CS 295A/395D: Artificial Intelligence

More MDPs

Prof. Emma Tosch

27 April 2022

The University of Vermont

Recall: MDPs

An MDP is defined by the model $\langle S, A, T, R, \gamma \rangle$ such that:

$S = \{s_1, \ldots, s_n\}$ is the set of *n* states

$A = \{ a_1, \ldots, a_m\}$ is the set of *m* actions (assume wlog we can take every action in every state)

T is a representation of the transition probability into a state given an action and a current state (i.e., $P(s \mid s', a)$, possibly represented by a $(\mathrm{m} * n) \times n$ matrix of transition probabilities such that each row represents some $v_i = \langle s^1, a^i \rangle$ and

$$p_{ij} = P\left(X_t = s_j \mid X_{t-1} = v_i(0), A = v_i(1) \right)$$

$R : S \times A \times S \rightarrow \mathfrak{R}$ is the reward function, which can be defined in terms of the current state, action, next state, or even as a probabilistic map – it is whatever you need it to be

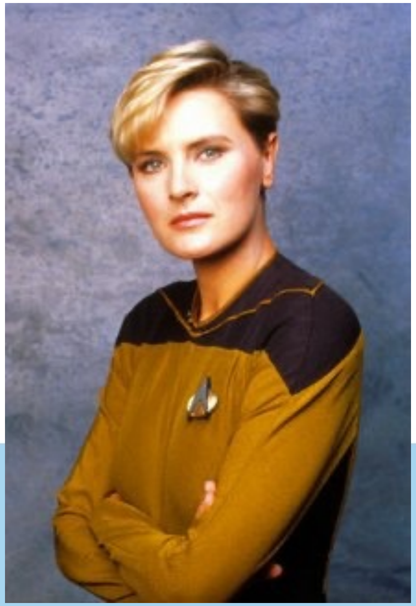
Recall: LTL Example

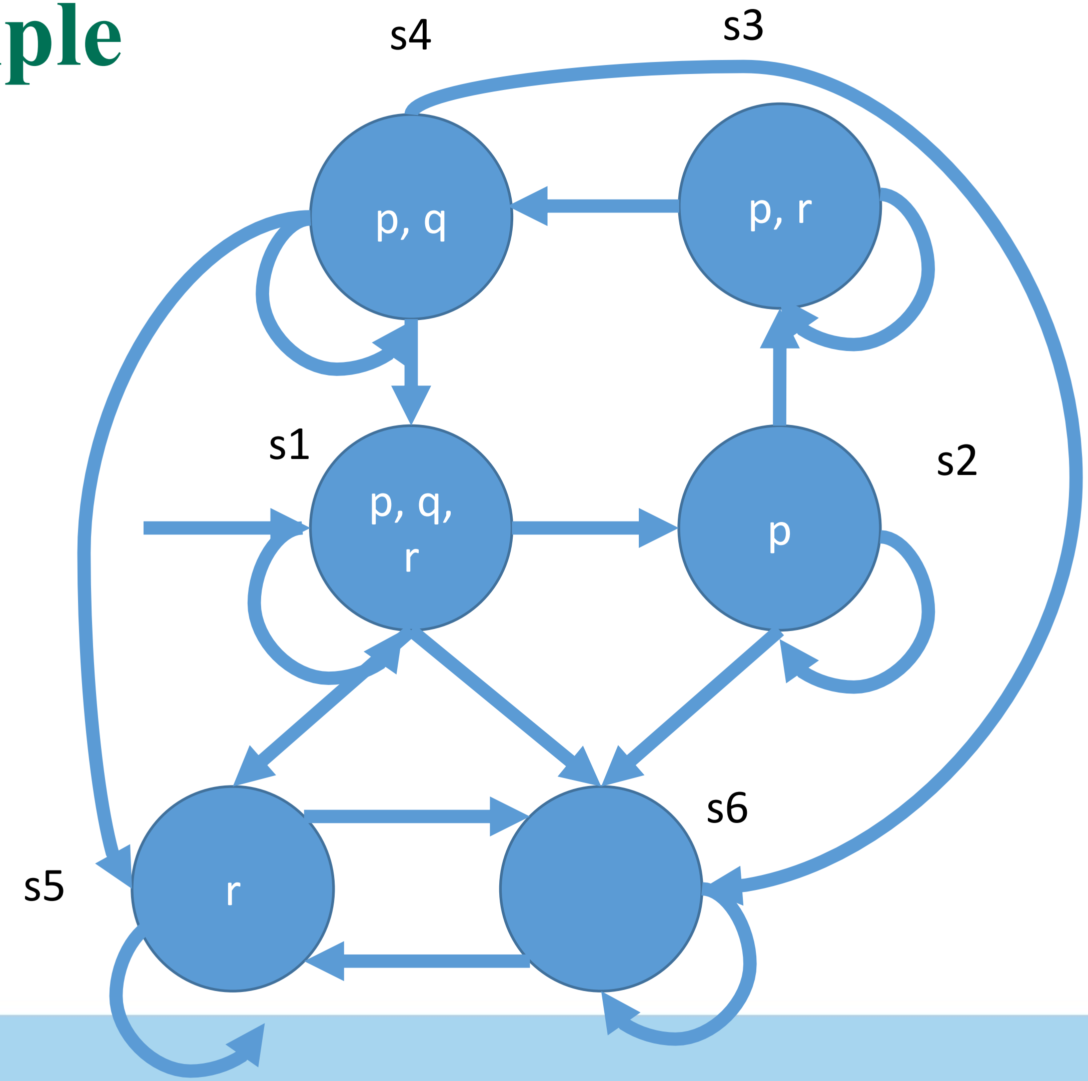
Example:

p = Yar alive

q = Yar on our ship

r = transporter ready





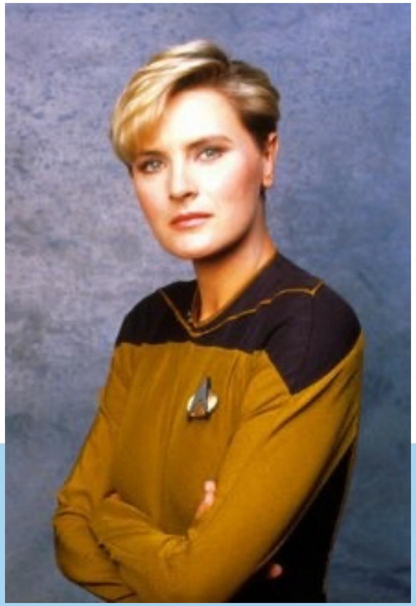
LTL Exercises

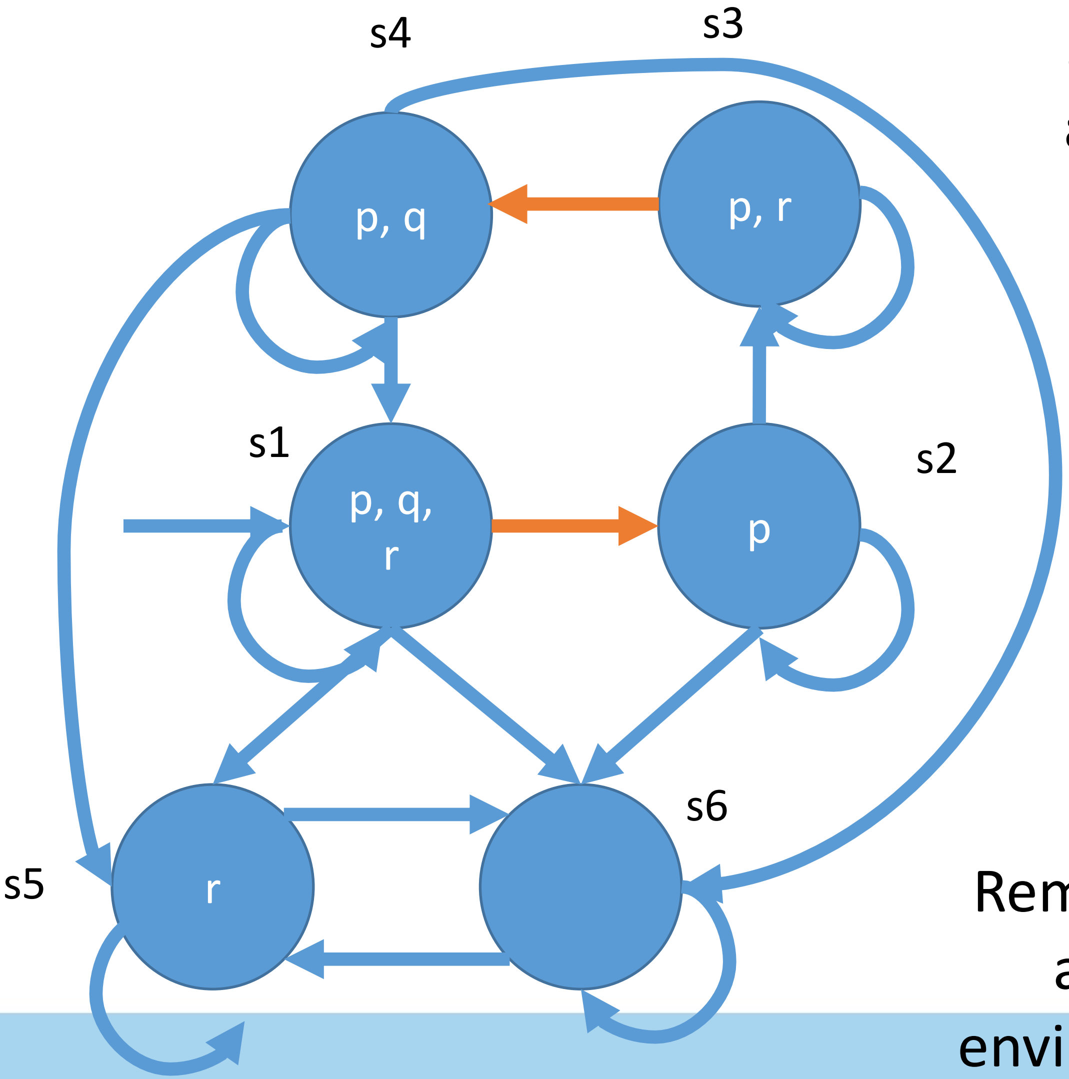
Example:

