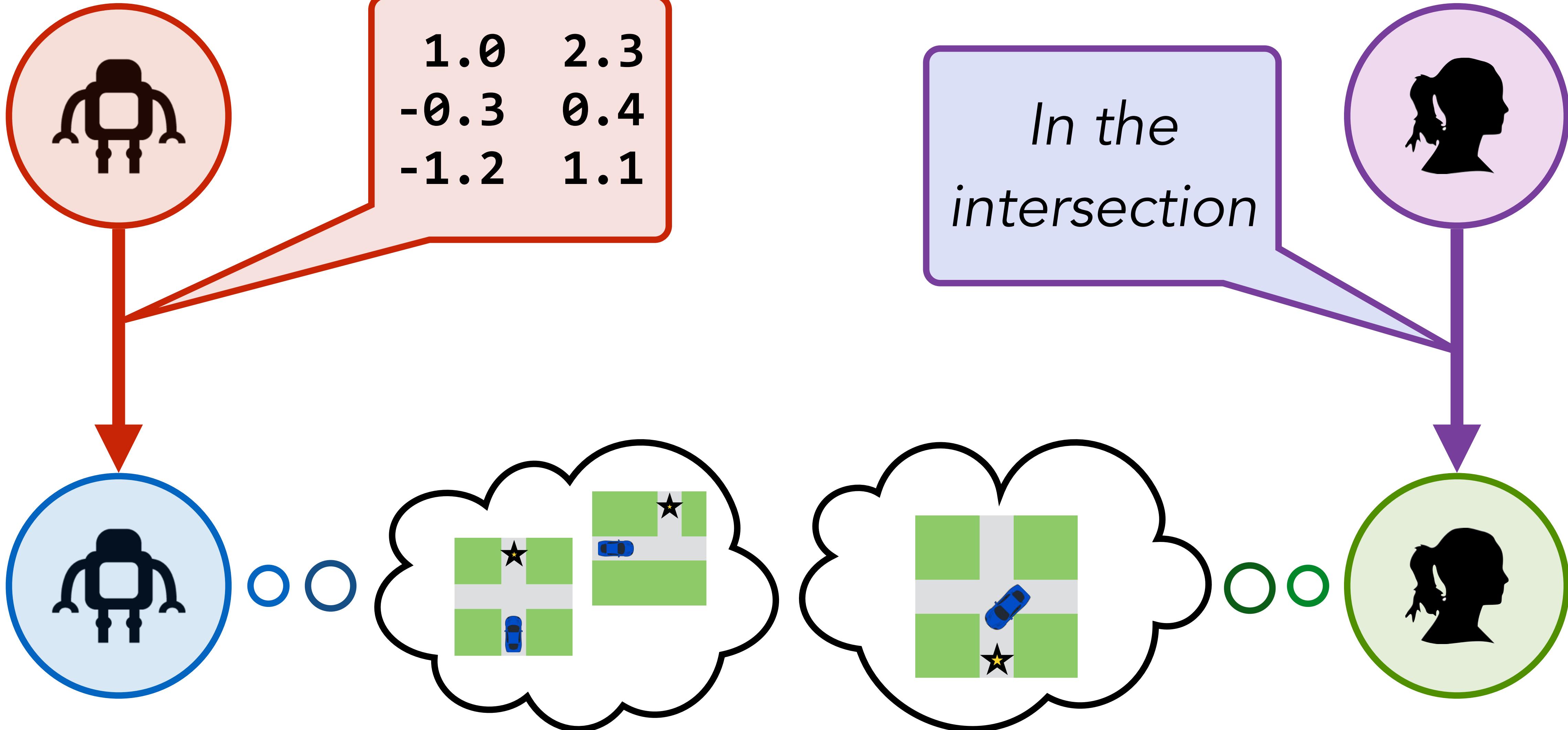


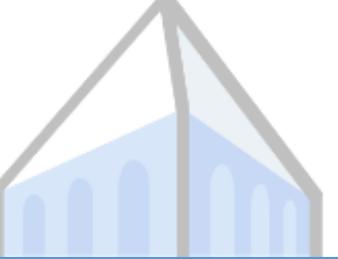
Translating with meaning



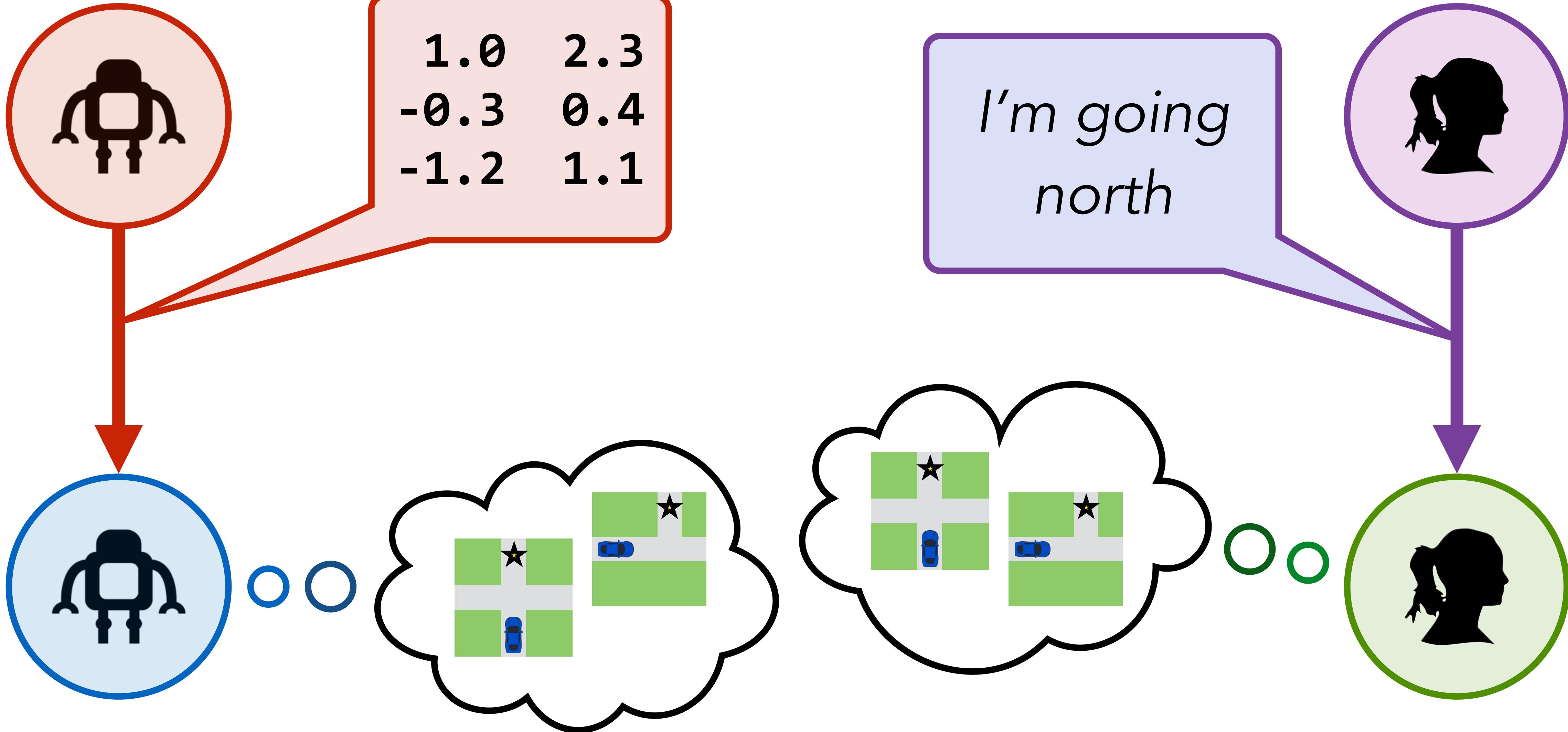


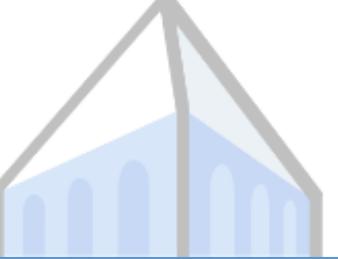
Translating with meaning



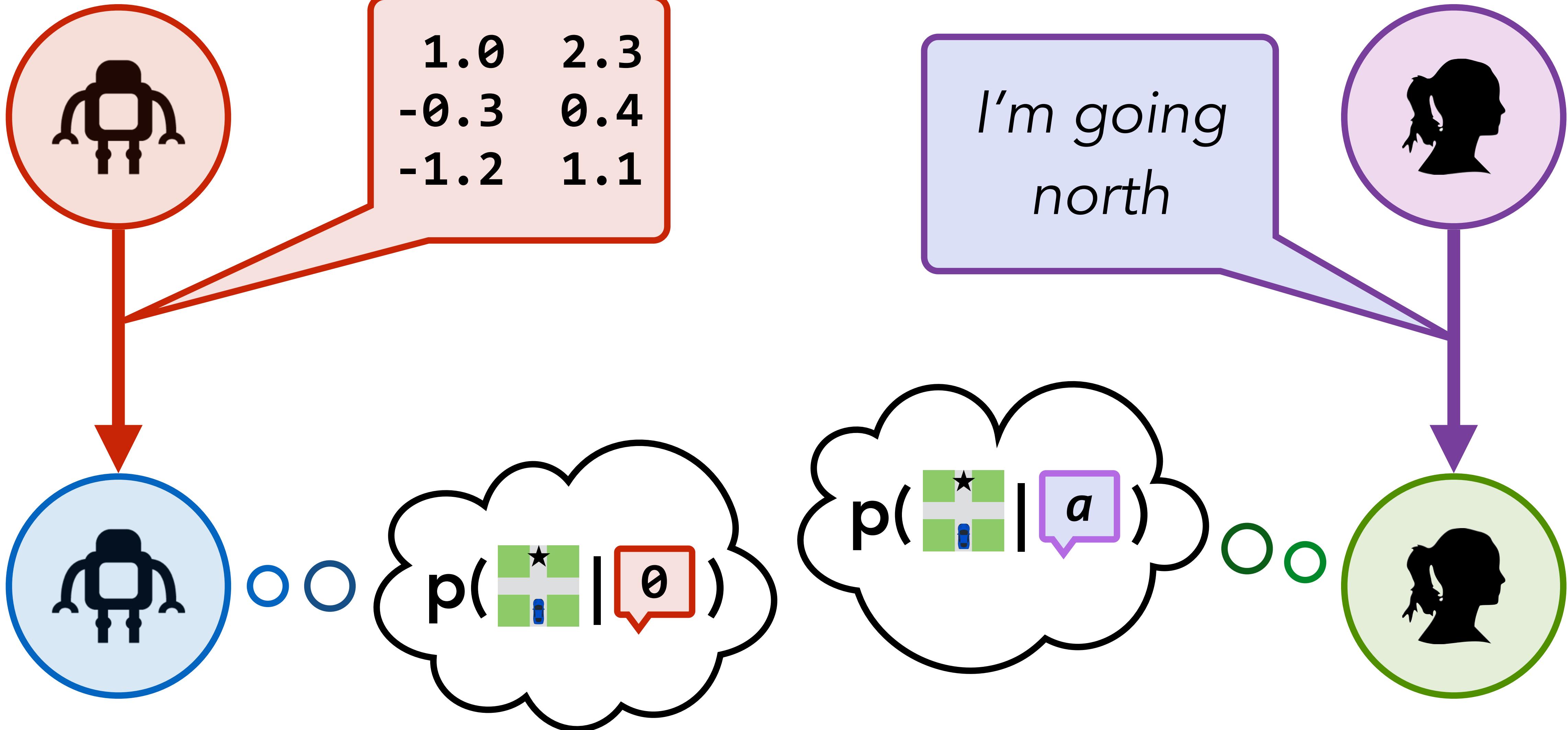


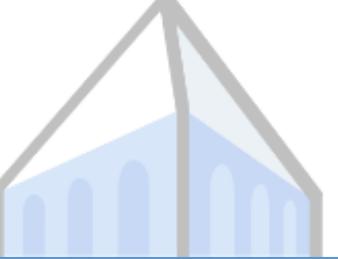
Translating with meaning



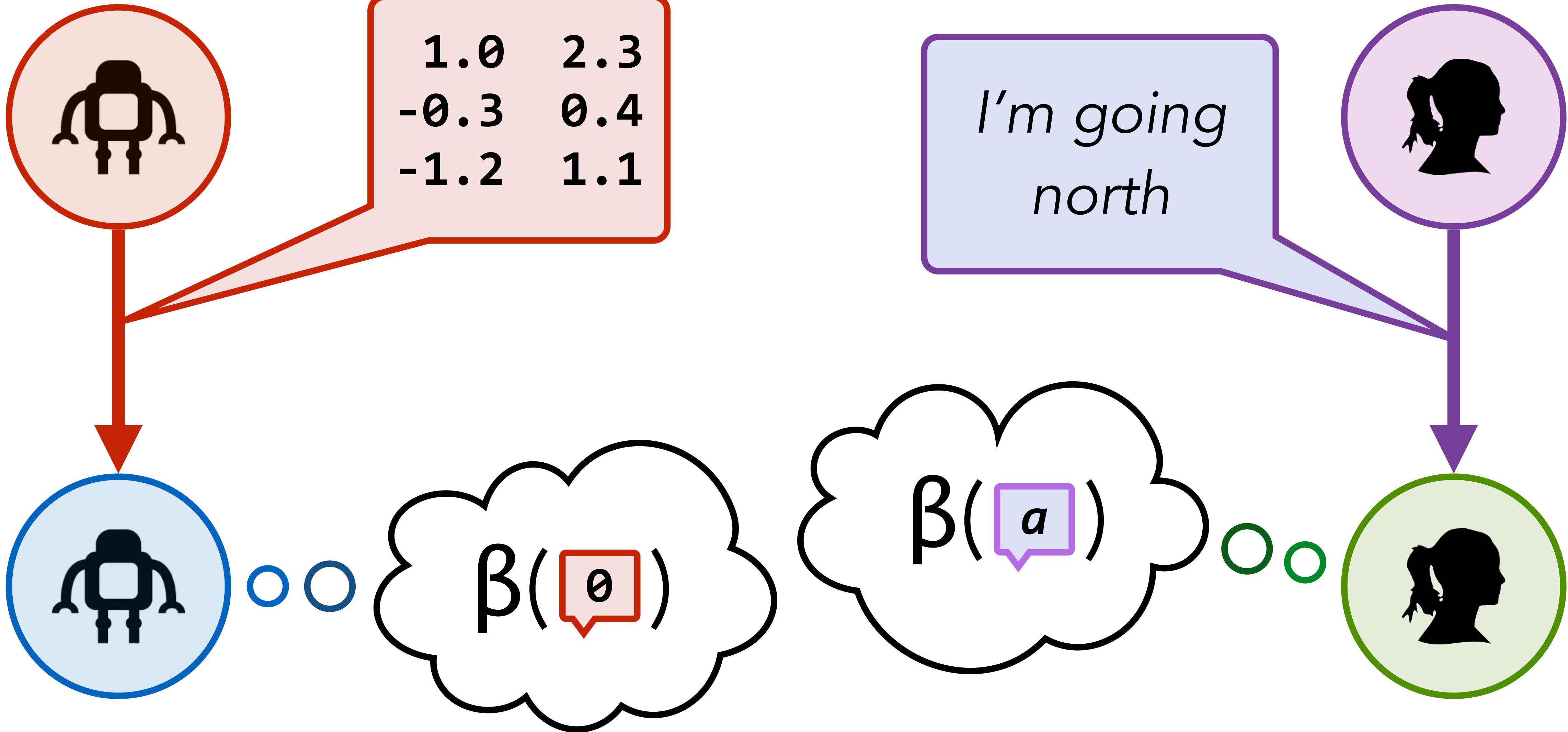


Translating with meaning



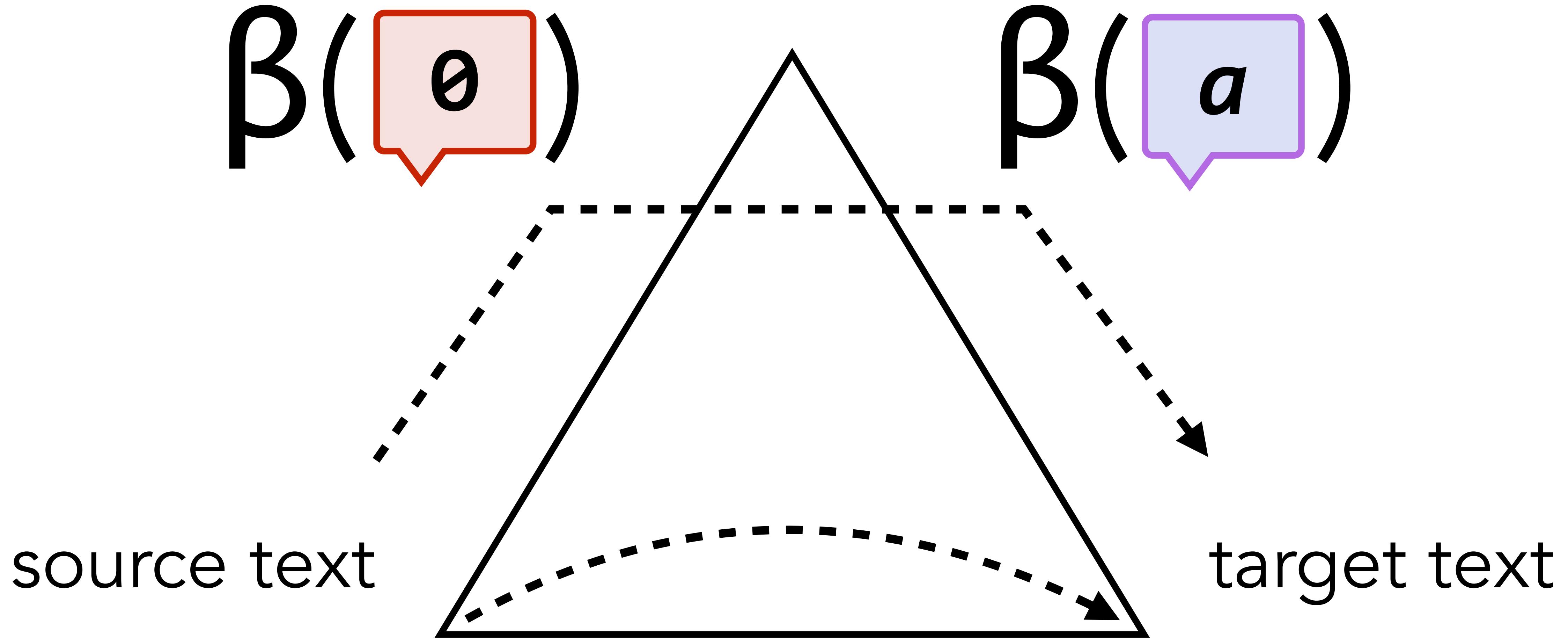


Translating with meaning





Interlingua!





Translation criterion