p = Yar alive

q = Yar on our ship

r = transporter ready





Some state transitions are associated with **actions** *we* take

Remaining transitions associated with environment dynamics

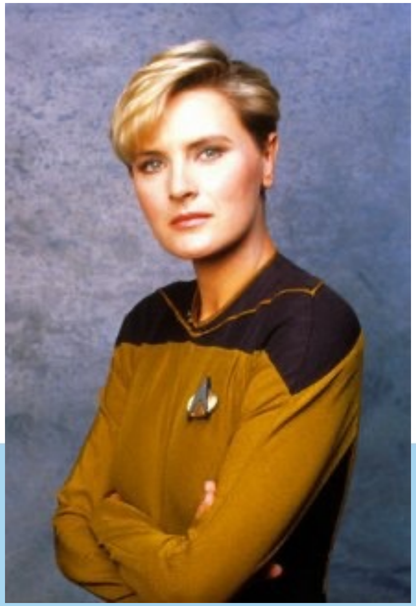
LTL Exercises

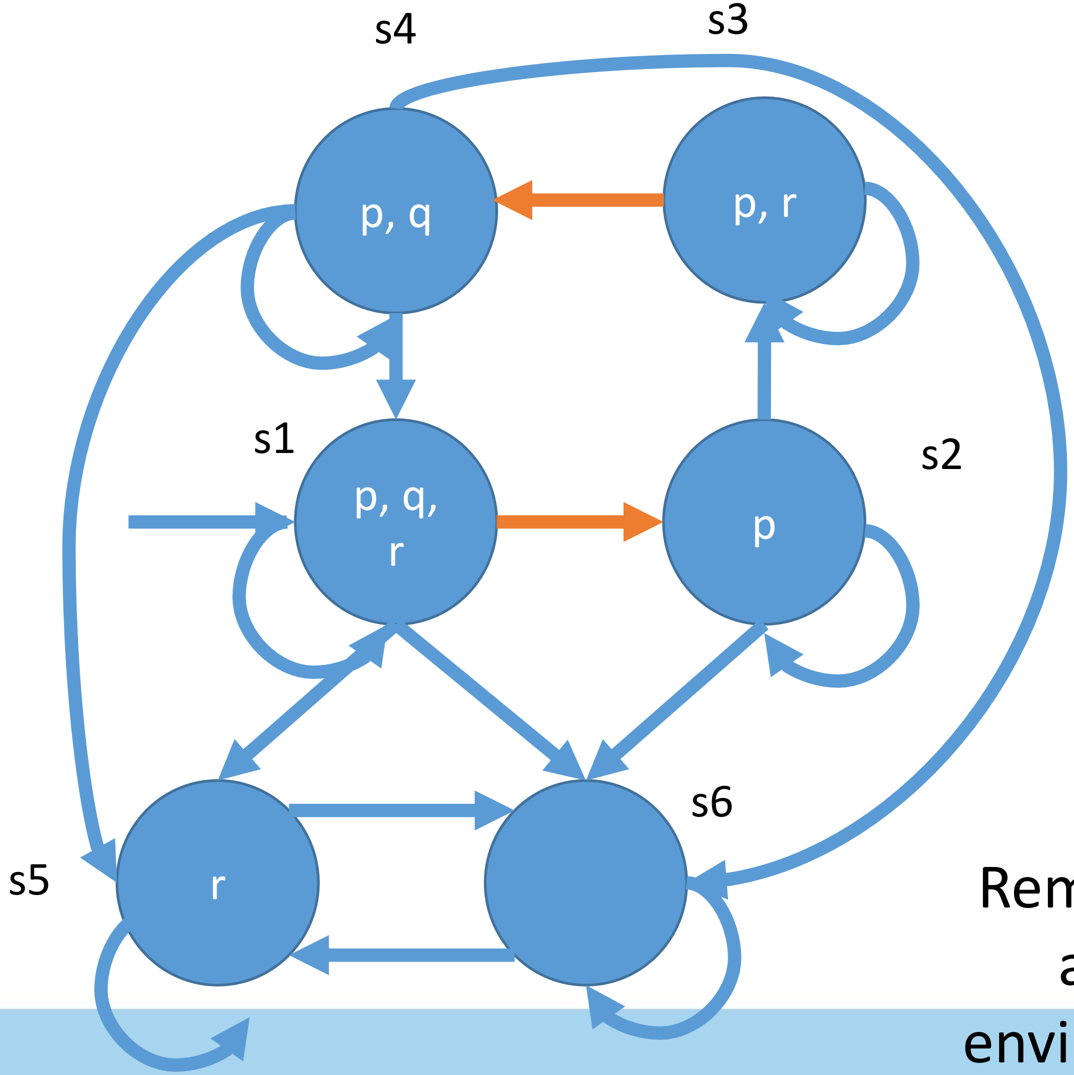
Example:

p = Yar alive

q = Yar on our ship

r = transporter ready





Orange = use transporter

Remaining transitions associated with environment dynamics

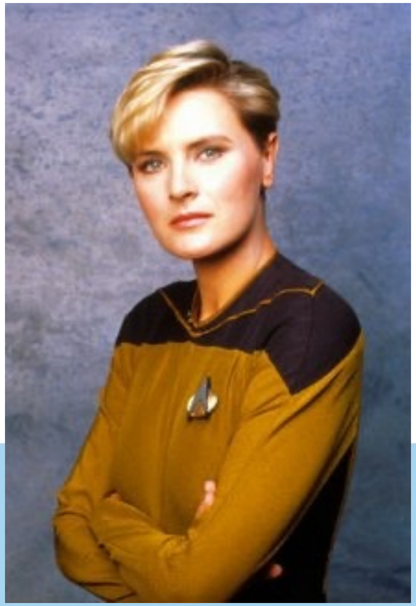
LTL Exercises

Example:

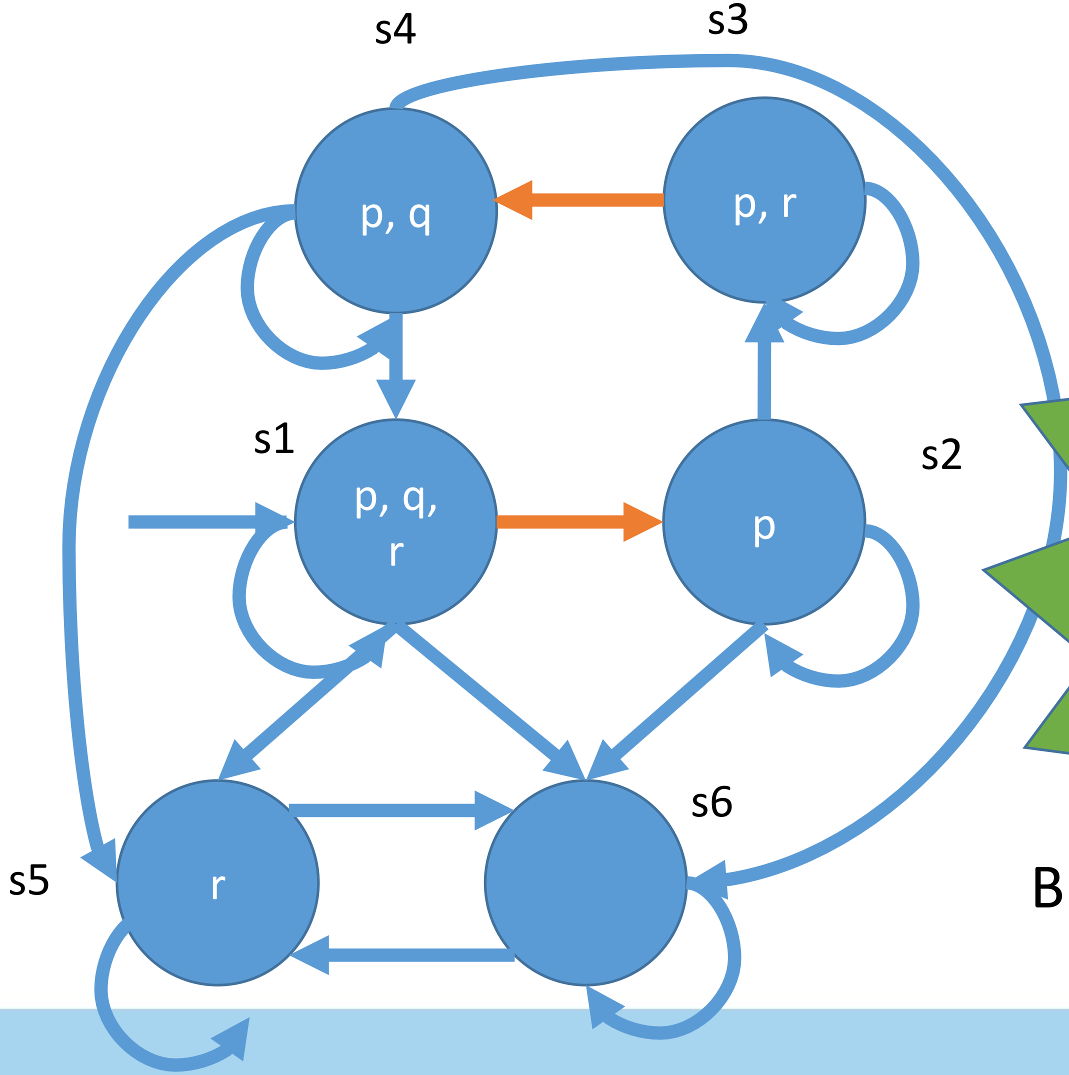
p = Yar alive

q = Yar on our ship

r = transporter ready



Orange = use transporter



Assign transition probabilities, starting with (s1, a1)

Blue = do nothing

Transporter as MDP

Actions: Orange (a1), Blue (a2)

$P(s1 \mid s1, a1) = 0$

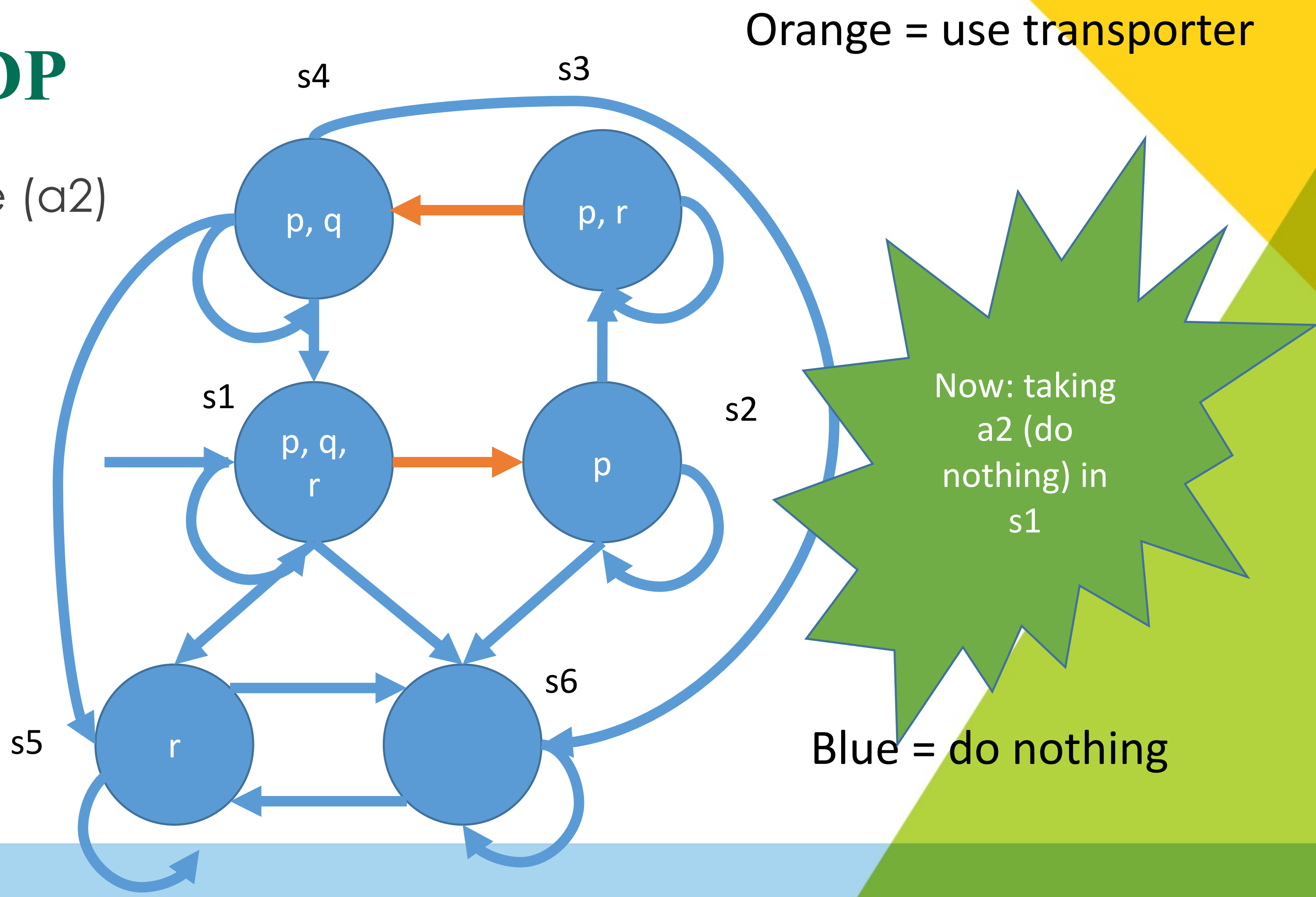
$P(s2 \mid s1, a1) = 1$

$P(s3 \mid s1, a1) = 0$

$P(s4 \mid s1, a1) = 0$

$P(s5 \mid s1, a1) = 0$

$P(s6 \mid s1, a1) = 0$



Transporter as MDP

Actions: Orange (a1), Blue (a2)

P(s1 | s1, a2) = 0.8

P(s2 | s1, a2) = 0

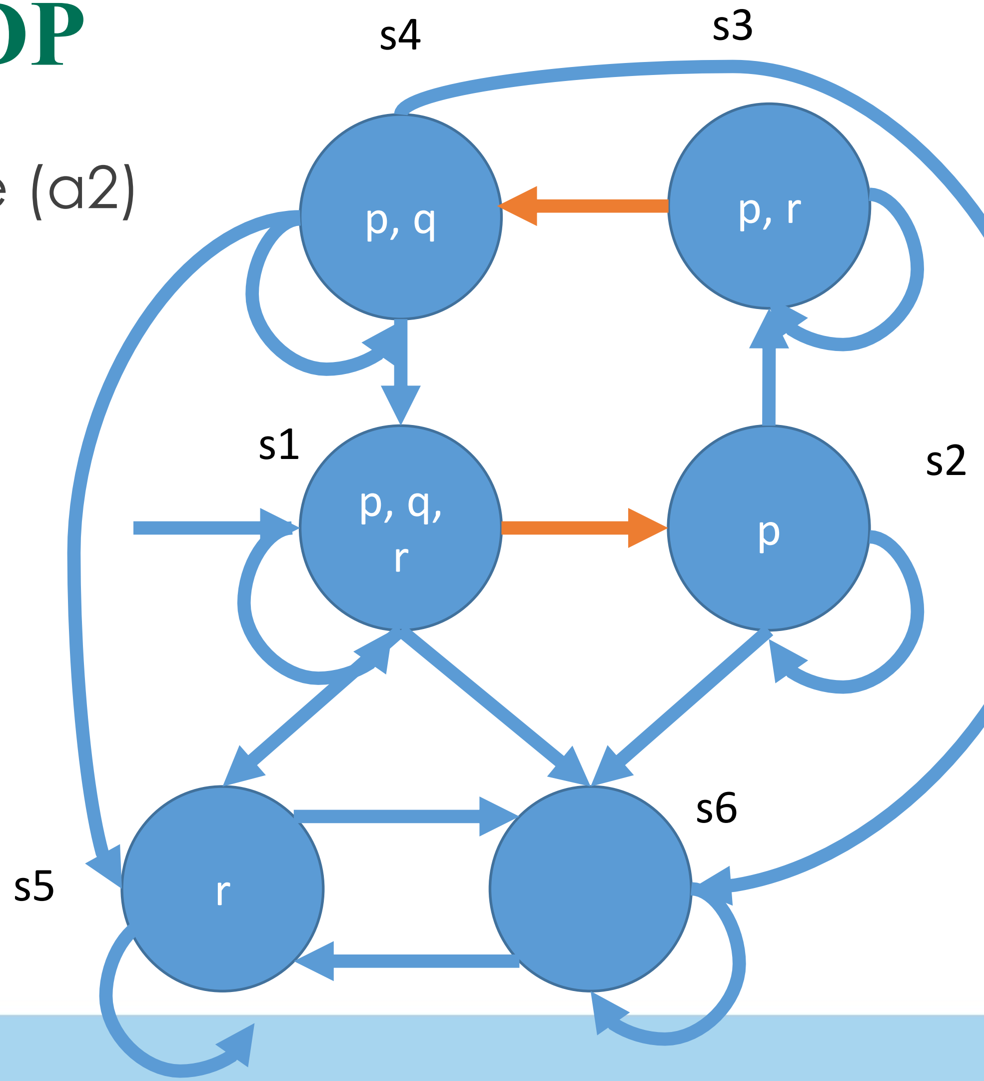
P(s3 | s1, a2) = 0

P(s4 | s1, a2) = 0

P(s5 | s1, a2) = 0.1

P(s6 | s1, a2) = 0.1

Orange = use transporter



Now: taking a2 (do nothing) in s1

Blue = do nothing

Transporter as MDP

Actions: Orange (a1), Blue (a2)

P(s1 | s2, a1) = 0

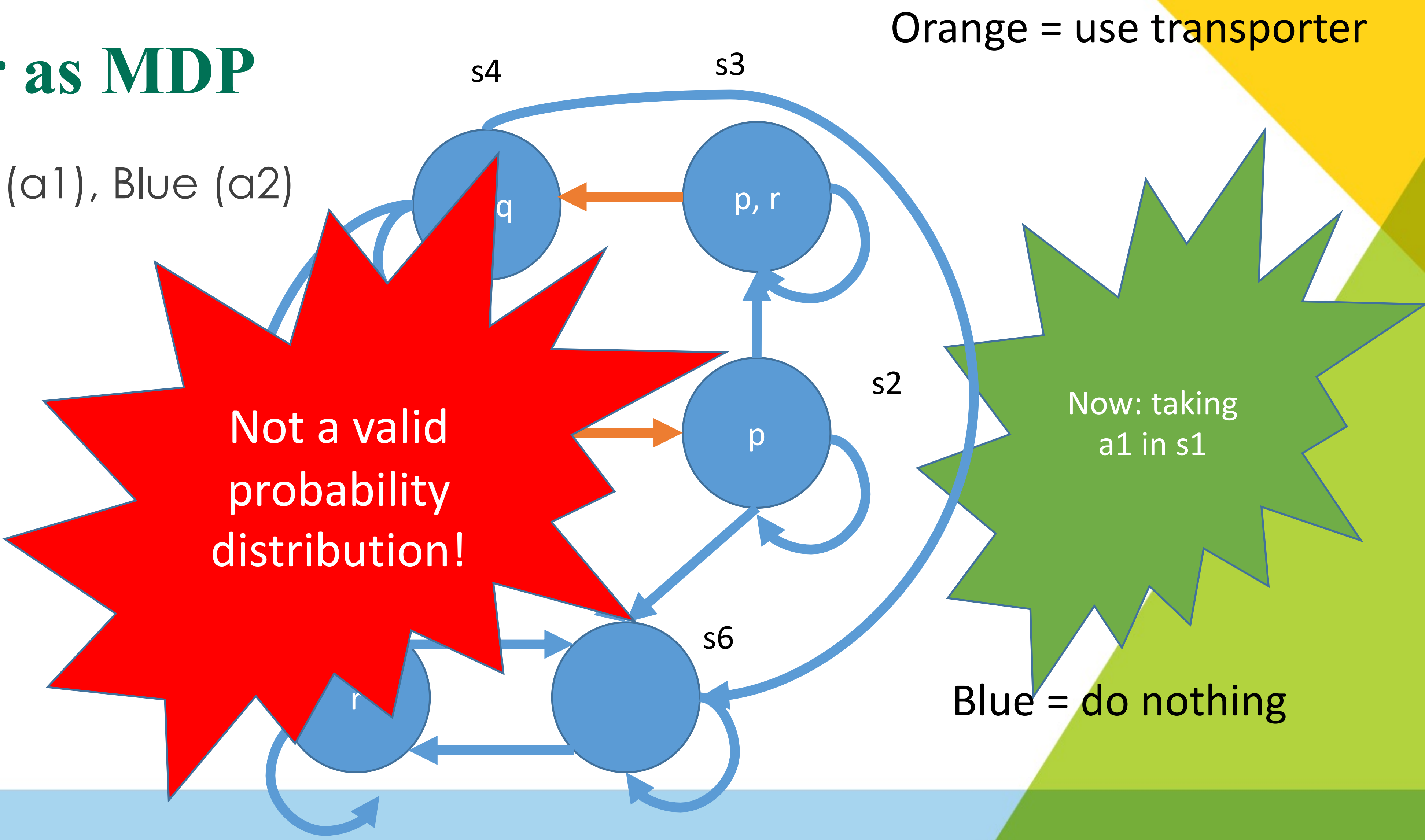
P(s2 | s2, a1) = 0

P(s3 | s2, a1) = 0

P(s4 | s2, a1) = 0

P(s5 | s2, a1) = 0

P(s6 | s2, a1) = 0



Transporter as MDP

Actions: Orange (a1), Blue (a2)

~~P(s1 | s2, a1) = 0~~

~~P(s2 | s2, a1) = 0~~

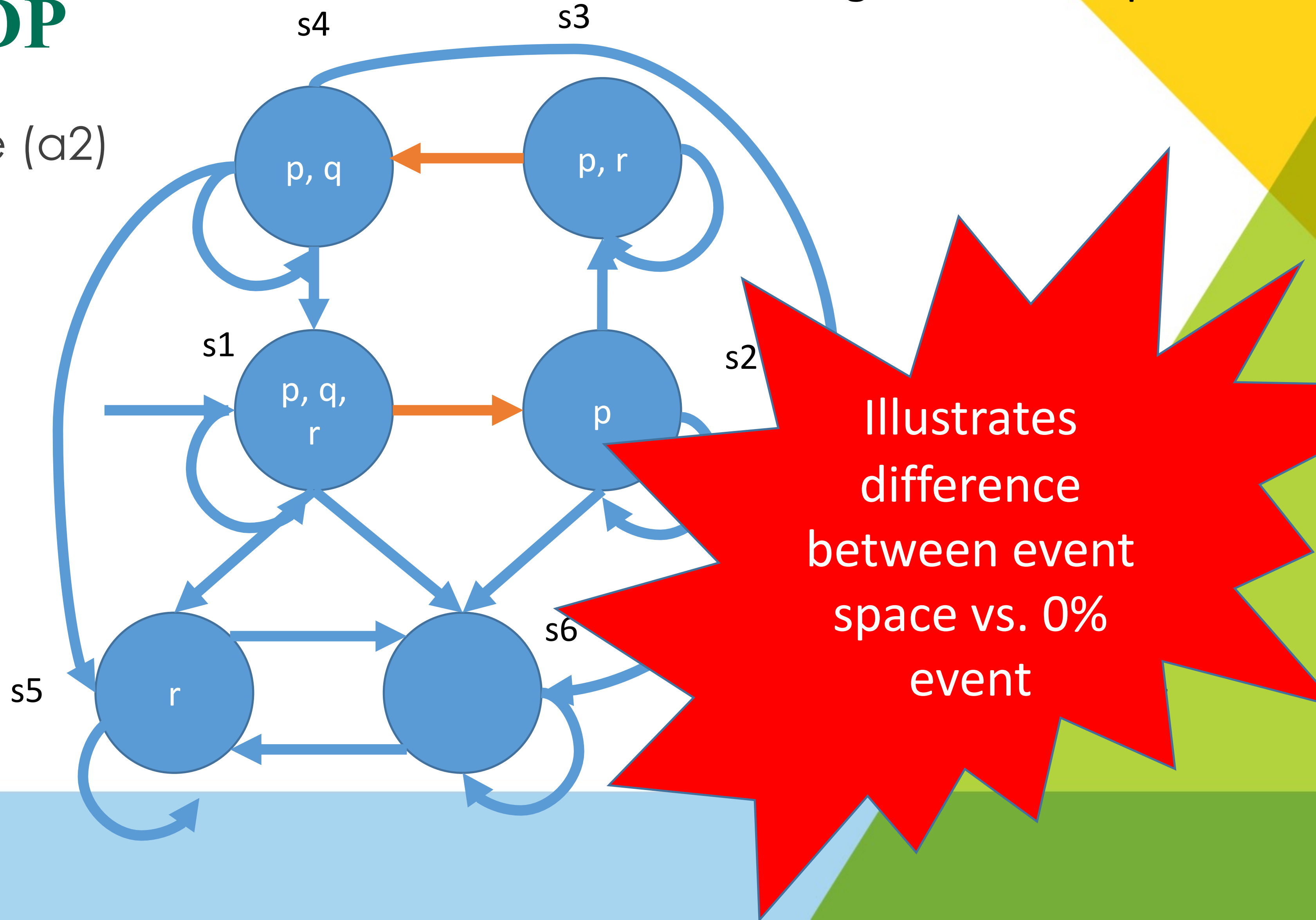
~~P(s3 | s2, a1) = 0~~

~~P(s4 | s2, a1) = 0~~

~~P(s5 | s2, a1) = 0~~

~~P(s6 | s2, a1) = 0~~

Orange = use transporter



Transporter as MDP

Actions: Orange (a1), Blue (a2)