argmin

a

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



Translation criterion

argmin

a

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



Translation criterion

argmin

a

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



Translation criterion

argmin

a

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



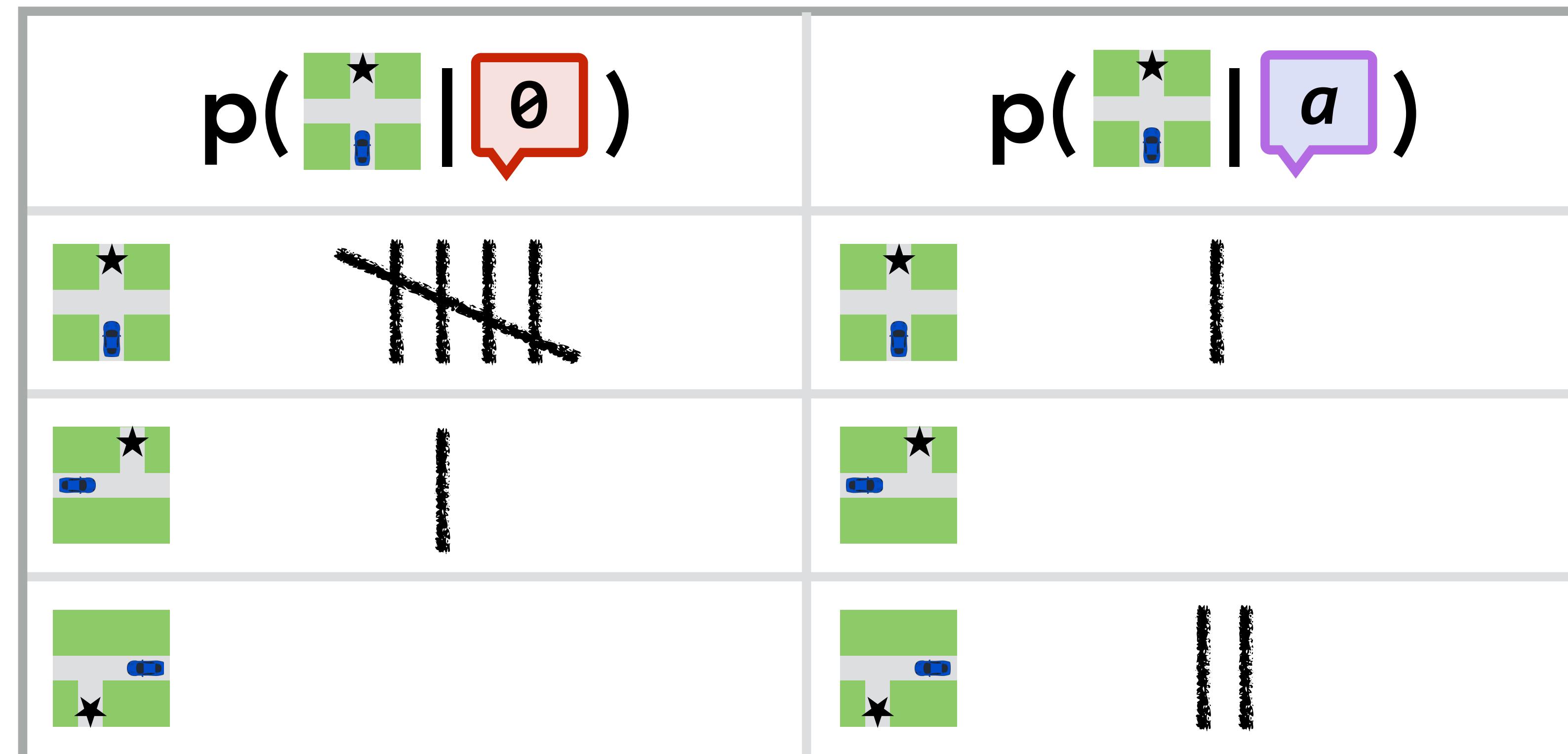
Computing representations

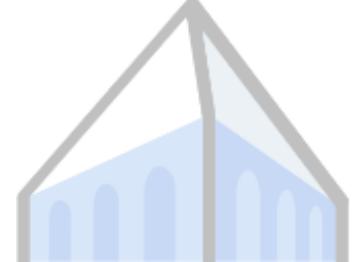
$$\operatorname{argmin}_a \text{KL}(\beta(\theta) \parallel \beta(a))$$