P(s1 | s2, a1) = 1

P(s2 | s2, a1) = 0

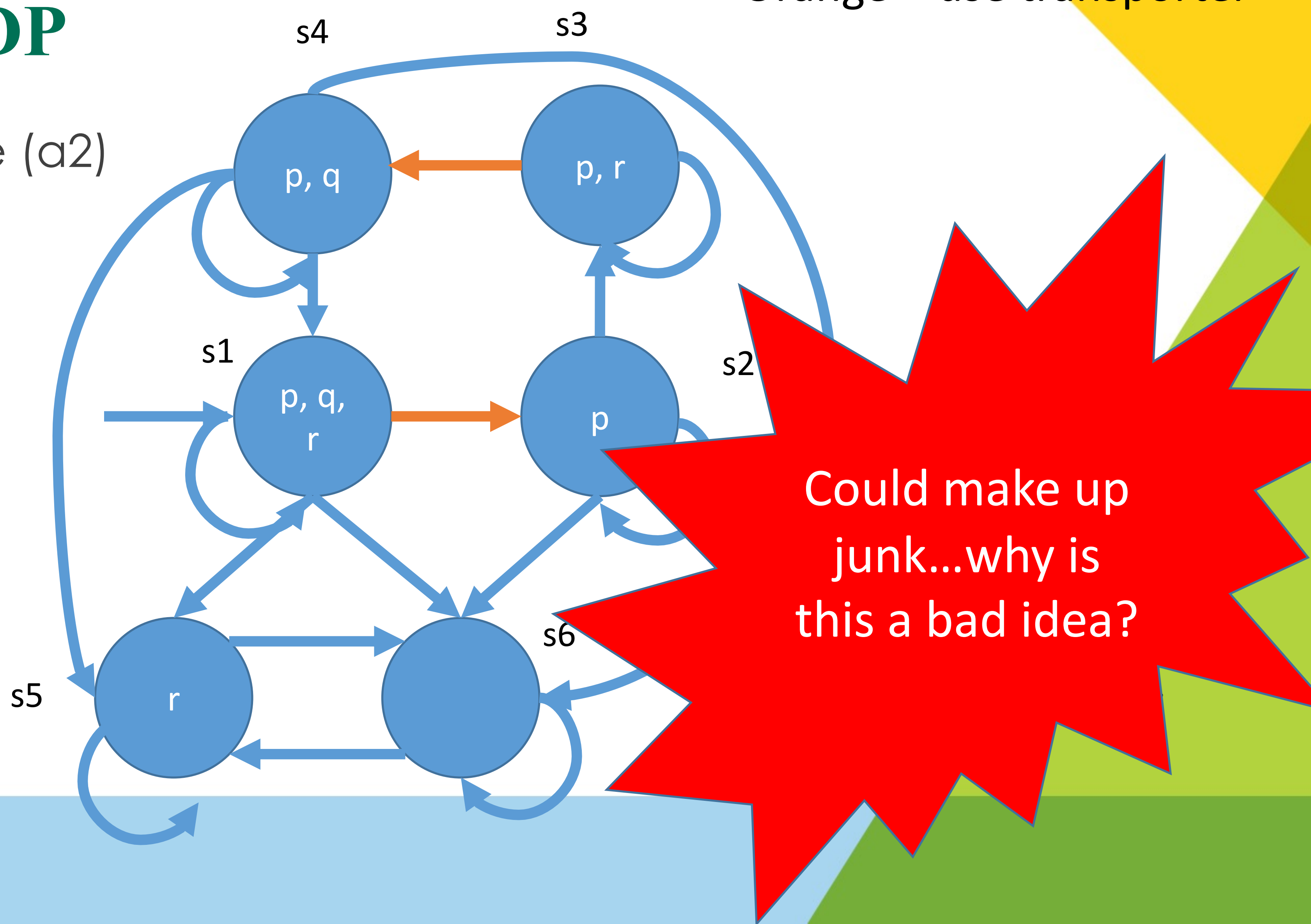
P(s3 | s2, a1) = 0

P(s4 | s2, a1) = 0

P(s5 | s2, a1) = 0

P(s6 | s2, a1) = 0

Orange = use transporter



Transporter as MDP

Orange = use transporter

Actions: Orange (a1), Blue (a2)

~~P(s1 | s2, a1) = 0~~

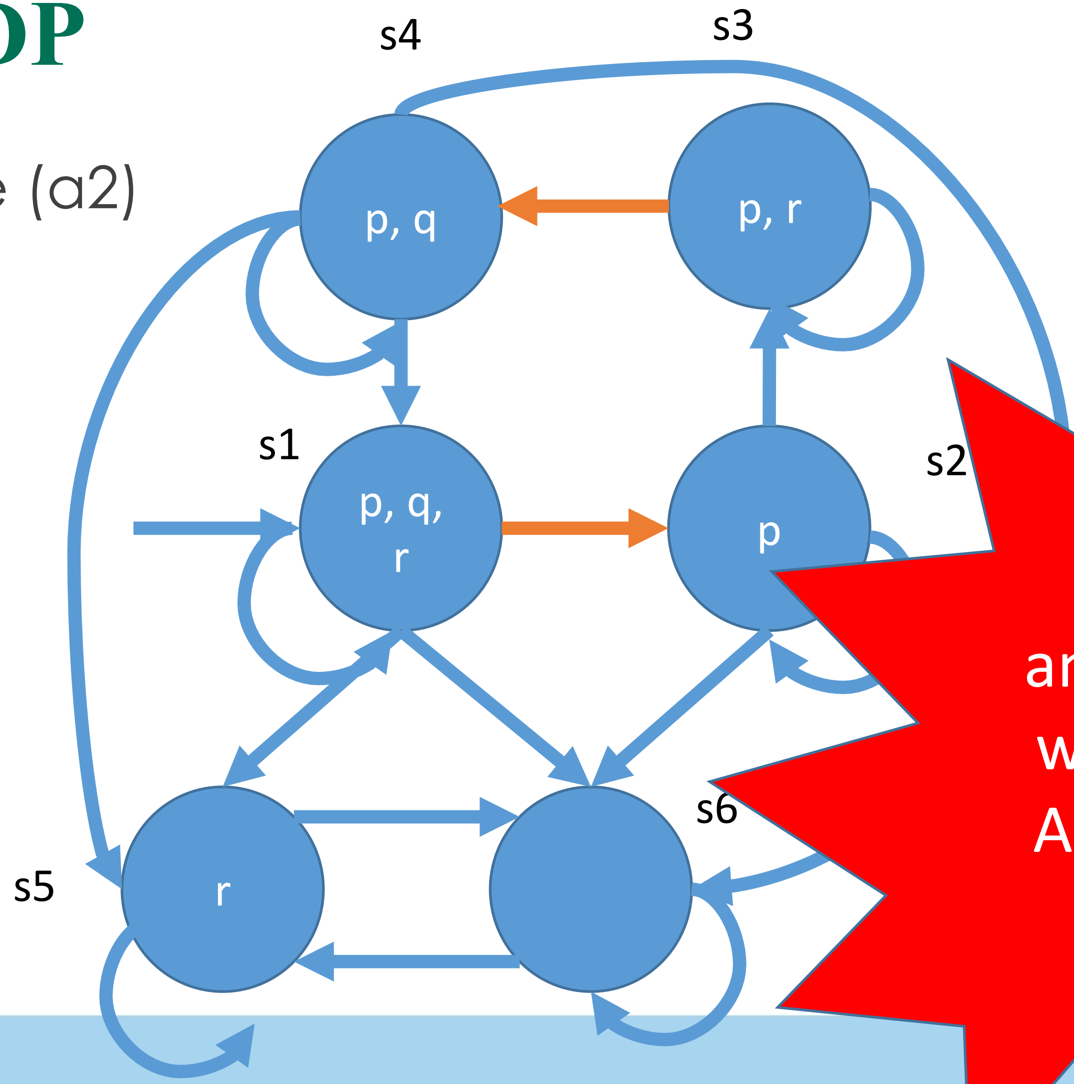
~~P(s2 | s2, a1) = 0~~

~~P(s3 | s2, a1) = 0~~

~~P(s4 | s2, a1) = 0~~

~~P(s5 | s2, a1) = 0~~

~~P(s6 | s2, a1) = 0~~



Illustrates another reason why we define A to be a map!

Transporter as MDP

Actions: Orange (a1), Blue (a2)

$P(s1 \mid s2, a2) = 0$

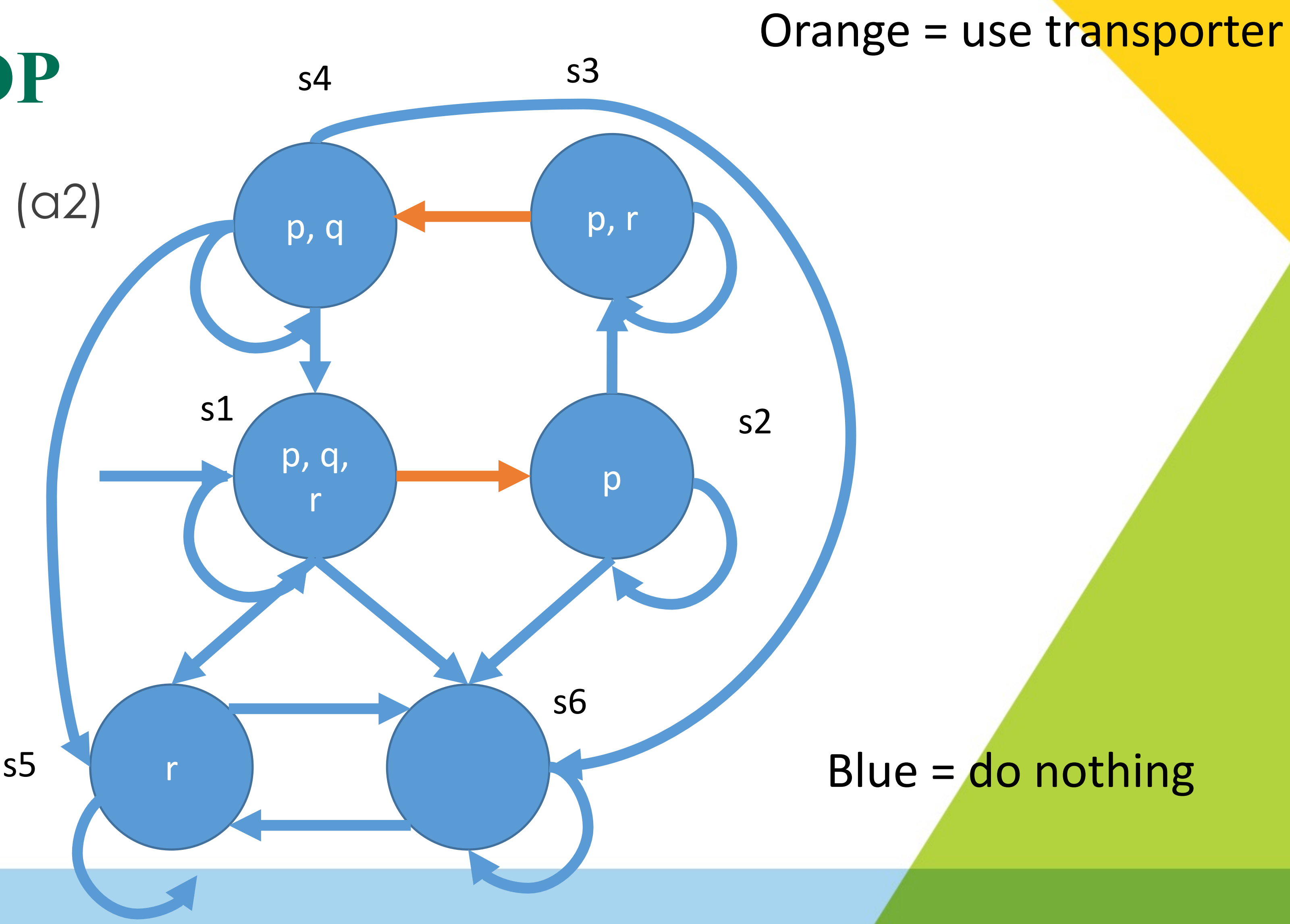
$P(s2 \mid s2, a2) = 0.5$

$P(s3 \mid s2, a2) = 0.4$

$P(s4 \mid s2, a2) = 0$

$P(s5 \mid s2, a2) = 0$

$P(s6 \mid s2, a2) = 0.1$



Transporter as MDP

Actions: Orange (a1), Blue (a2)

$P(s1 \mid s3, a1) = 0$

$P(s2 \mid s3, a1) = 0$

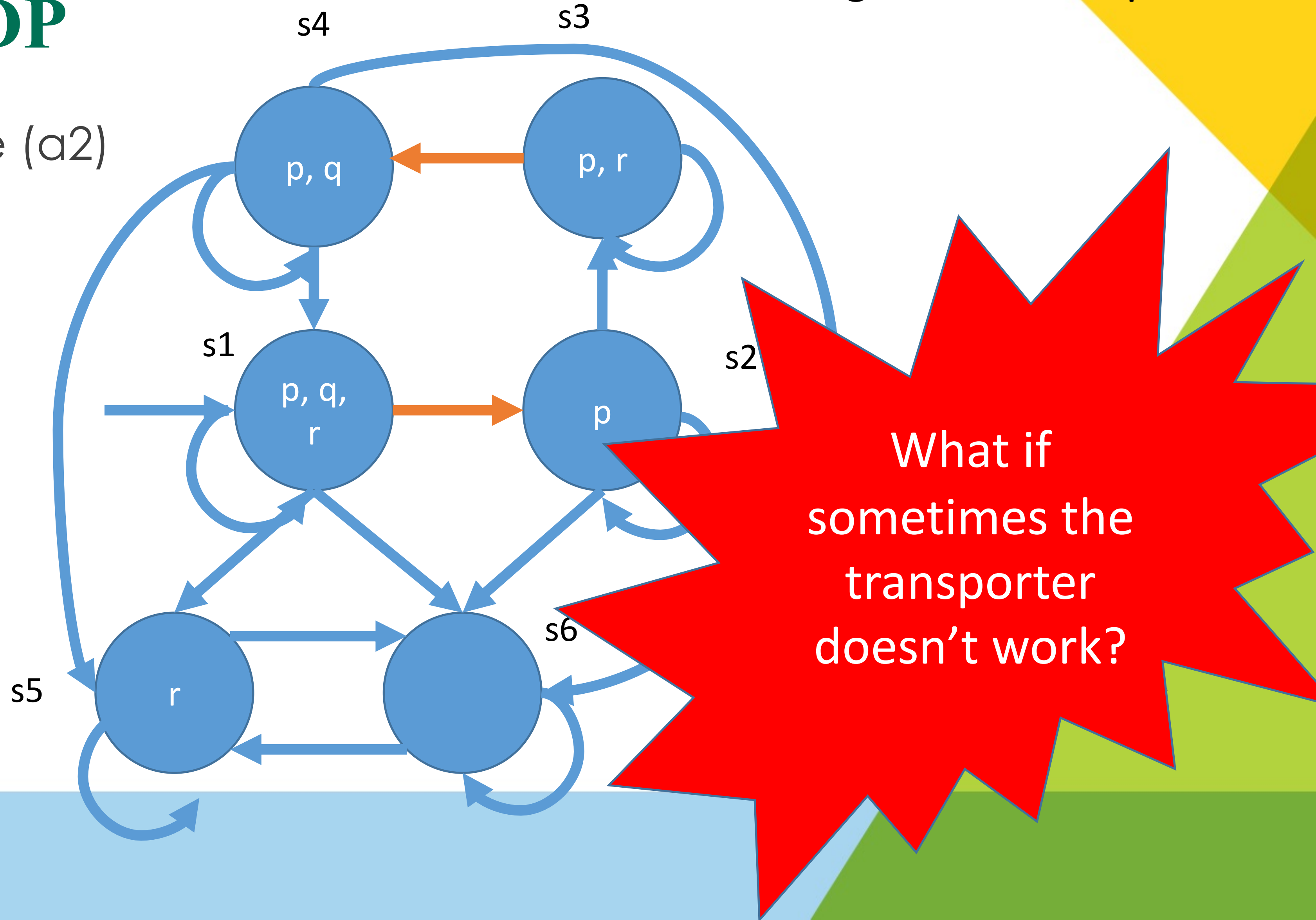
$P(s3 \mid s3, a1) = 0.1$

$P(s4 \mid s3, a1) = 0.9$

$P(s5 \mid s3, a1) = 0$

$P(s6 \mid s3, a1) = 0$

Orange = use transporter



Sources of randomness and variability

1. Environment dynamics :
    1. Aleatory: Probabilistic transitions between states
    2. Epistemic: Uncertainty over state values

POMDP

2. Probability of action succeeding vs. taking an action
    1. Former: Can also be modeled as environmental!

MDP

    2. Latter: Randomness comes from the agent

RL

MDPs vs RL

MDPs

- Problem: Sequential Decision-Making
- Domain: environments having a particular property (Markovian)
- Describes framework over which to search for decisions in that environment

RL

- Problem: Learning to act over time (agent-based)
- Domain: any environment that emits a signal
- Describes a *framework* for *learning* over sequences

Transporter as MDP

Actions: Orange (a1), Blue (a2)

P(s1 | s3, a1) = 0.9

P(s2 | s3, a1) = 0

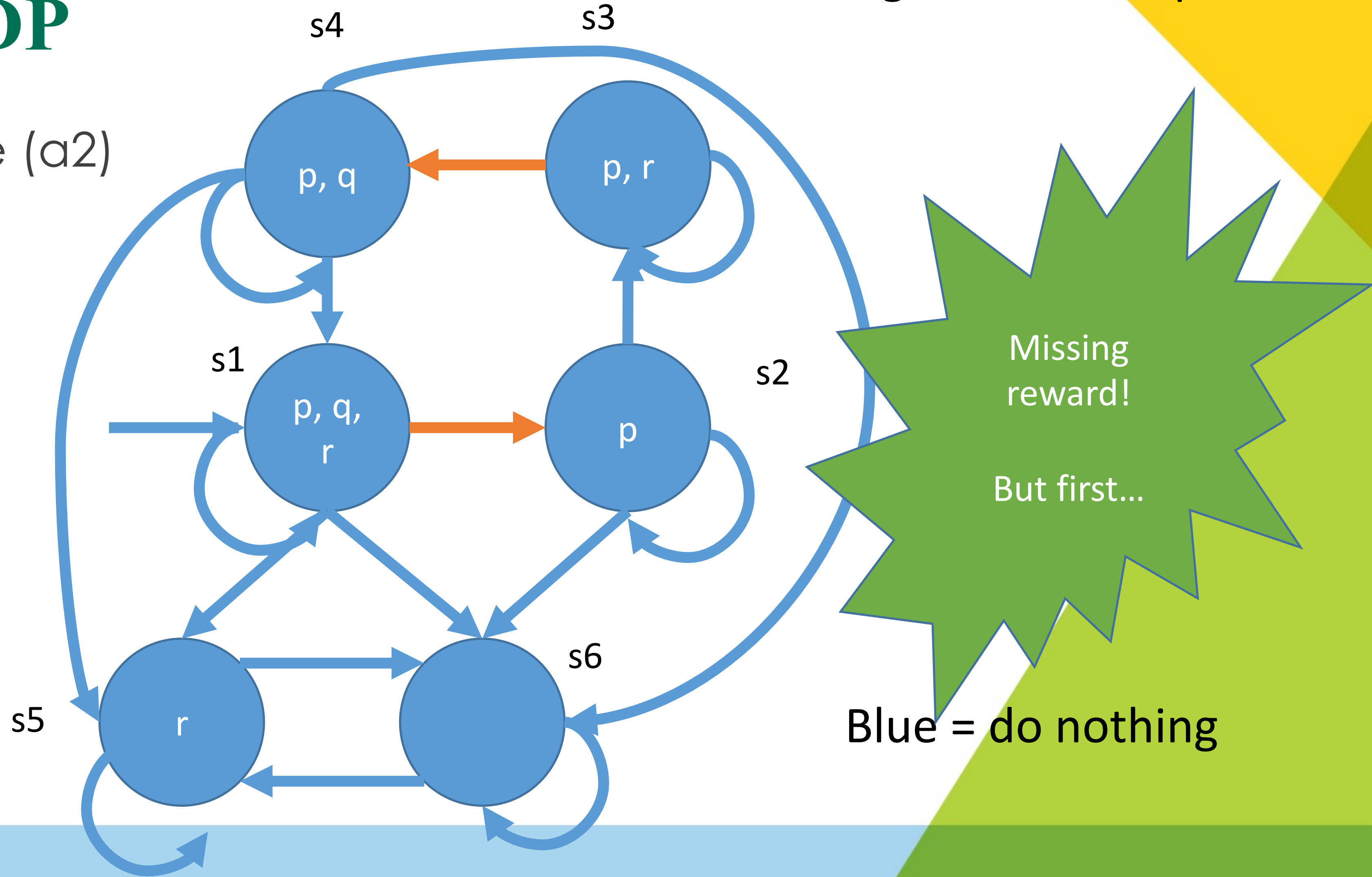
P(s3 | s3, a1) = 0.1

P(s4 | s3, a1) = 0

P(s5 | s3, a1) = 0

P(s6 | s3, a1) = 0

Orange = use transporter



Blue = do nothing

What does it mean to be in a particular state?

- State completely abstracted in the MDP framework
- How do we know we are in a particular state?
  - This is a deep question that connects back to the ontology discussion at the start of the semester
  - Recall: extensional vs. intensional definitions of sets
- Now: intensional (assume we can read features that allow us to evaluate predicates…)