Computing representations: sparsity

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





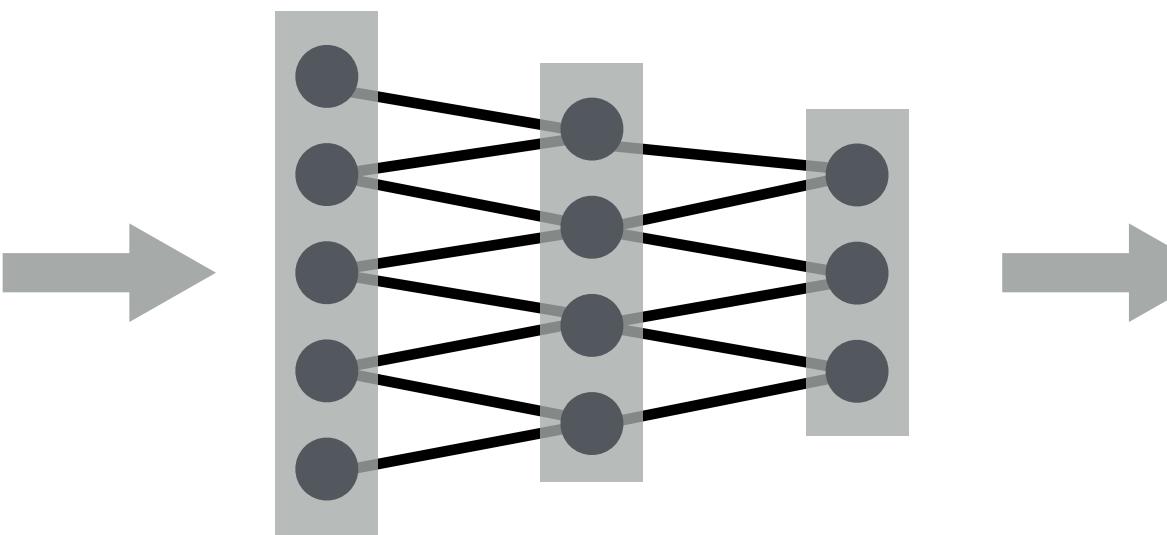
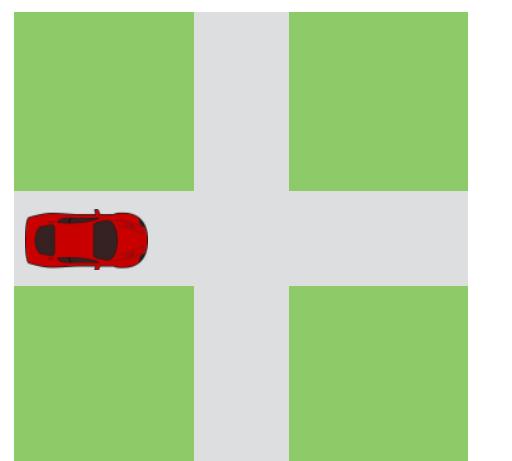
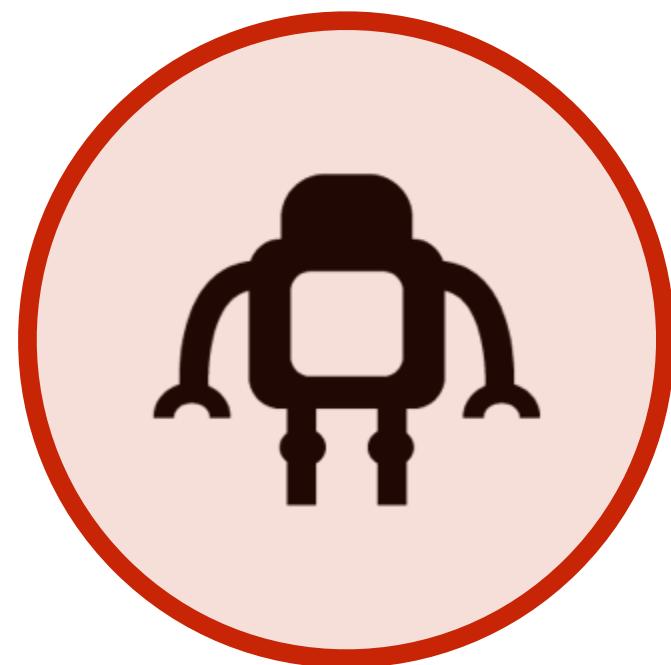
Computing representations: smoothing

argmin

a

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

agent
policy



actions &
messages



Computing representations: smoothing

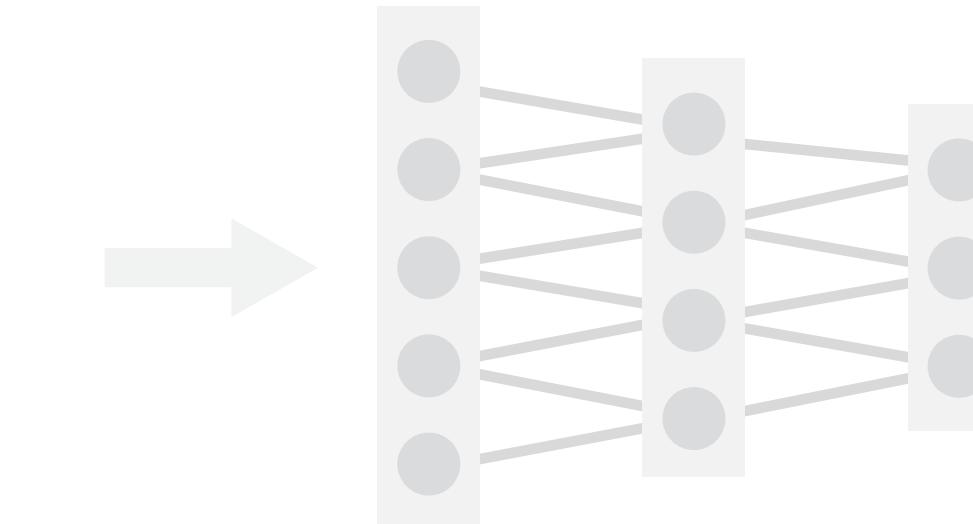
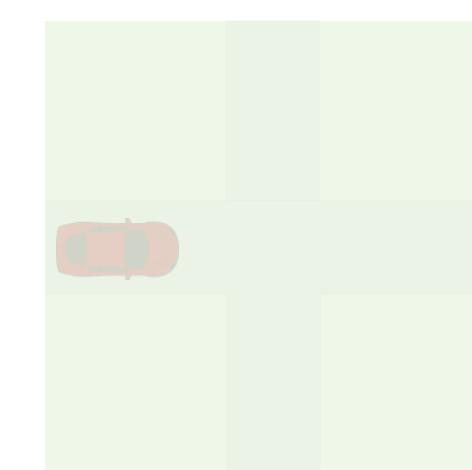
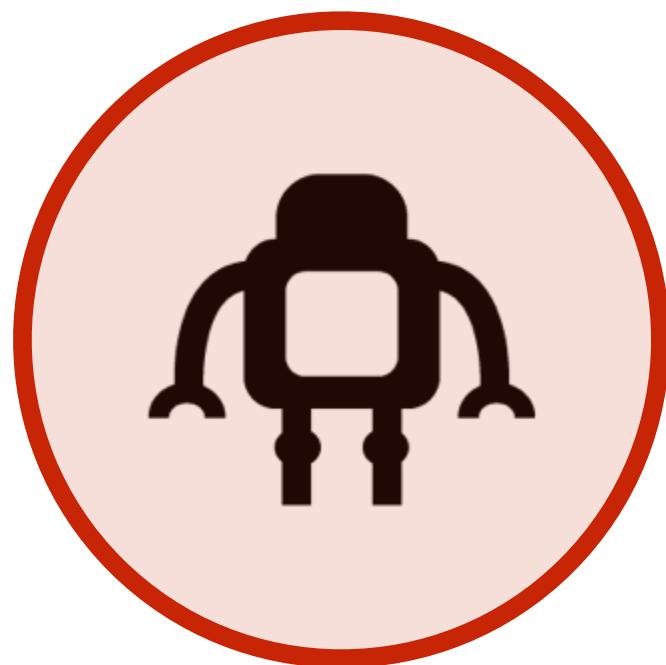
argmin

a

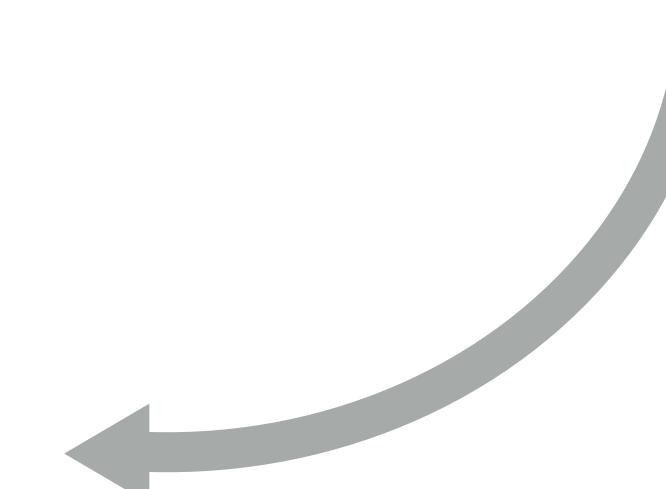
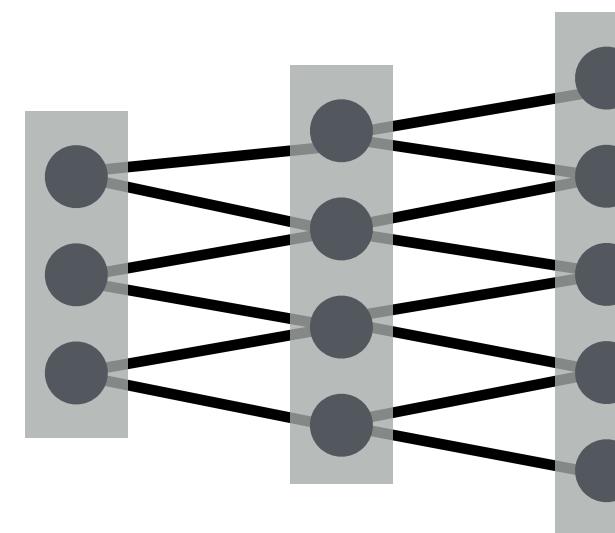
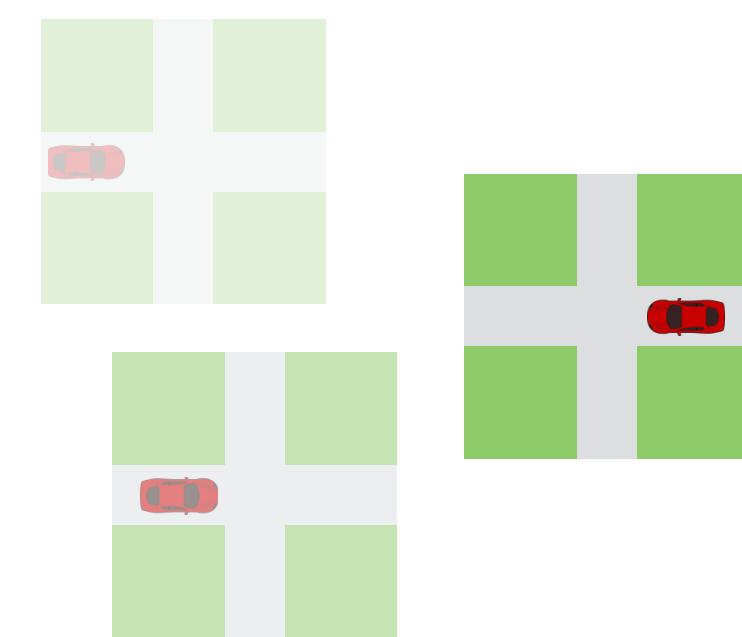
$$\text{KL}(\beta(\theta) \parallel \beta(a))$$

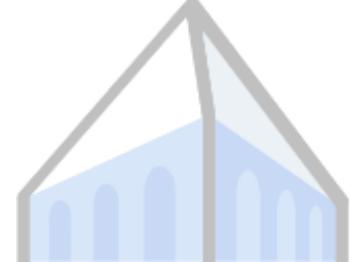
agent
policy

agent
model



actions &
messages





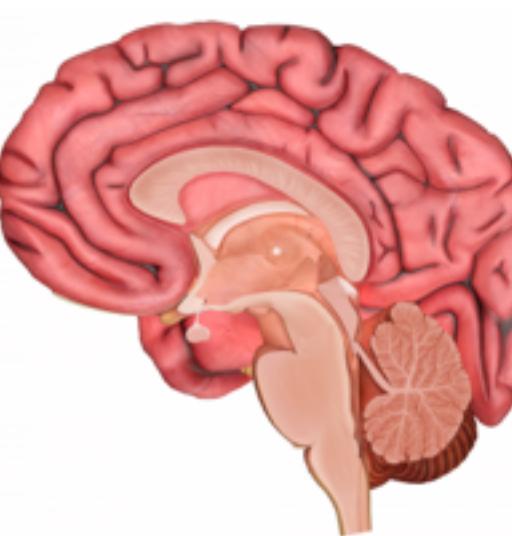
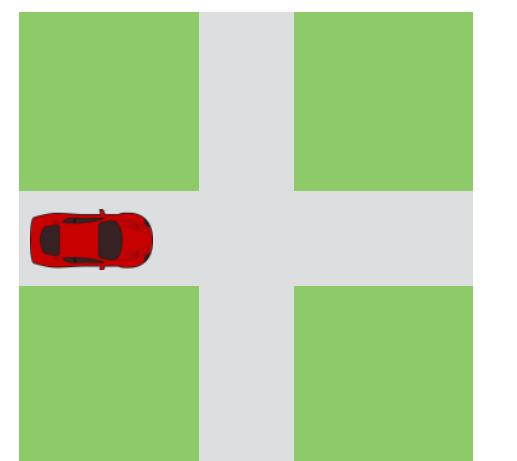
Computing representations: smoothing

argmin

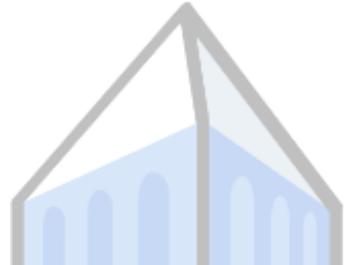
a

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

human



actions &
messages



Computing representations: smoothing

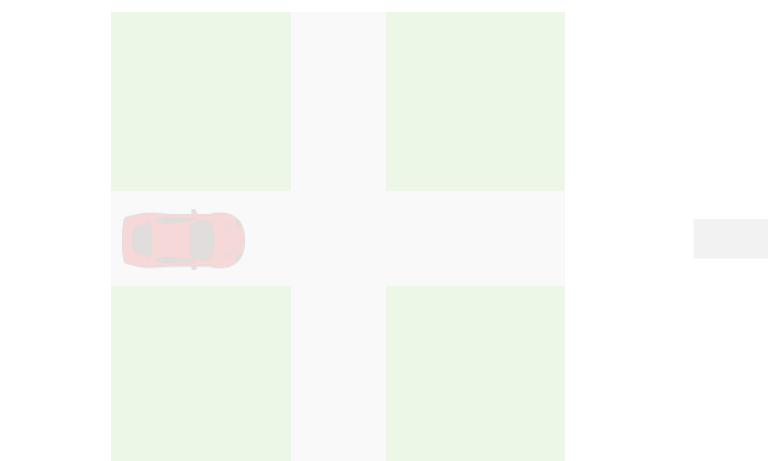
argmin

a

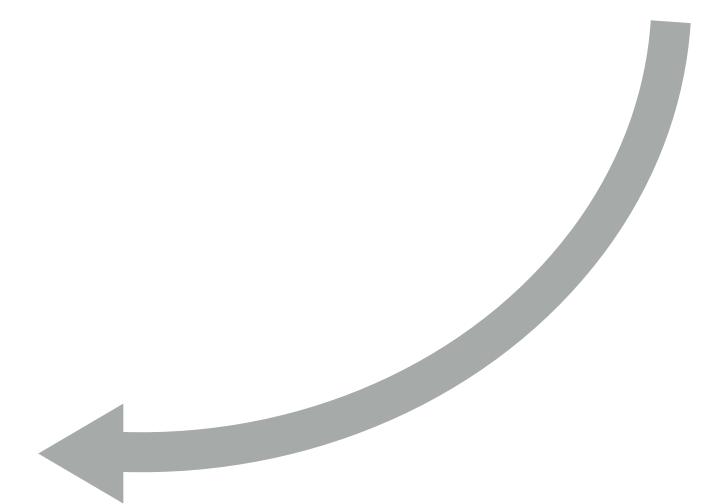
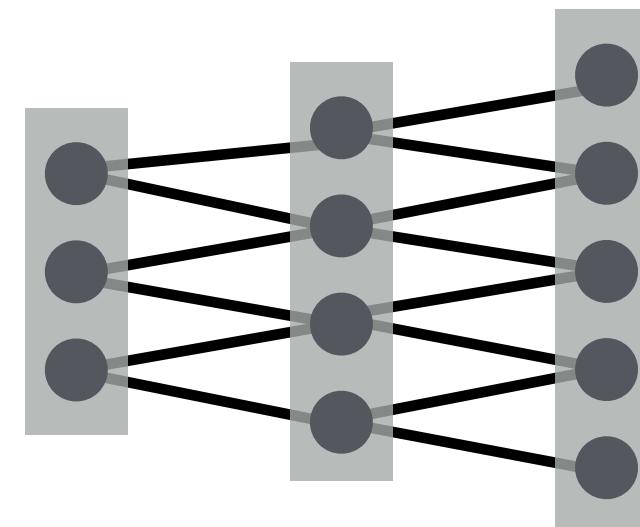
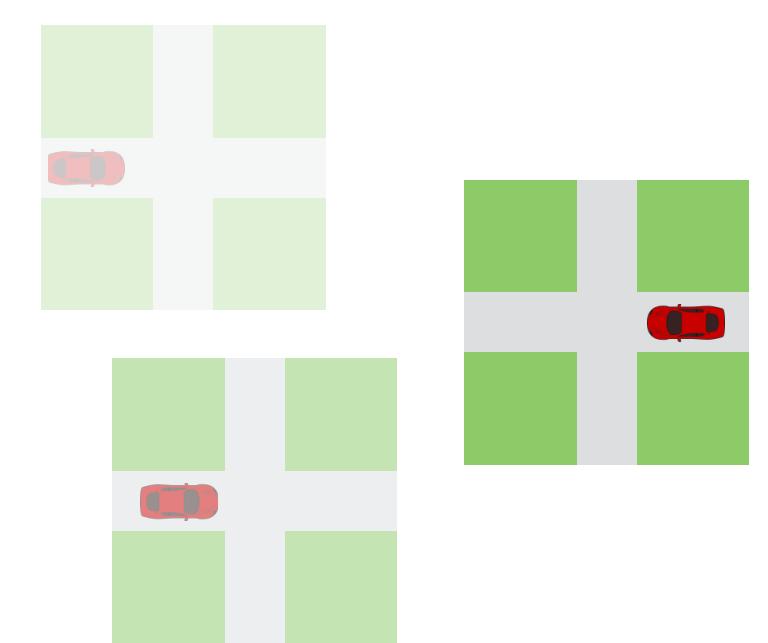
$$\text{KL}(\beta(\theta) \parallel \beta(a))$$

human
policy

human
model



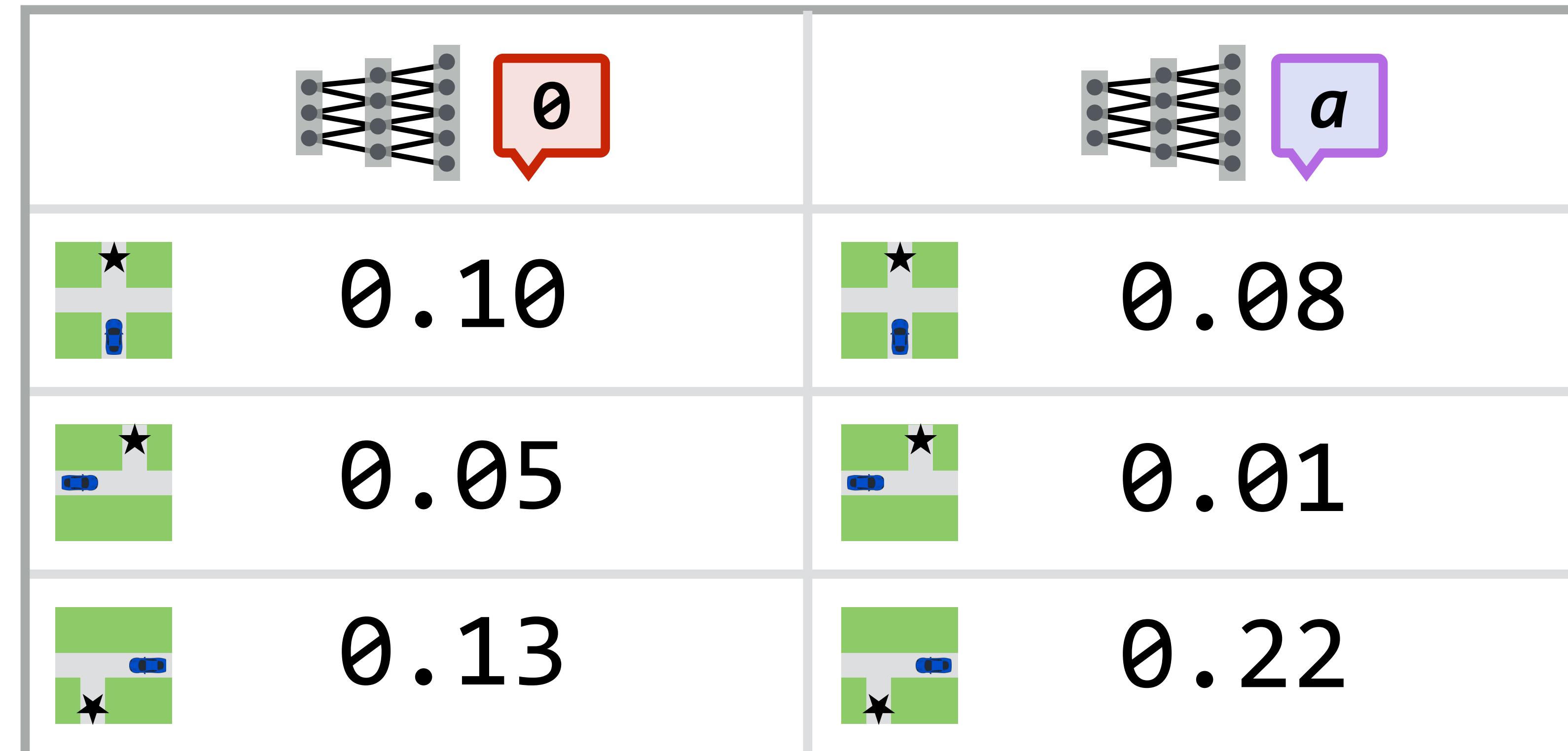
actions &
messages





Computing representations: smoothing

argmin_{*a*} KL($\beta(\theta)$ || $\beta(a)$)





Computing KL

argmin

a

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



Computing KL

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

$$\text{KL}(p \parallel q) = \mathbb{E}_p \frac{p(\text{car})}{q(\text{car})}$$



Computing KL: sampling

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

$$\text{KL}(p \parallel q) = \sum_i p(\star_i) \log \frac{p(\star_i)}{q(\star_i)}$$



Finding translations

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



Finding translations: brute force

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

| | | |
|----------------------------------|---|-----|
| <i>going north</i> | → | 0.5 |
| <i>crossing the intersection</i> | → | 2.3 |
| <i>I'm done</i> | → | 0.2 |
| <i>after you</i> | → | 9.7 |