Transporter as MDP

$(p, q, r) \mapsto s_1$

$(p, q, \neg r) \mapsto s_4$

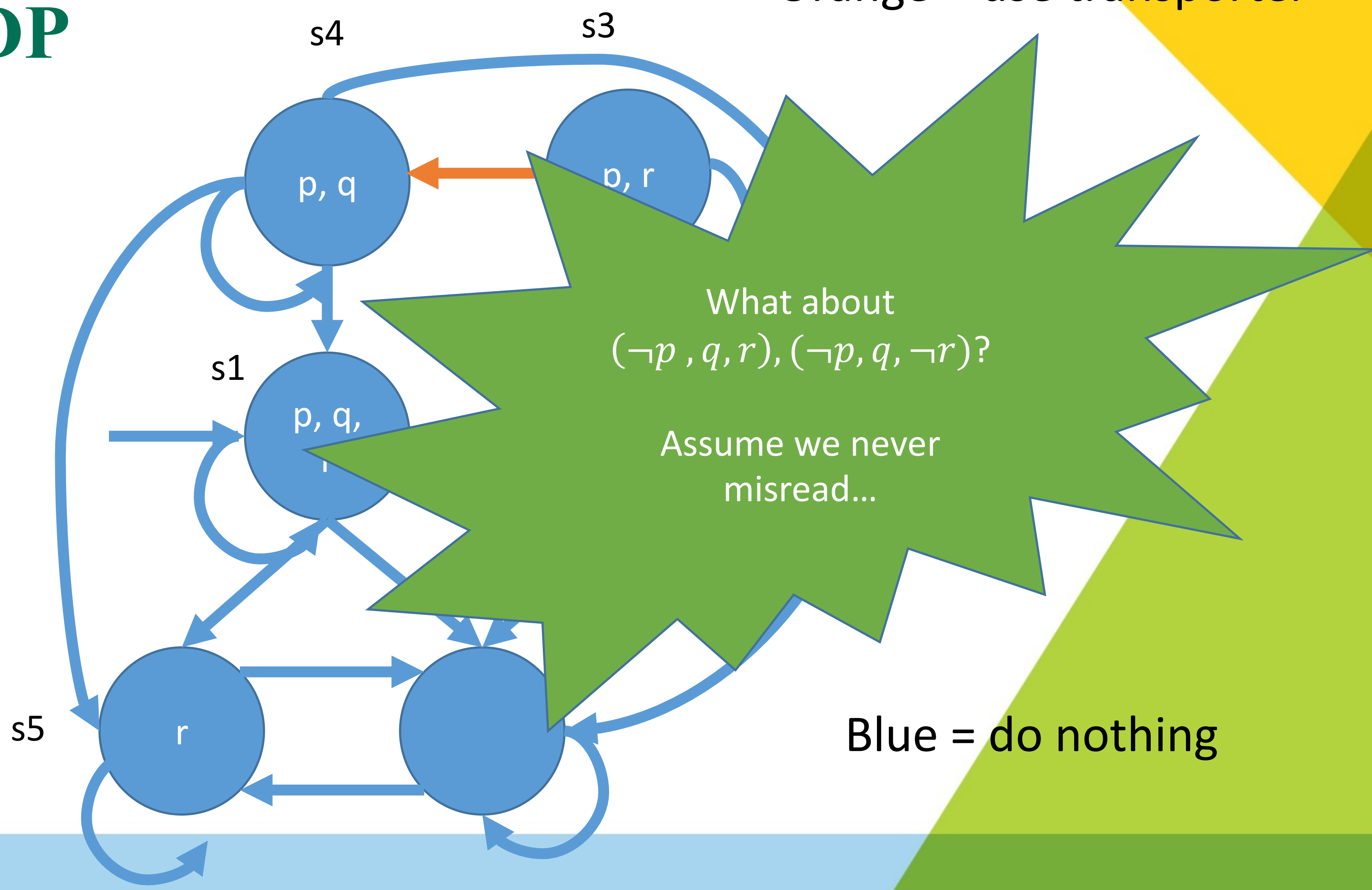
$(p, \neg q, r) \mapsto s_3$

$(p, \neg q, \neg r) \mapsto s_2$

$(\neg p, \neg q, r) \mapsto s_5$

$(\neg p, \neg q, \neg r) \mapsto s_6$

Orange = use transporter



Blue = do nothing

Transporter as MDP

$(p, q, r) \mapsto s_1$

$(p, q, \neg r) \mapsto s_4$

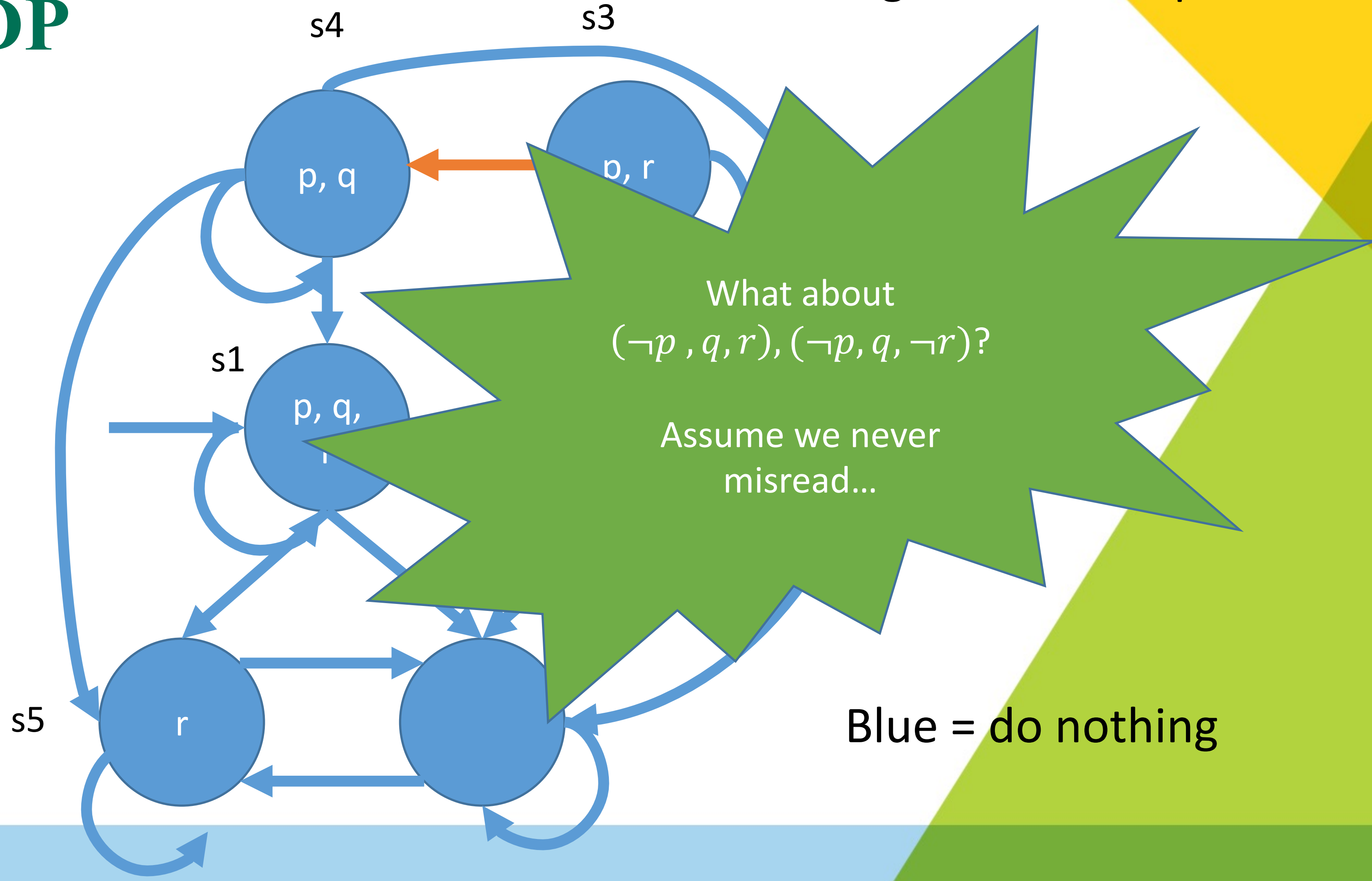
$(p, \neg q, r) \mapsto s_3$

$(p, \neg q, \neg r) \mapsto s_2$

$(\neg p, \neg q, r) \mapsto s_5$

$(\neg p, \neg q, \neg r) \mapsto s_6$

Orange = use transporter



Blue = do nothing

Transporter Policy

Example:

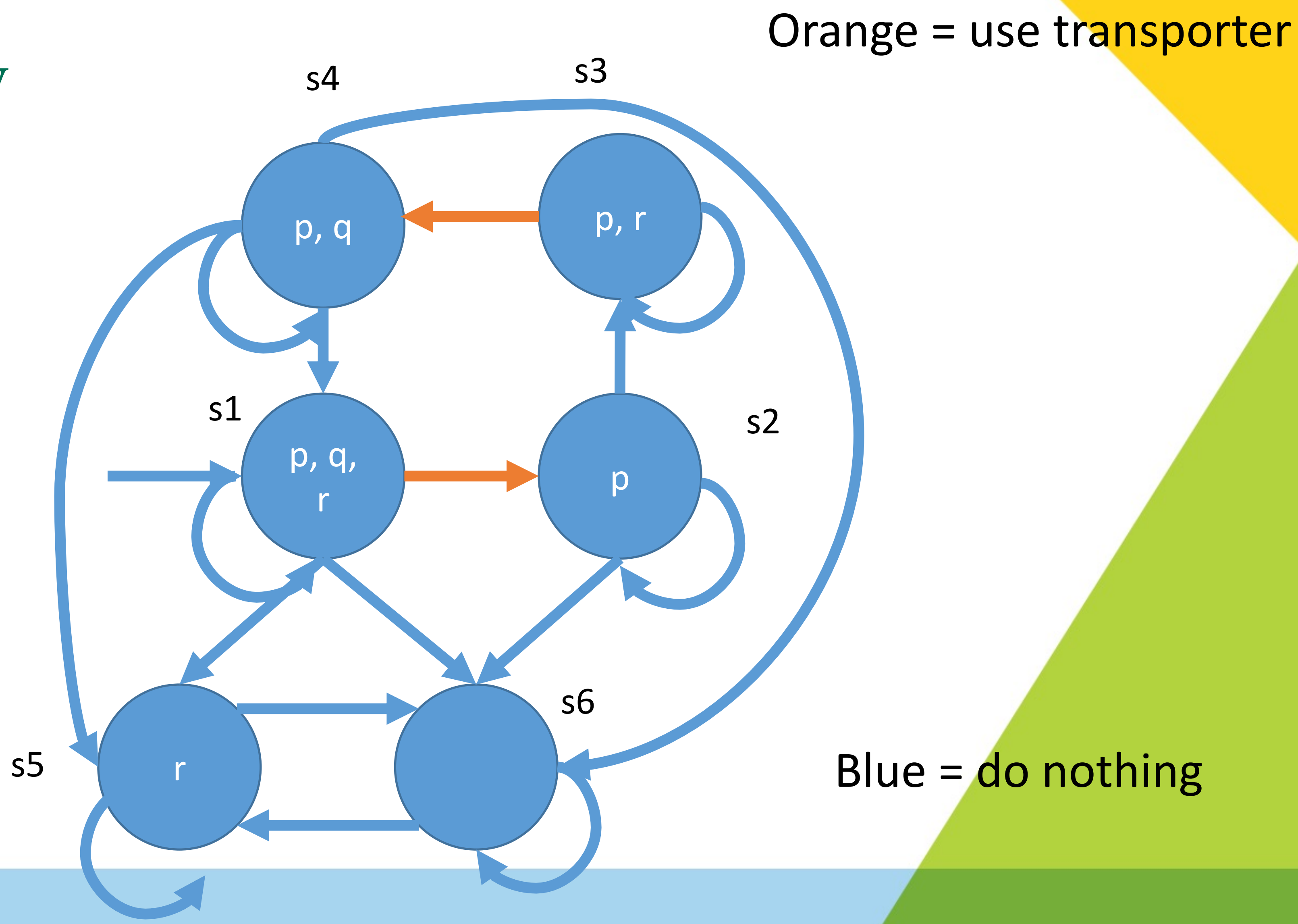
p = Yar alive

q = Yar on our ship

r = transporter ready

$$\phi = G[(r \wedge q \rightarrow X\neg q) \vee (r \wedge \neg q \rightarrow X\, q)]$$

- note: death happens



Transporter Policy

Can express ϕ as π:

$\phi = \; G[(r \wedge q \rightarrow X \neg q$

$\quad\quad \vee \, (r \wedge \neg q \rightarrow X \, q)]$

$\pi(s_1) = a_1$

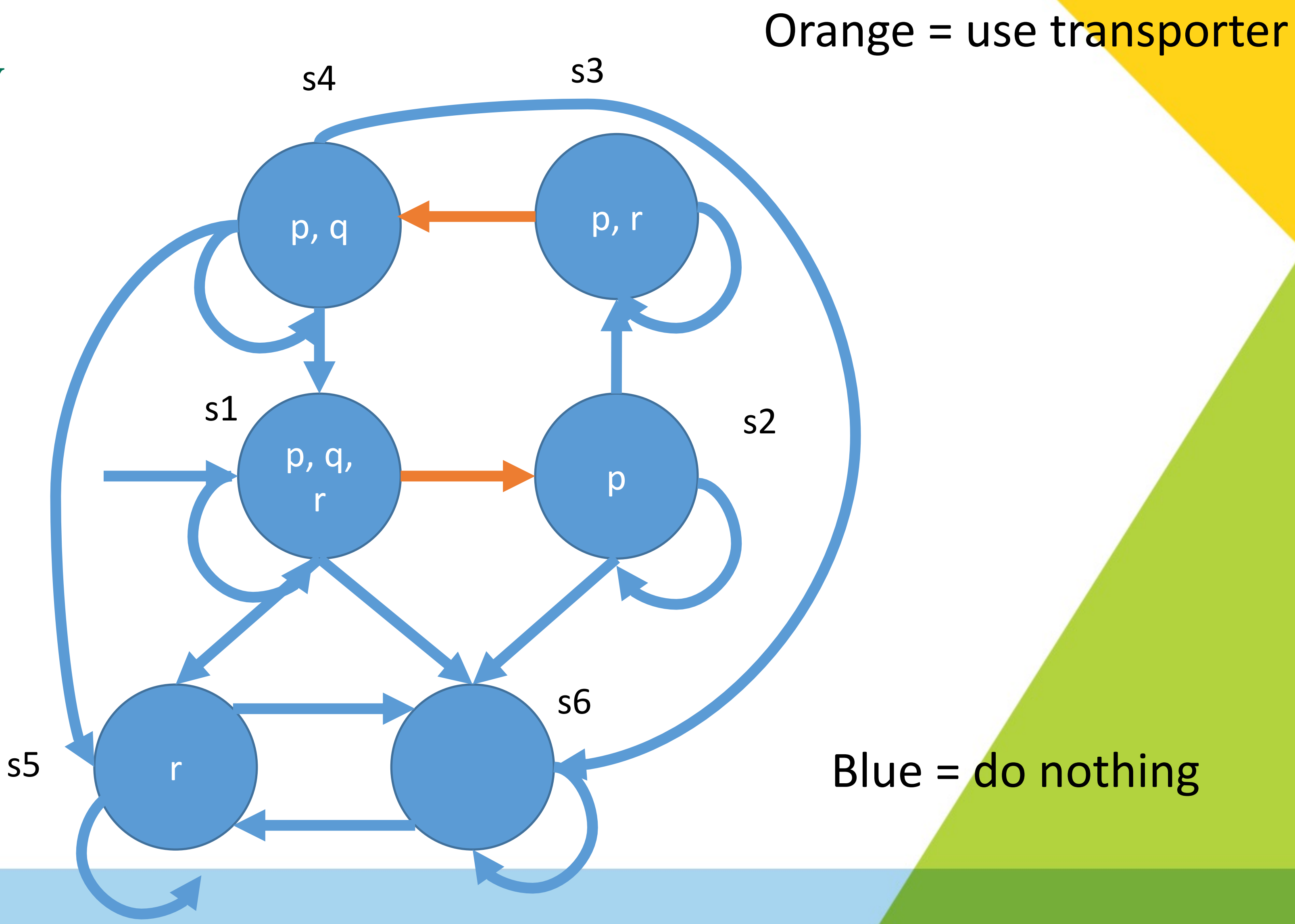
$\pi(s_2) = a_2$

$\pi(s_3) = a_1$

$\pi(s_4) = a_2$

$\pi(s_5) = a_1$

$\pi(s_6) = a_1$



Transporter Policy

Can express ϕ as π:

$$\phi = G[(r \wedge q \rightarrow X \neg q$$

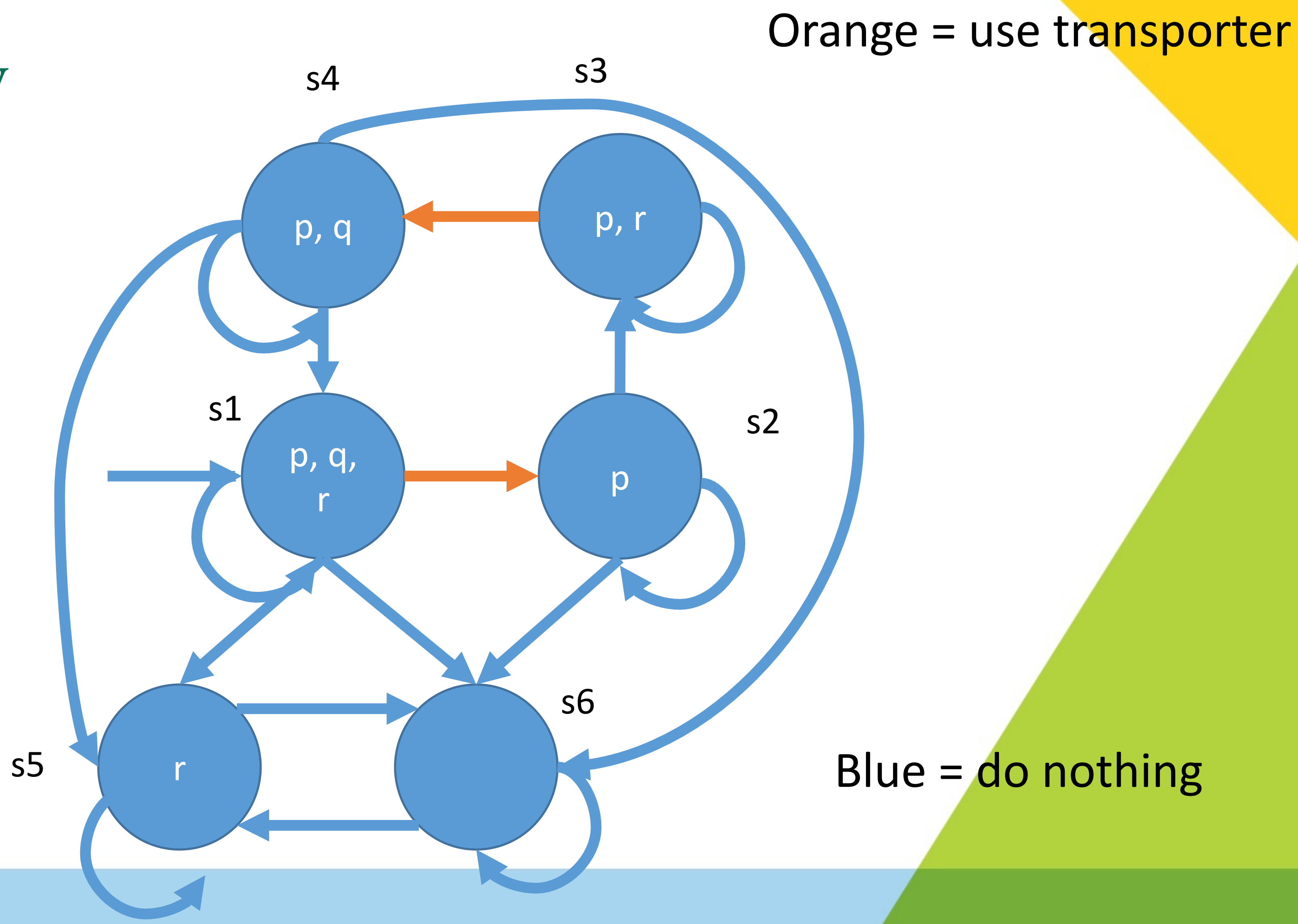
$$\vee (r \wedge \neg q \rightarrow X\, q)]$$

$\pi(a_1 \mid s_1) = 1$

$\pi(a_2 \mid s_1) = 0$

$\pi(a_1 \mid s_2) = 1$

…



Transporter Policy

Can express ϕ as π:

$$\phi = G[(r \wedge q \rightarrow F \neg q$$

$$\vee (r \wedge \neg q \rightarrow F\, q)]$$

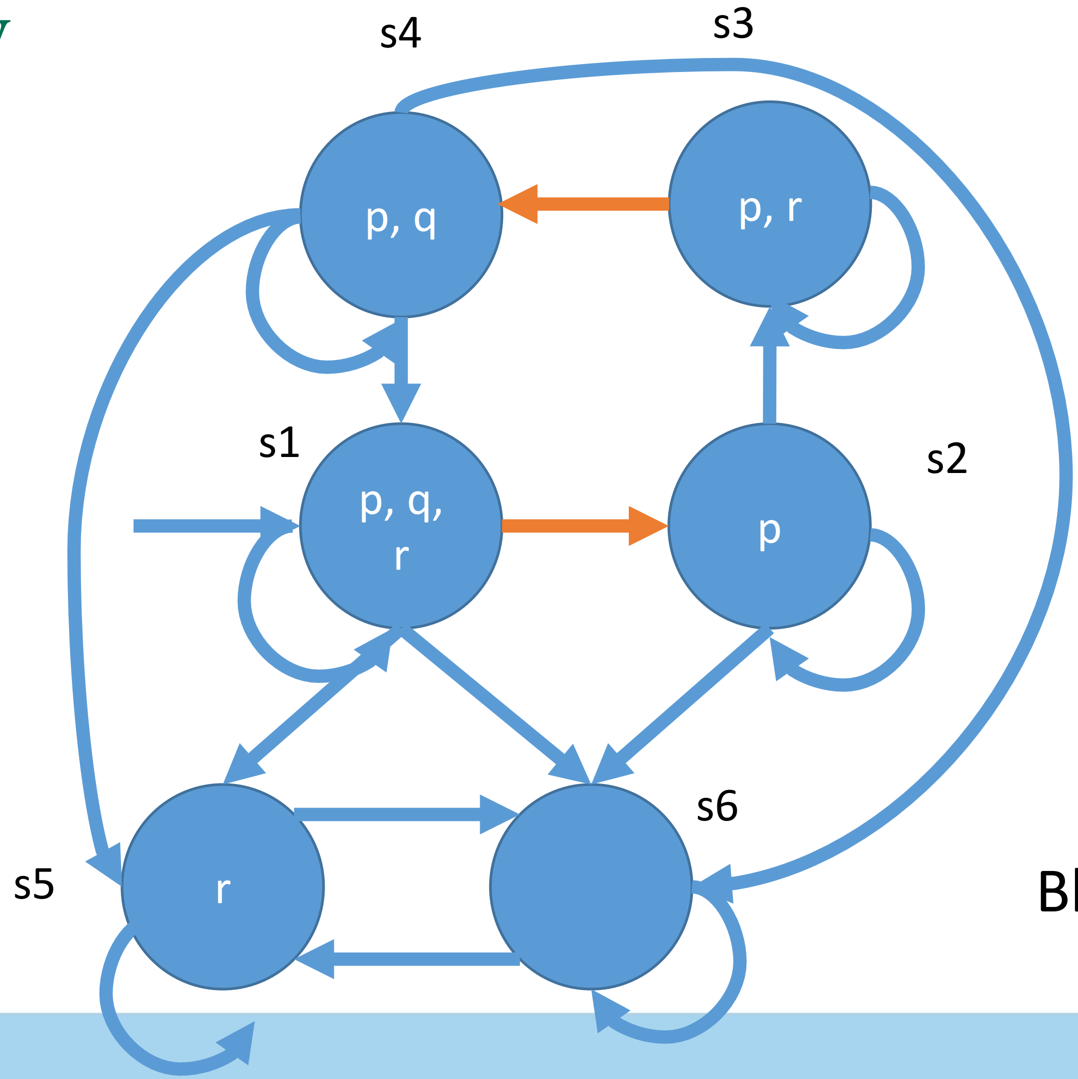
$\pi(a_1 \mid s_1) = 1$

$\pi(a_2 \mid s_1) = 0$

$\pi(a_1 \mid s_2) = 1$

…

Orange = use transporter



Blue = do nothing

Transporter Policy

Can express ϕ as π:

$\phi = G[(r \land q \rightarrow F \neg q$

$\lor (r \land \neg q \rightarrow F\ q)]$

$\pi(a_1 \mid s_1) = 0.5$

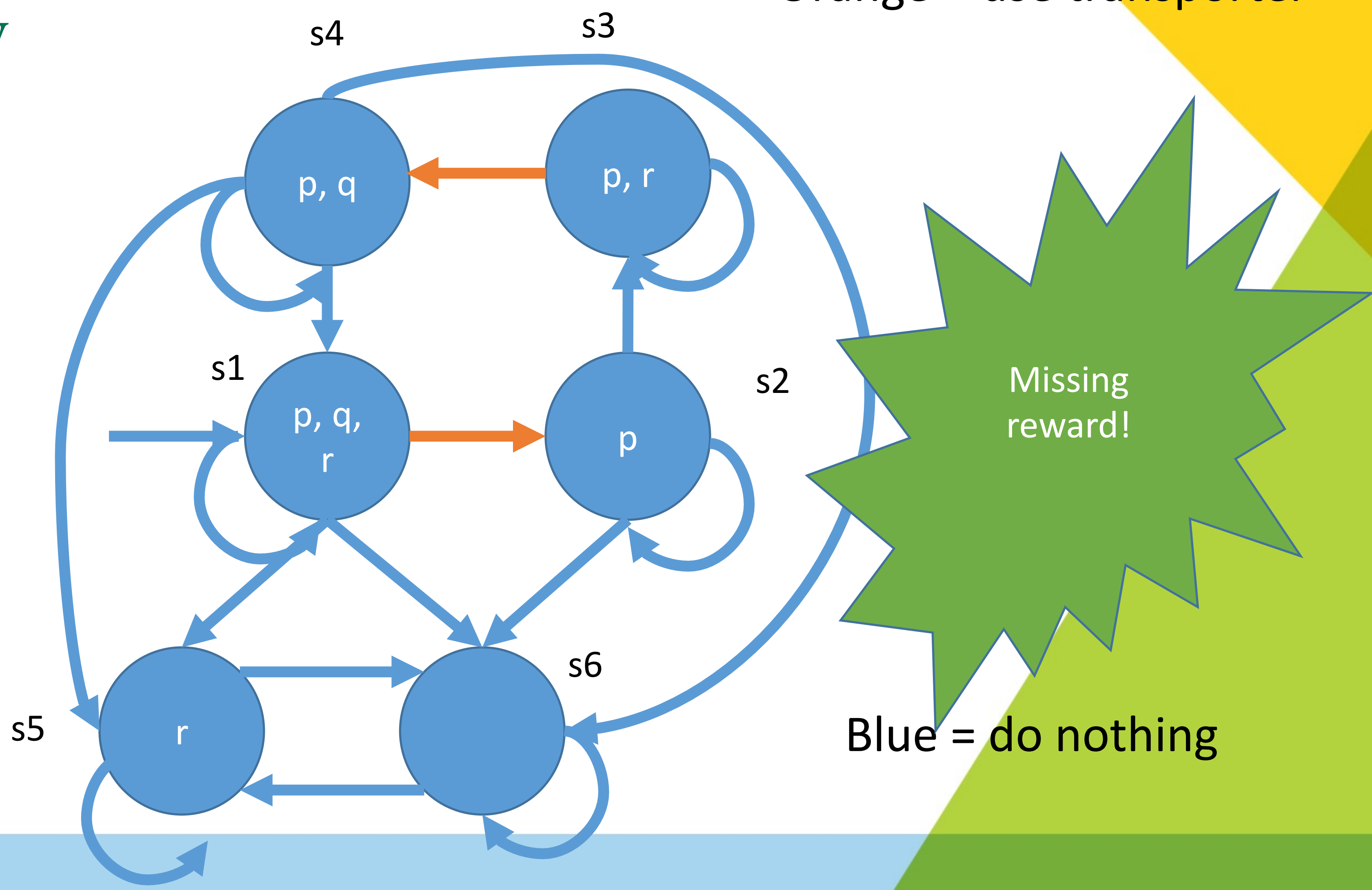
$\pi(a_2 \mid s_1) = 0.5$

Why might we do this?

What if

$P(s_6 \mid s_4) = 0.99$?