Finding translations: brute force

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

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

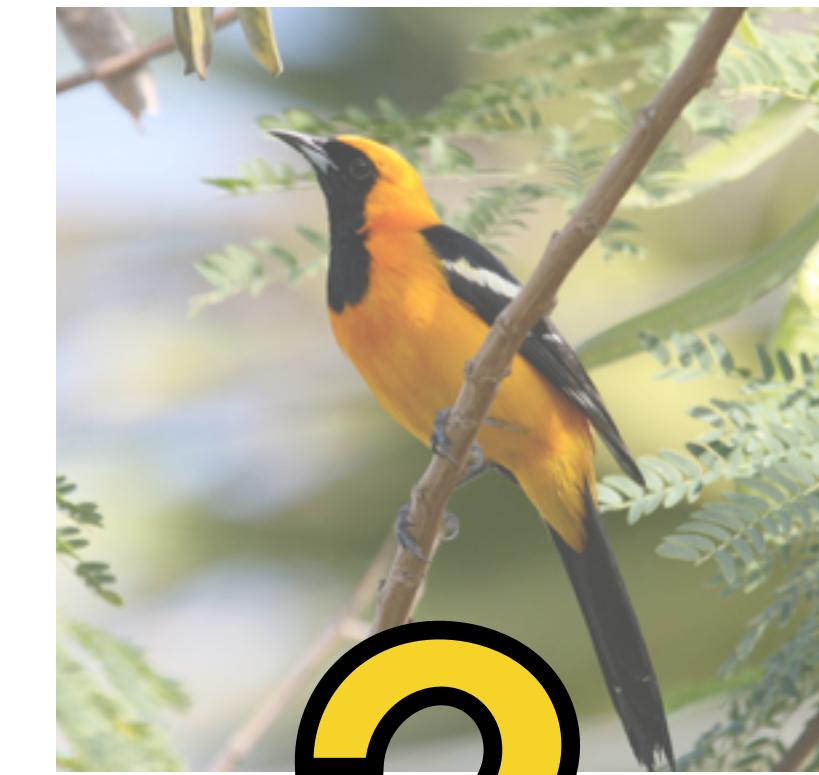
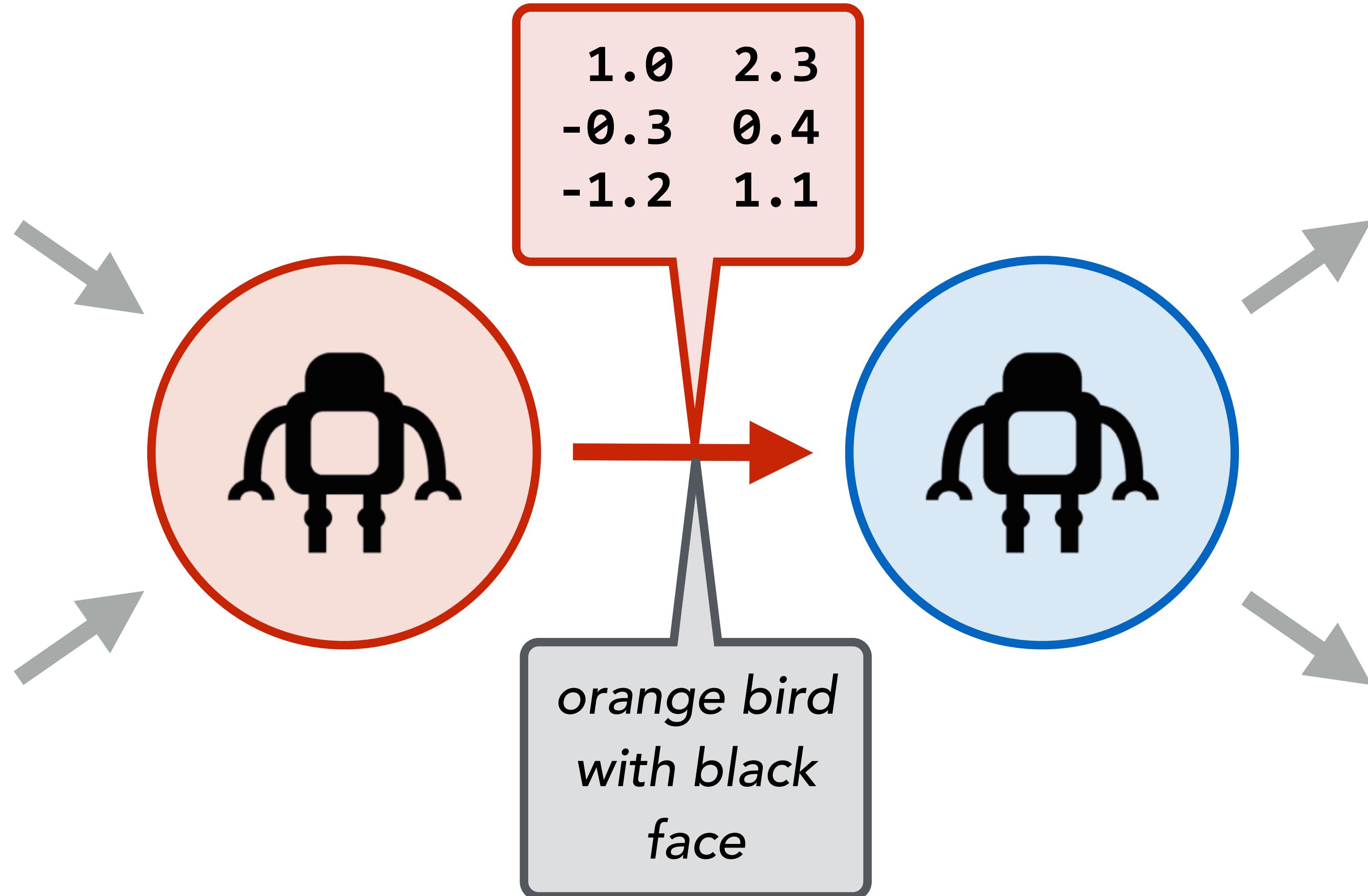
Semantic machine translation

Implementation details

Evaluation

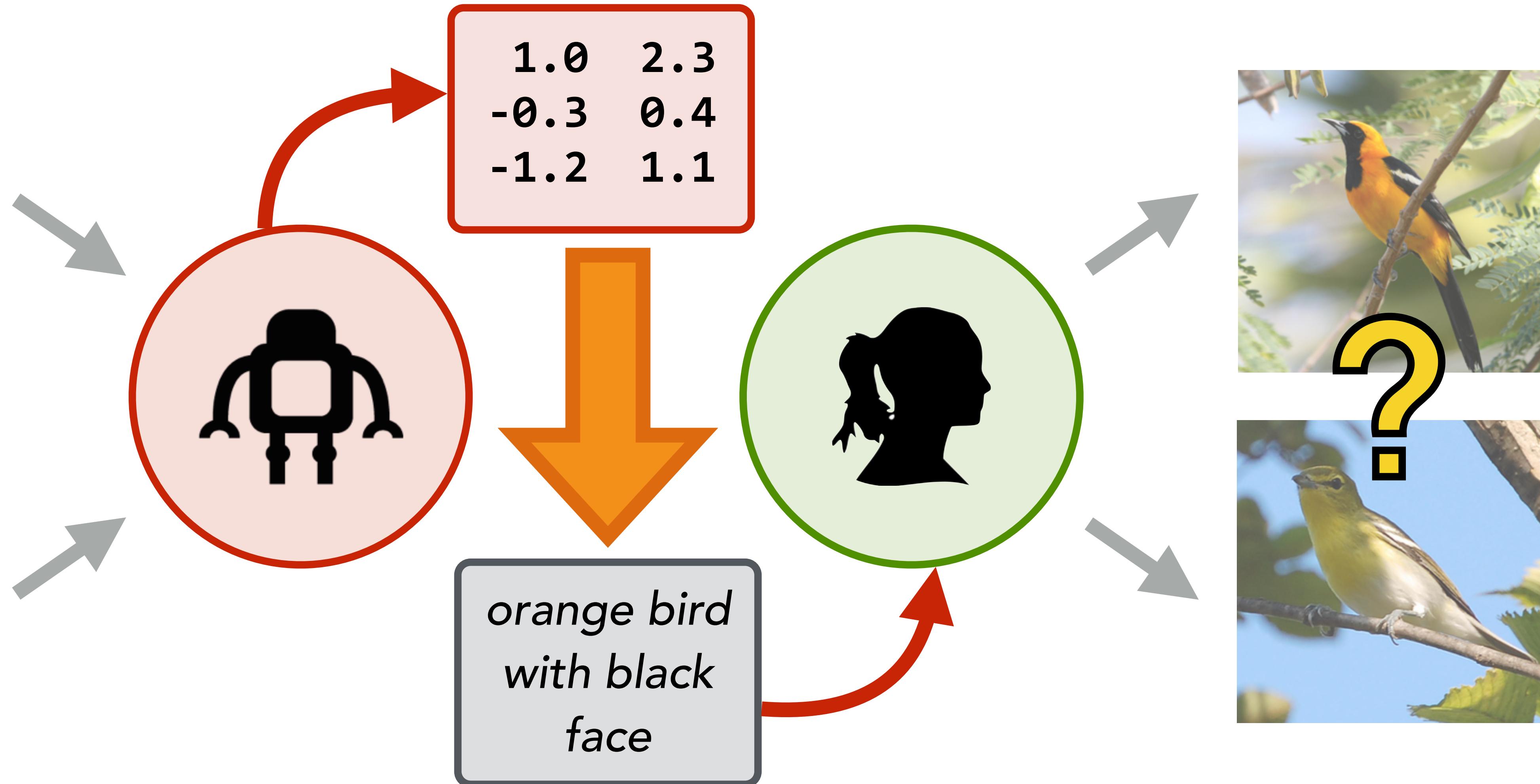


Referring expression games



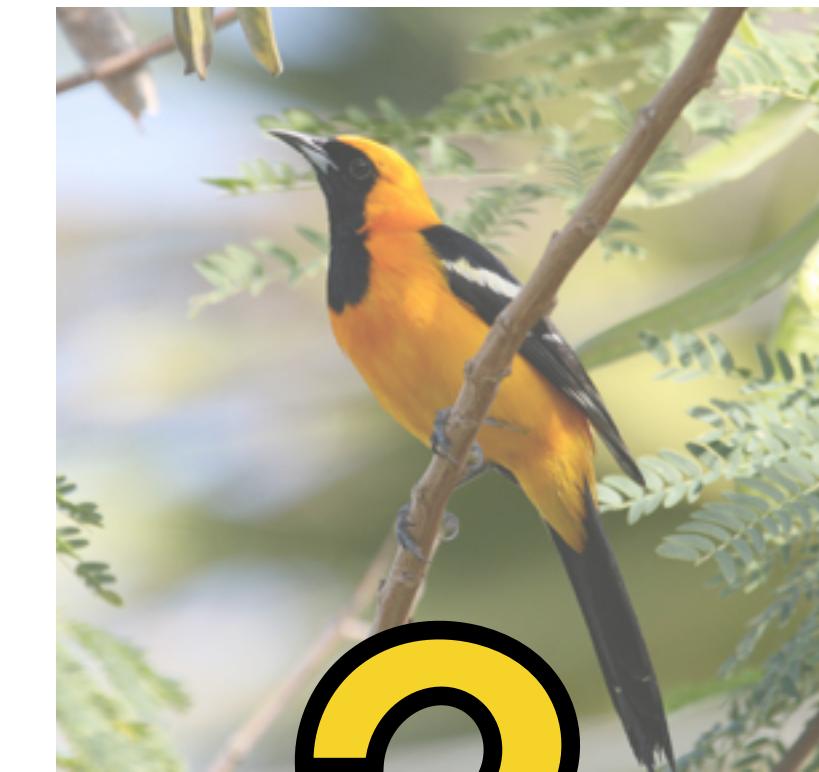
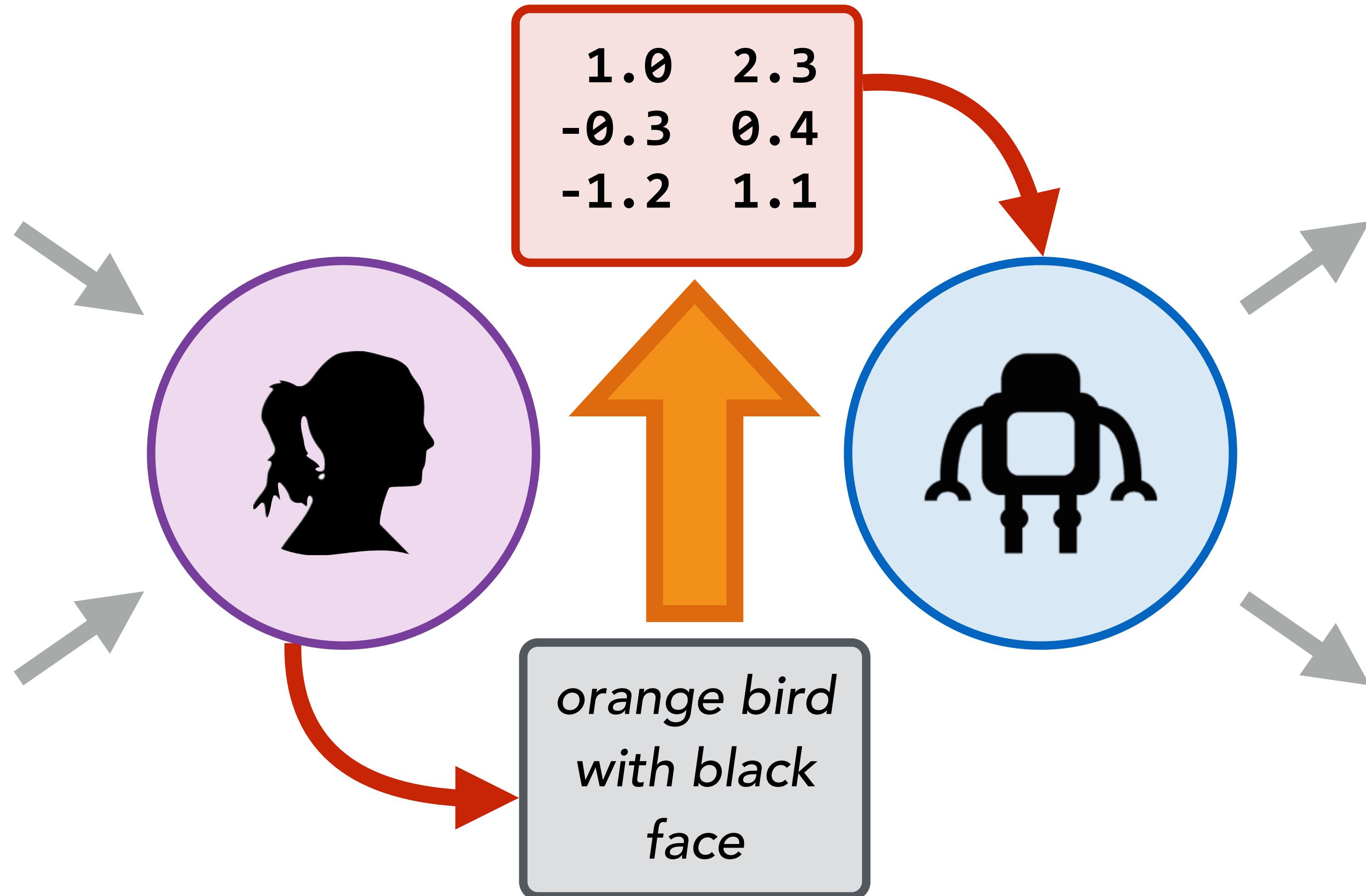


Evaluation: translator-in-the-loop



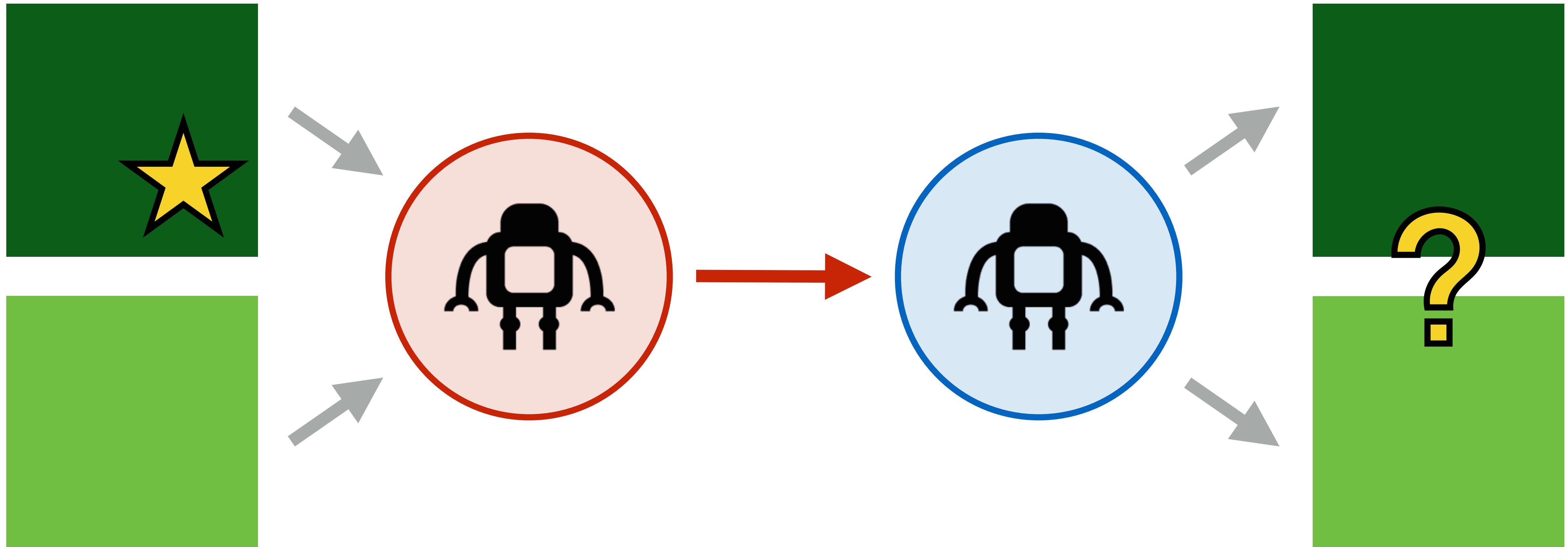


Evaluation: translator-in-the-loop



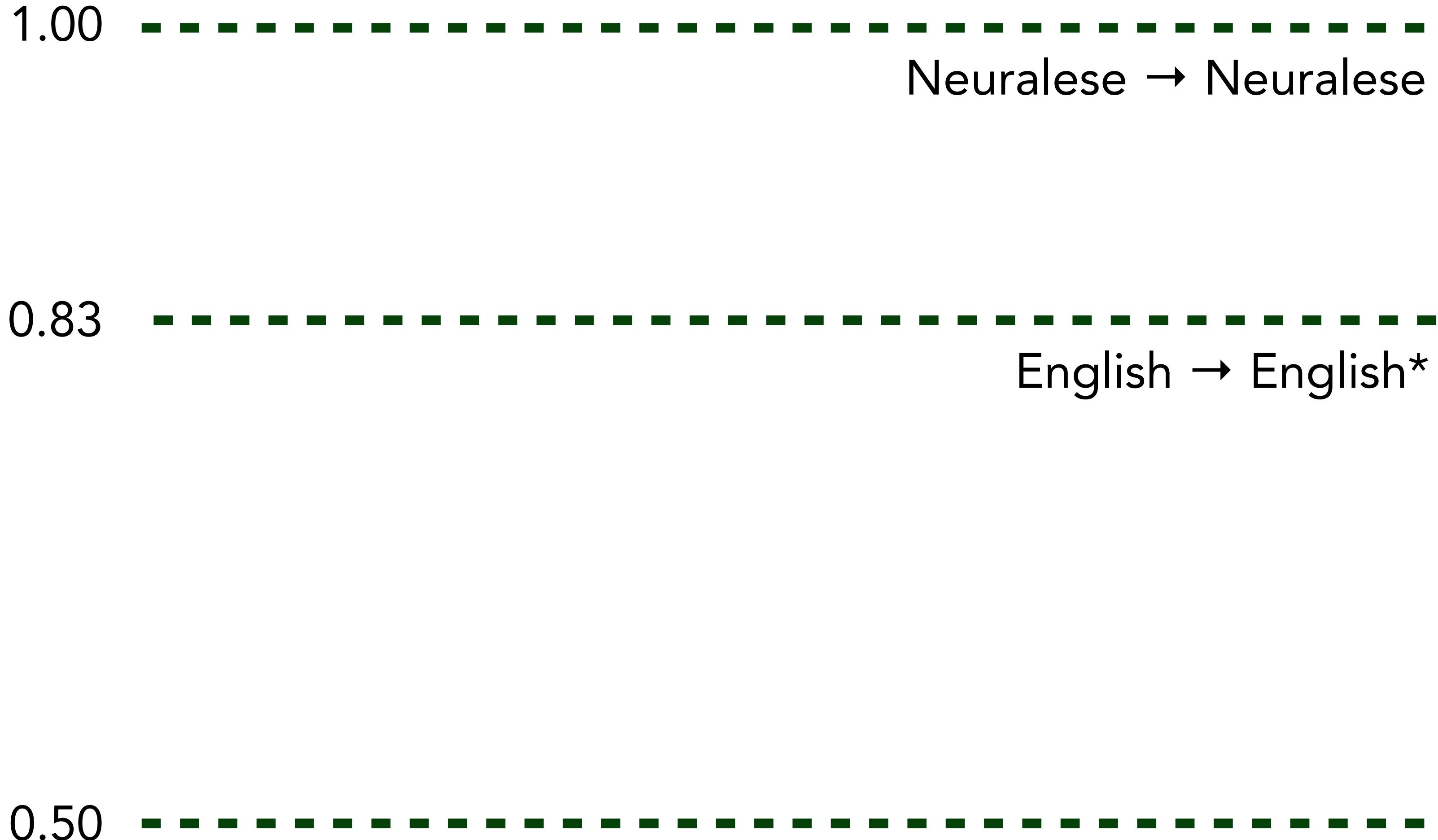


Experiment: color references



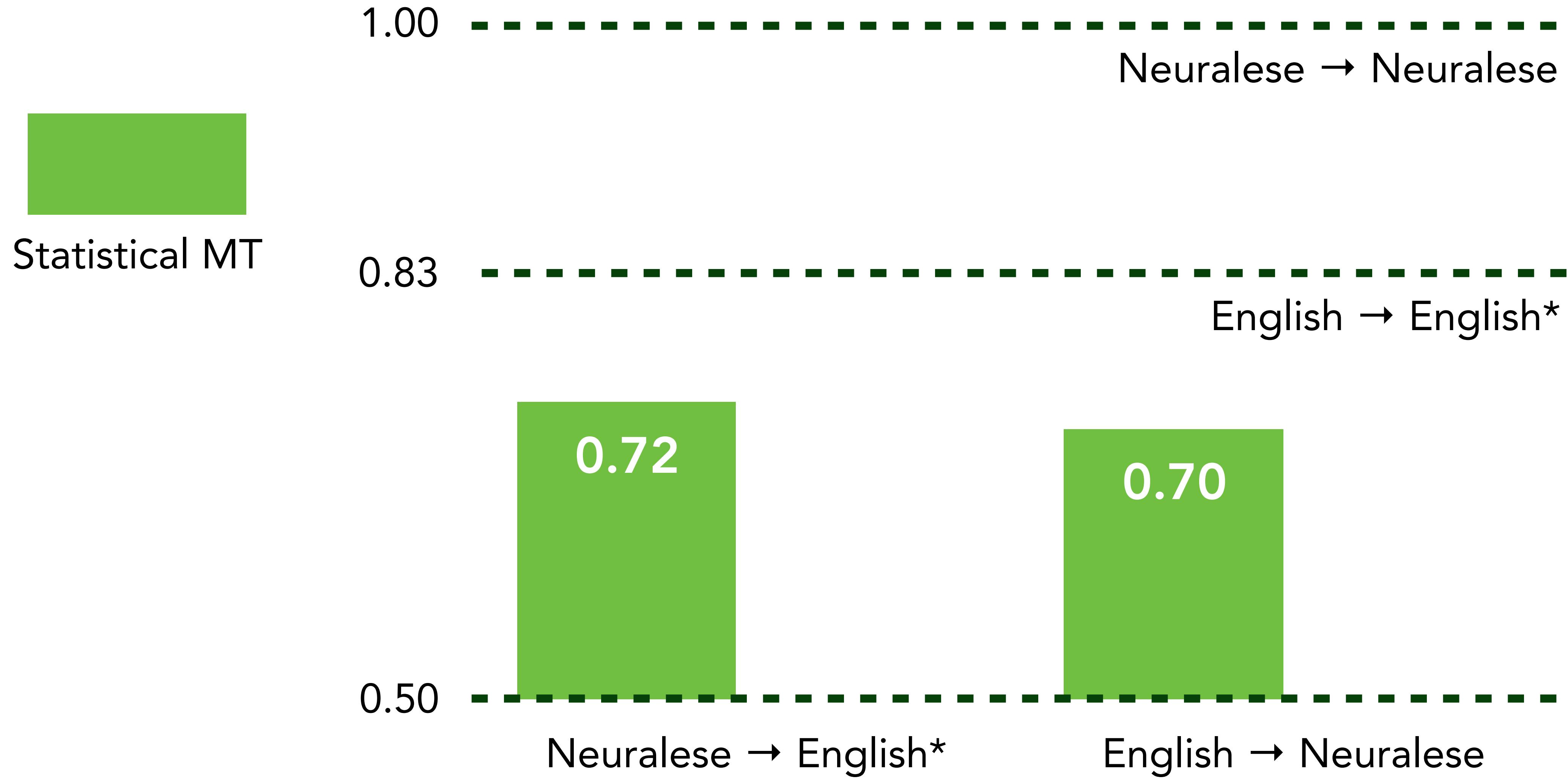


Experiment: color references



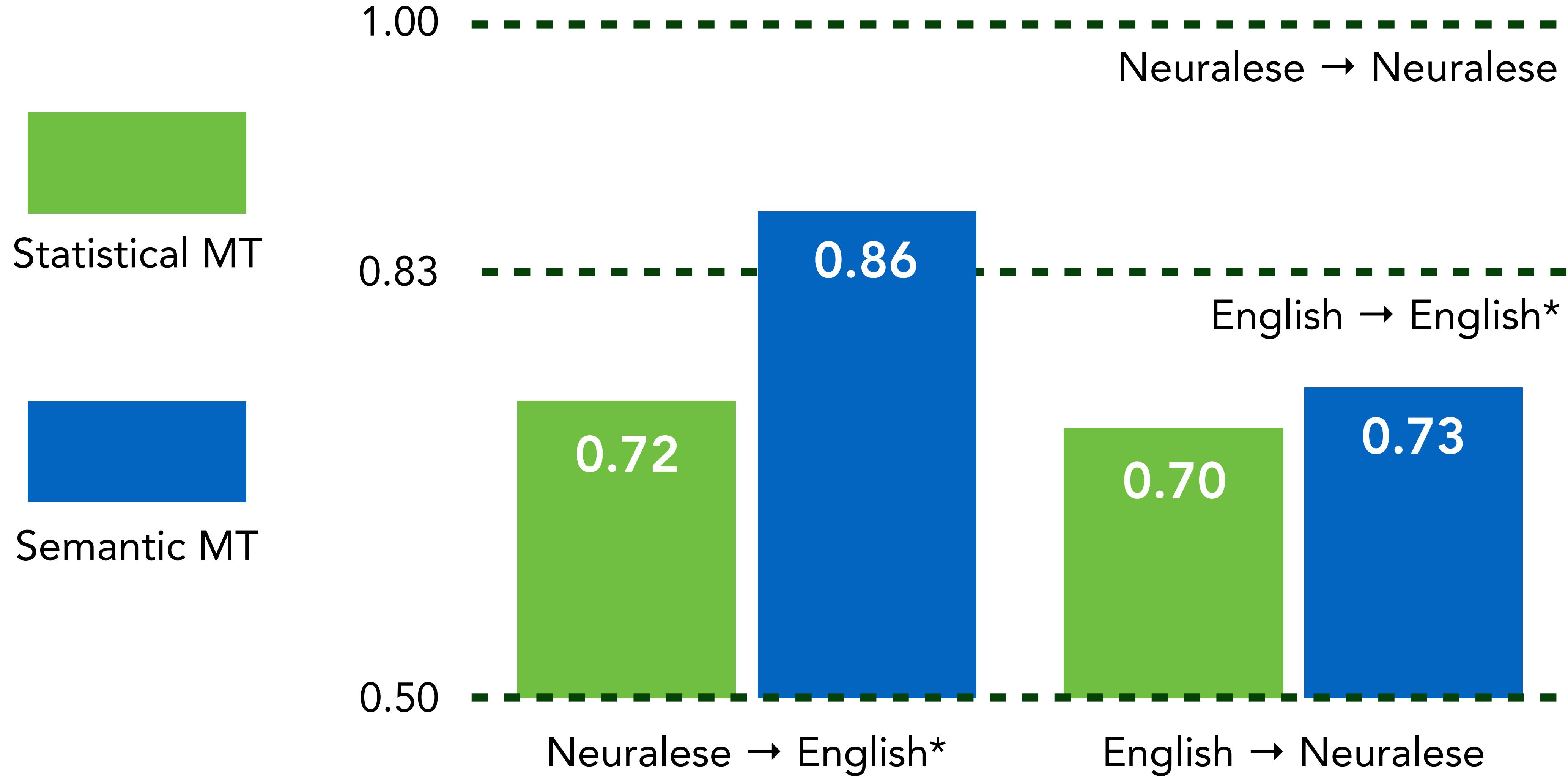


Experiment: color references



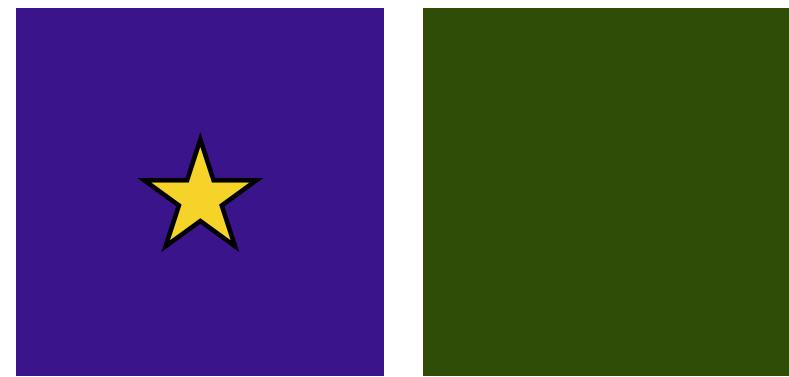


Experiment: color references





Experiment: color references



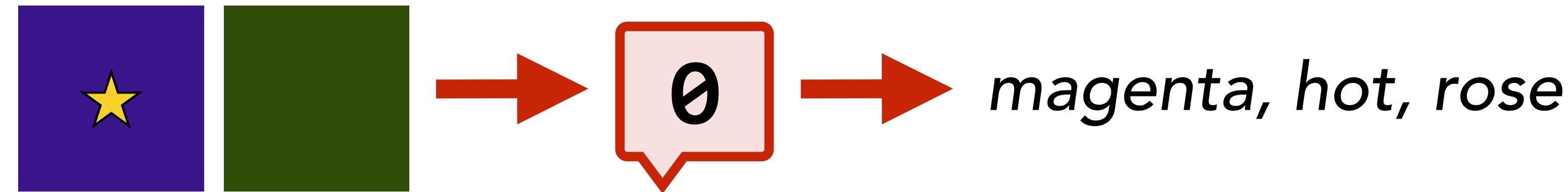


Experiment: color references



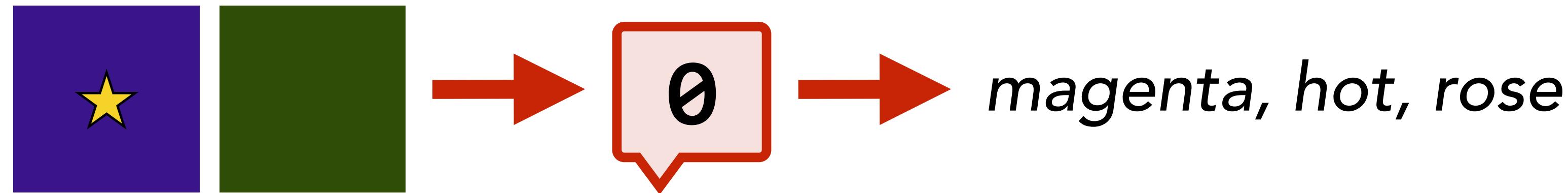


Experiment: color references



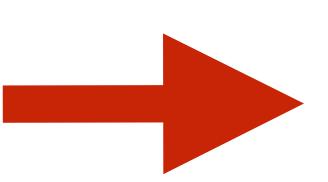
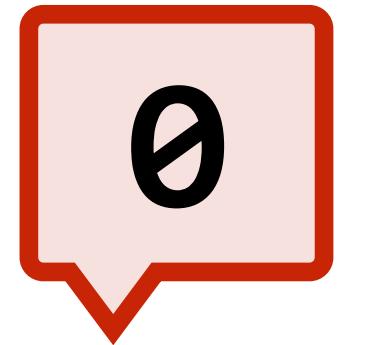
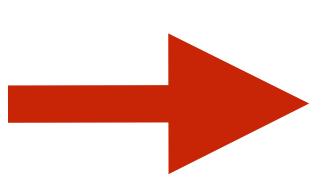
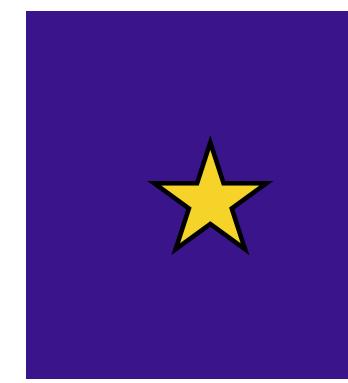


Experiment: color references

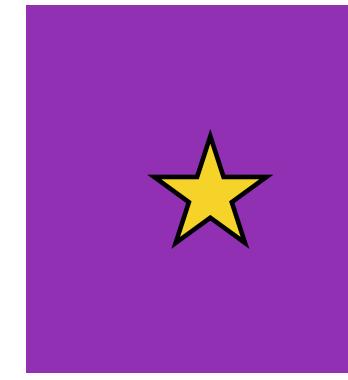




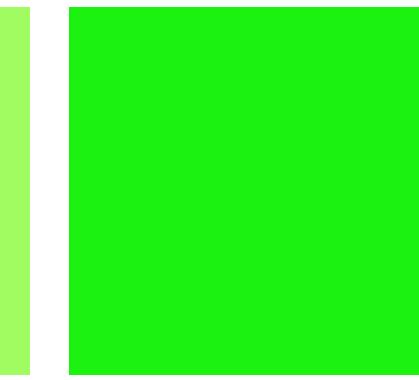
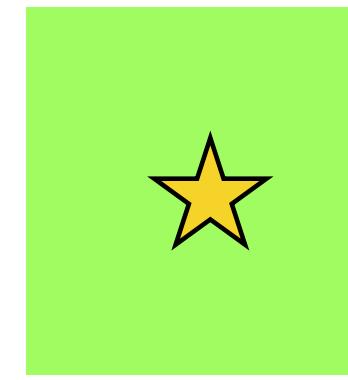
Experiment: color references



magenta, hot, rose



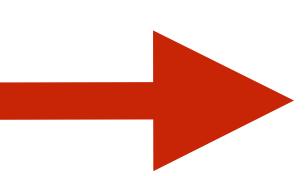
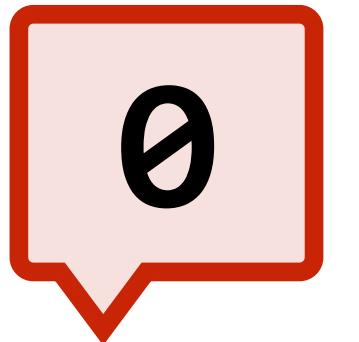
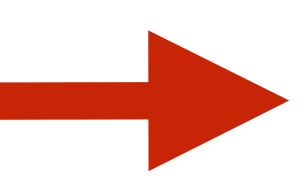
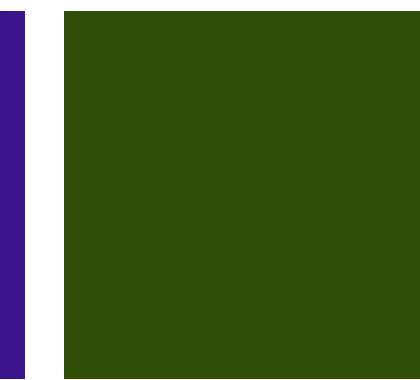
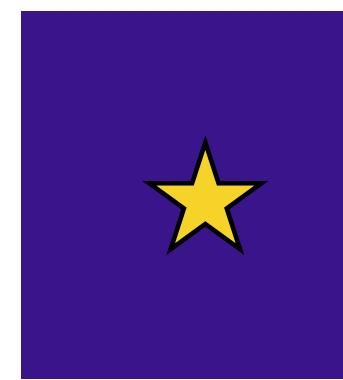
magenta, hot, violet



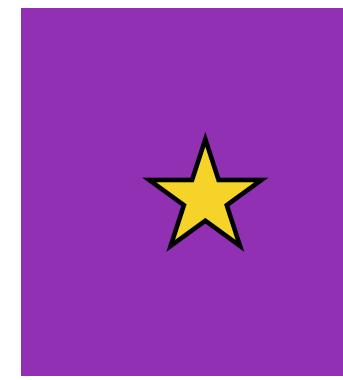
olive, puke, pea



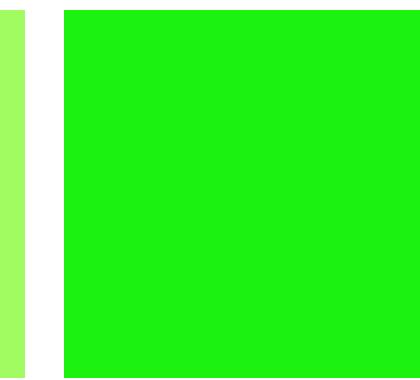
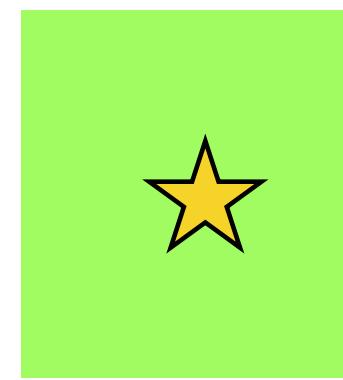
Experiment: color references



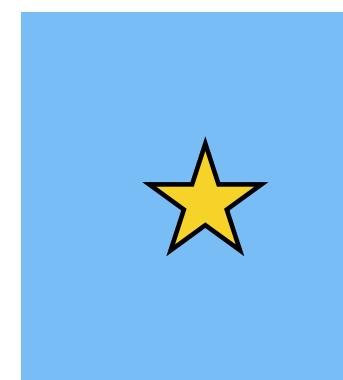
magenta, hot, rose



magenta, hot, violet



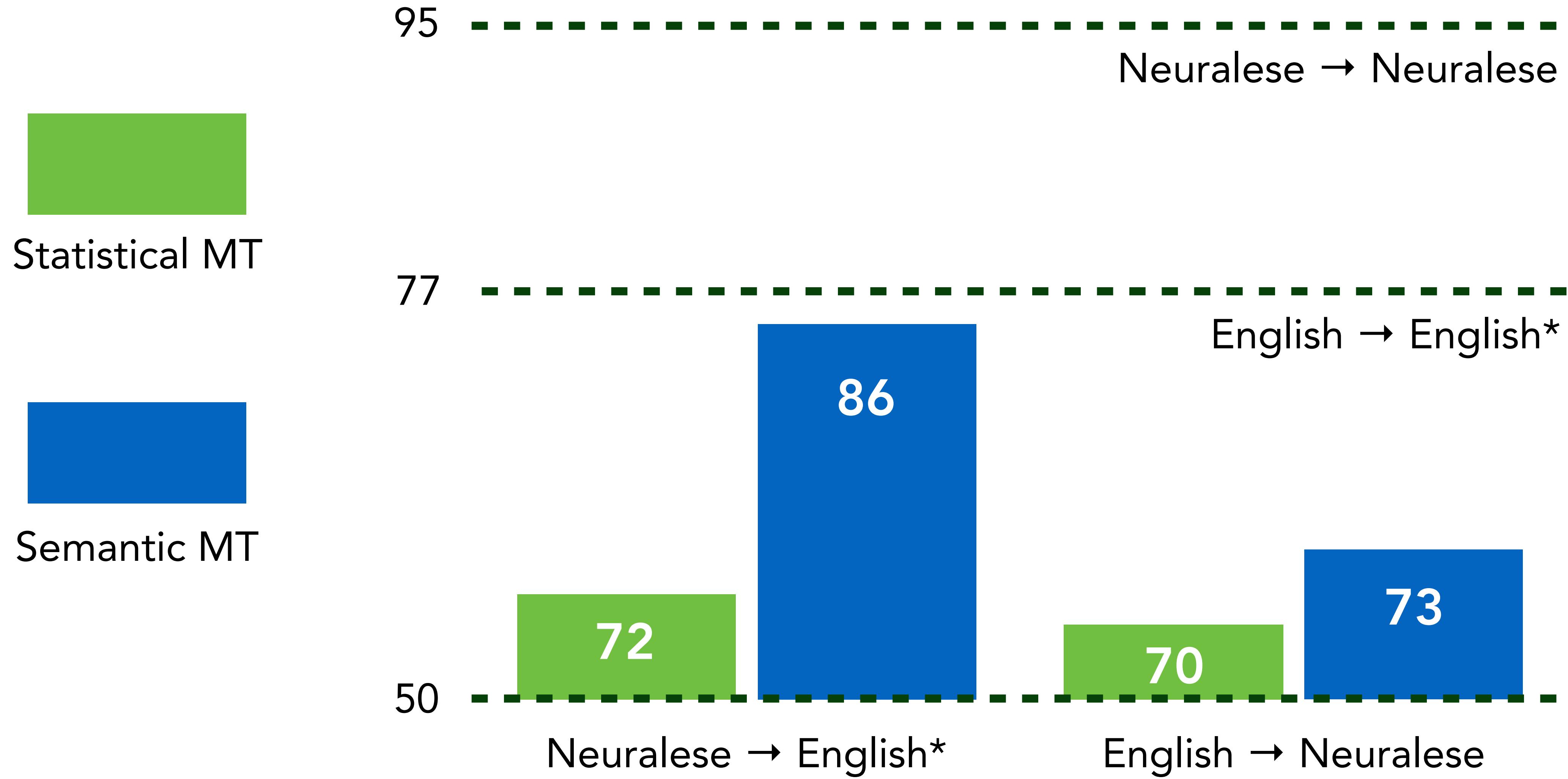
olive, puke, pea



pinkish, grey, dull



Experiment: image references





Experiment: image references



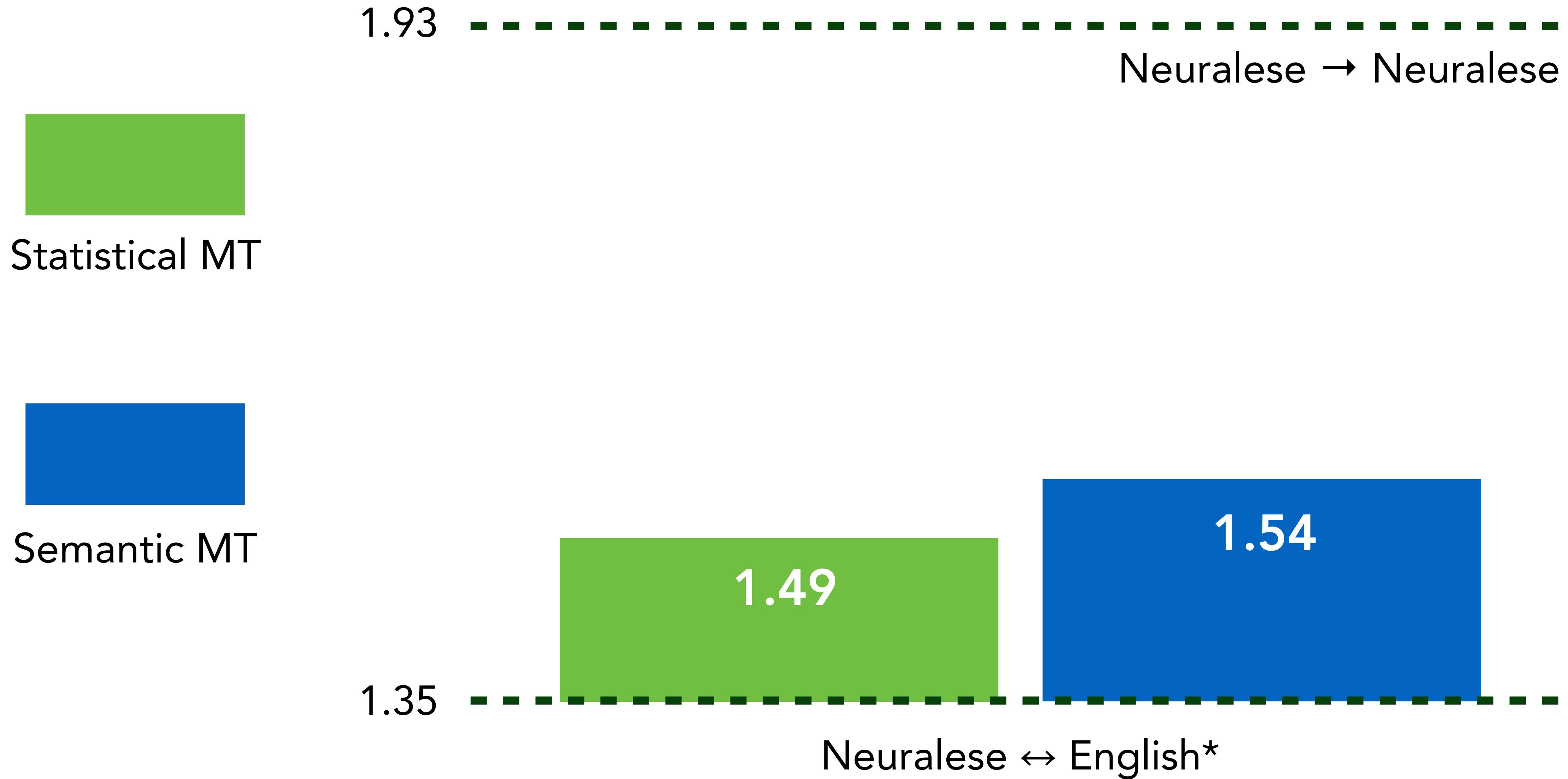
large bird, black wings, black crown



small brown, light brown, dark brown



Experiment: driving game





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!



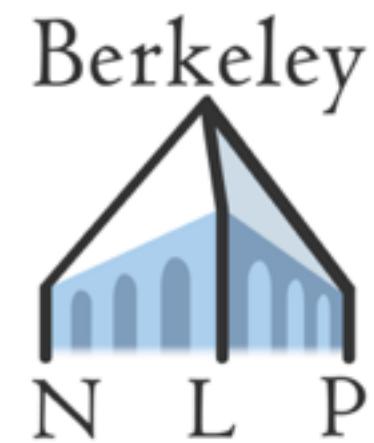
Conclusions

- 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!

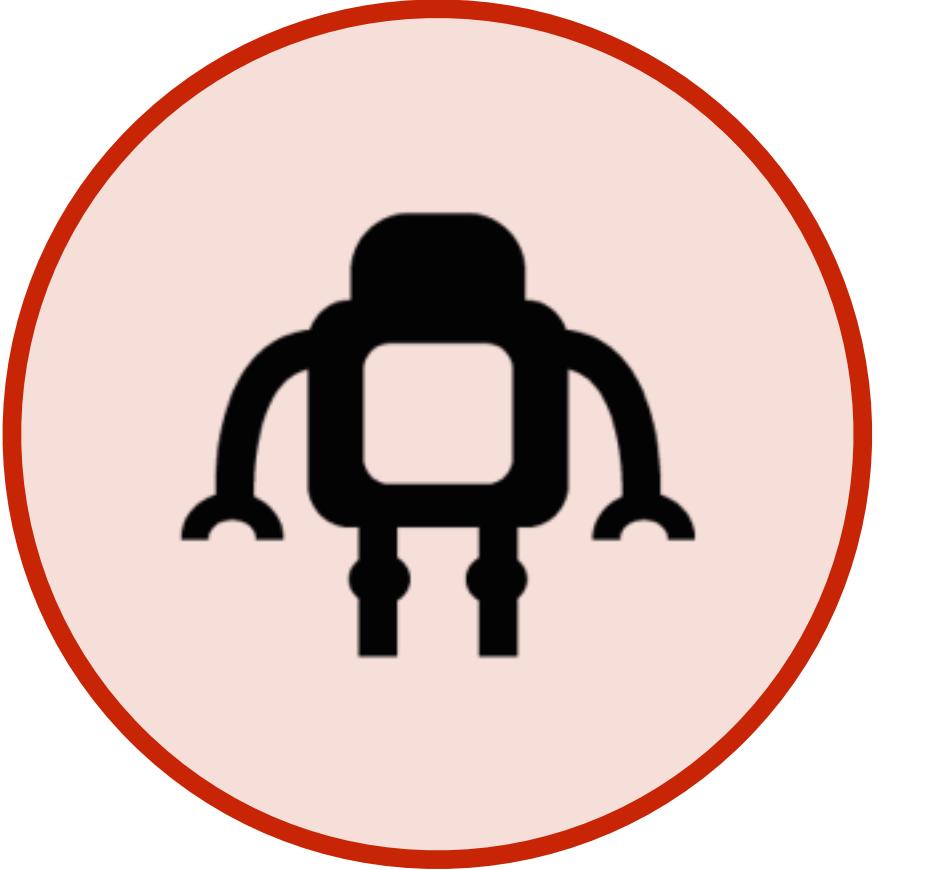


What about compositionality?

Analogs of linguistic structure
in deep representations



Jacob Andreas and Dan Klein



| | |
|------|-----|
| 1.0 | 2.3 |
| -0.3 | 0.4 |
| -1.2 | 1.1 |

A black silhouette of a person's head in profile, facing right. The head is enclosed within a circular frame that has a thick purple border and a light blue fill.

Thank you!