Orange = use transporter



Blue = do nothing

Transporter Policy

Can express ϕ as π:

$\phi = G[(r \wedge q \rightarrow F \neg q$

$\vee (r \wedge \neg q \rightarrow F\ q)]$

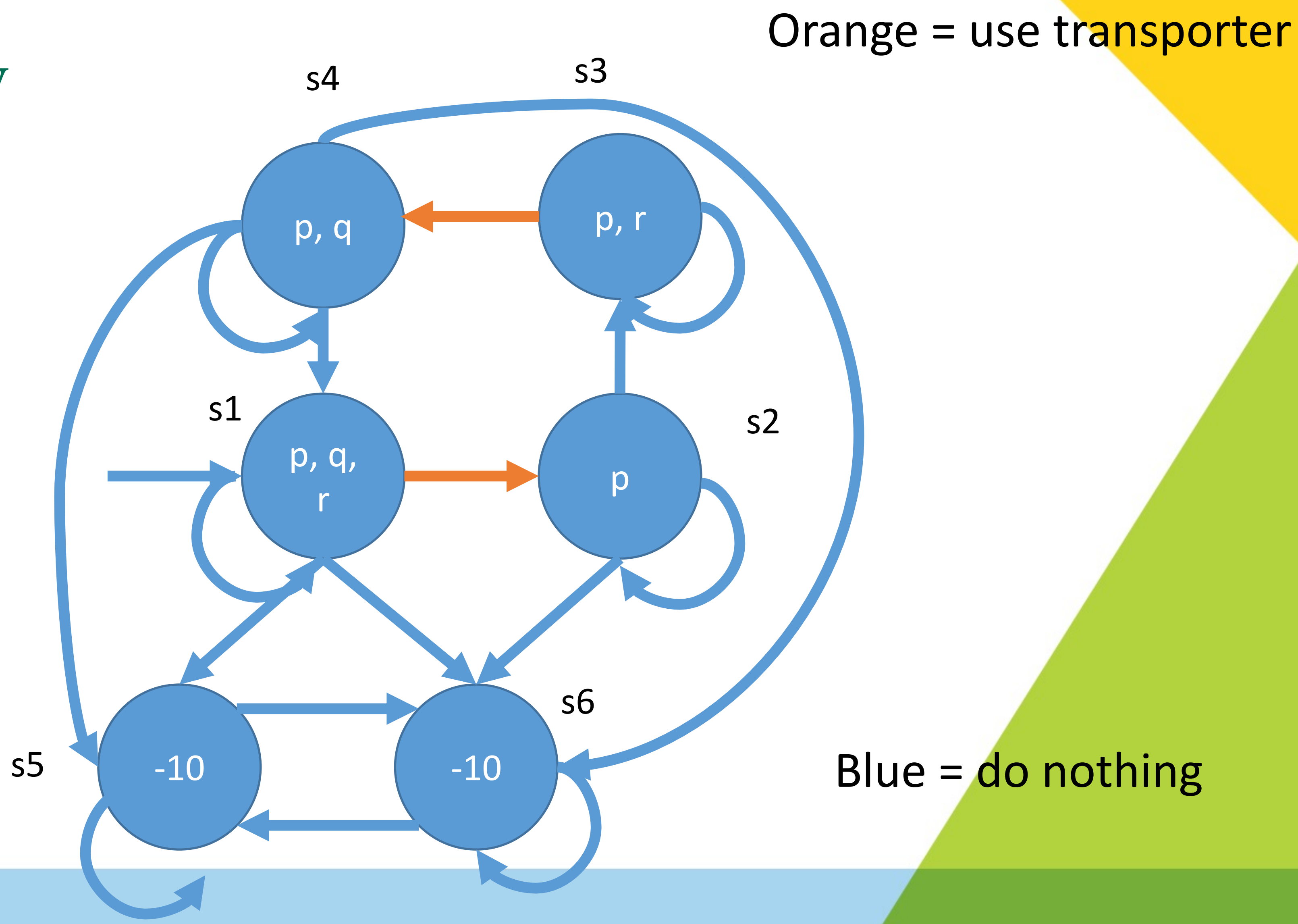
$\pi(a_1 \mid s_1) = 0.5$

$\pi(a_2 \mid s_1) = 0.5$

Why might we do this?

What if

$P(s_6 \mid s_4) = 0.99$?



Transporter Policy

Can express ϕ as π:

$\phi = \; G[(r \wedge q \rightarrow F\neg q$

$\vee \; (r \wedge \neg q \rightarrow F\; q)]$

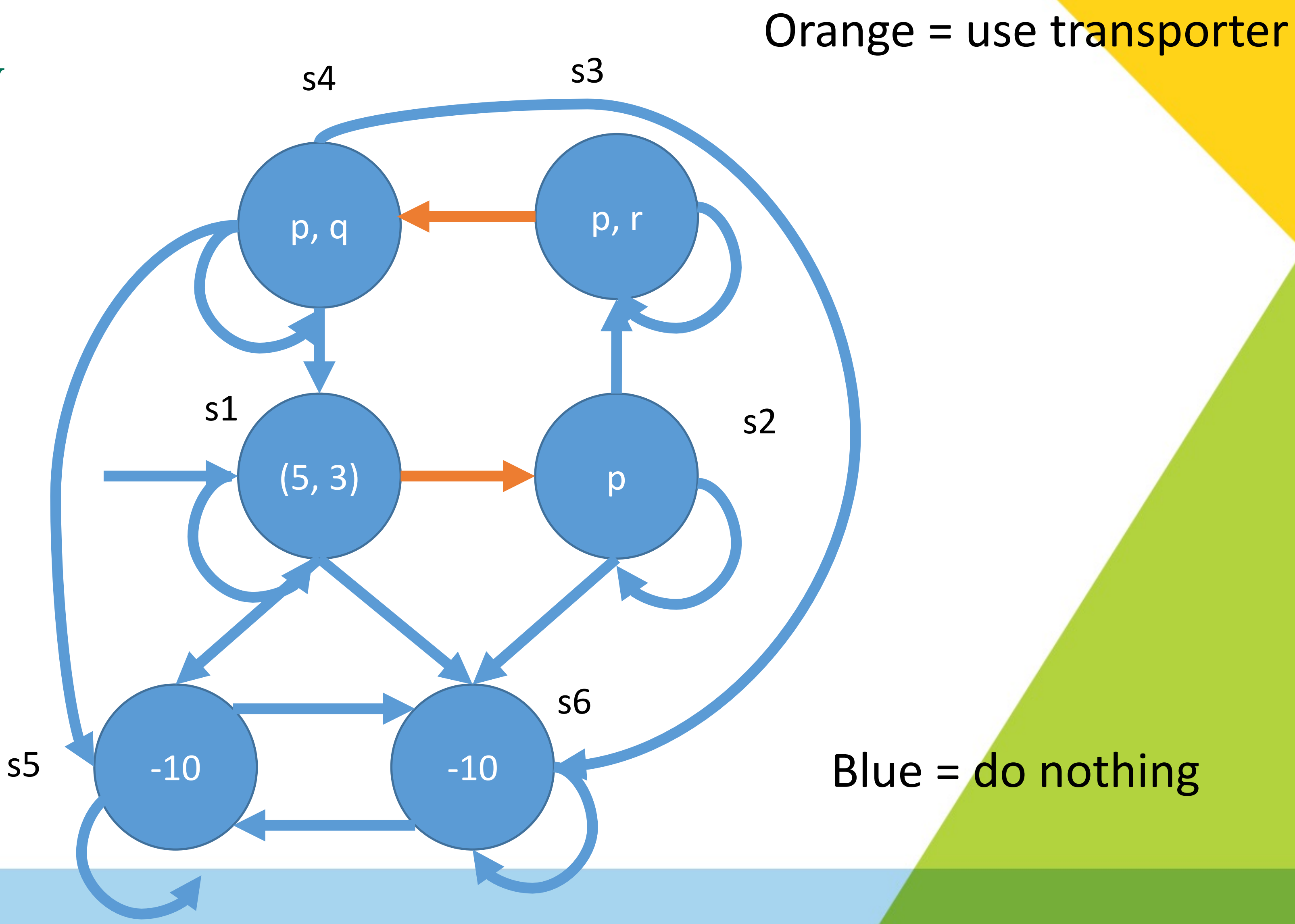
$\pi(\, a_1 \mid s_1\,) = 0.5$

$\pi(\, a_2 \mid s_1\,) = 0.5$

Why might we do this?

What if

$P(\, s_6 \mid s_4\,) = 0.99$?



Transporter Policy

Can express ϕ as π:

$\phi = G[(r \wedge q \rightarrow F \neg q$

$\vee (r \wedge \neg q \rightarrow F\ q)]$

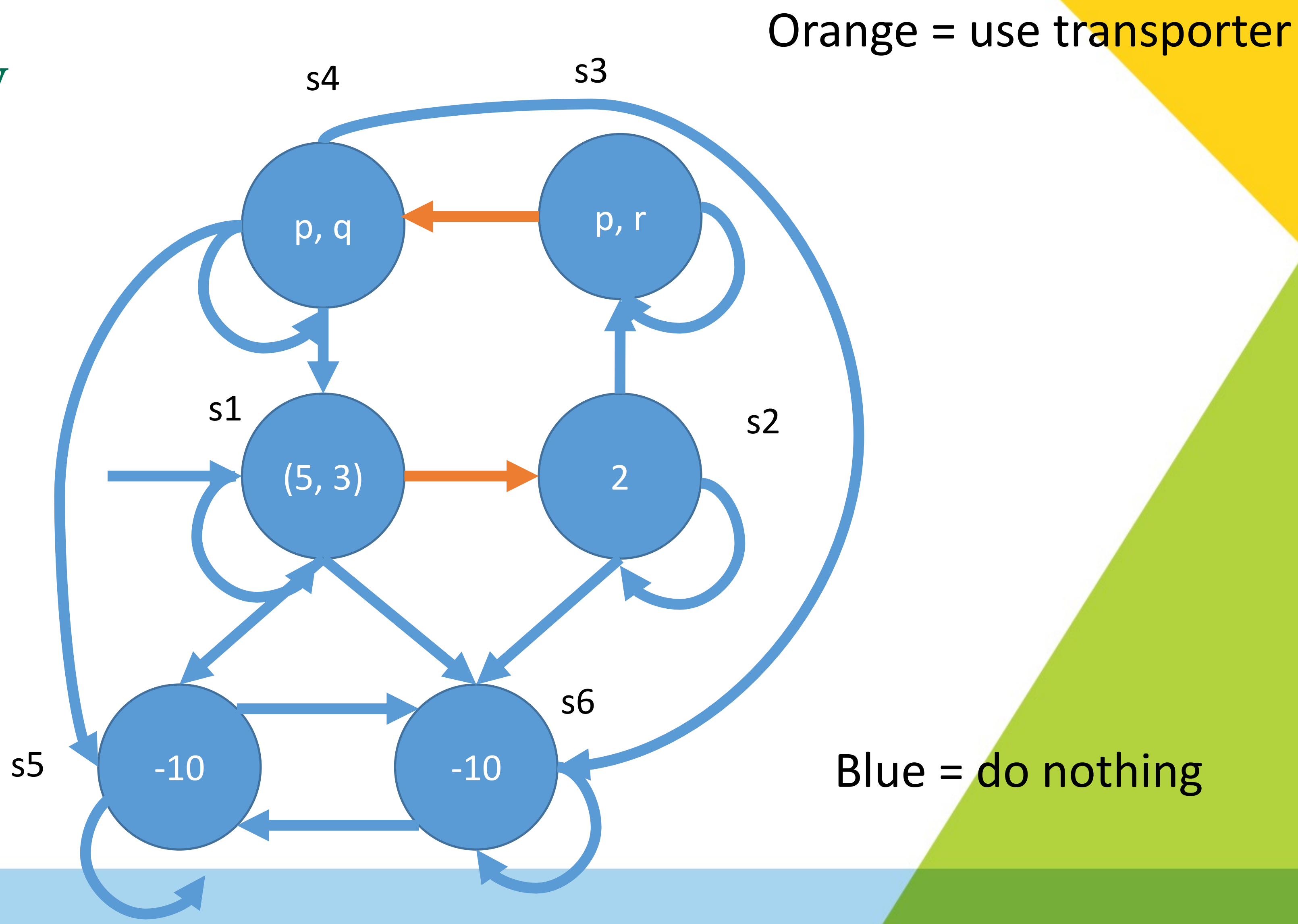
$\pi(a_1 \mid s_1) = 0.5$

$\pi(a_2 \mid s_1) = 0.5$

Why might we do this?

What if

$P(s_6 \mid s_4) = 0.99$?



Transporter Policy

Can express ϕ as π:

$\phi = G[(r \wedge q \rightarrow F\neg q$

$\vee (r \wedge \neg q \rightarrow F\, q)]$

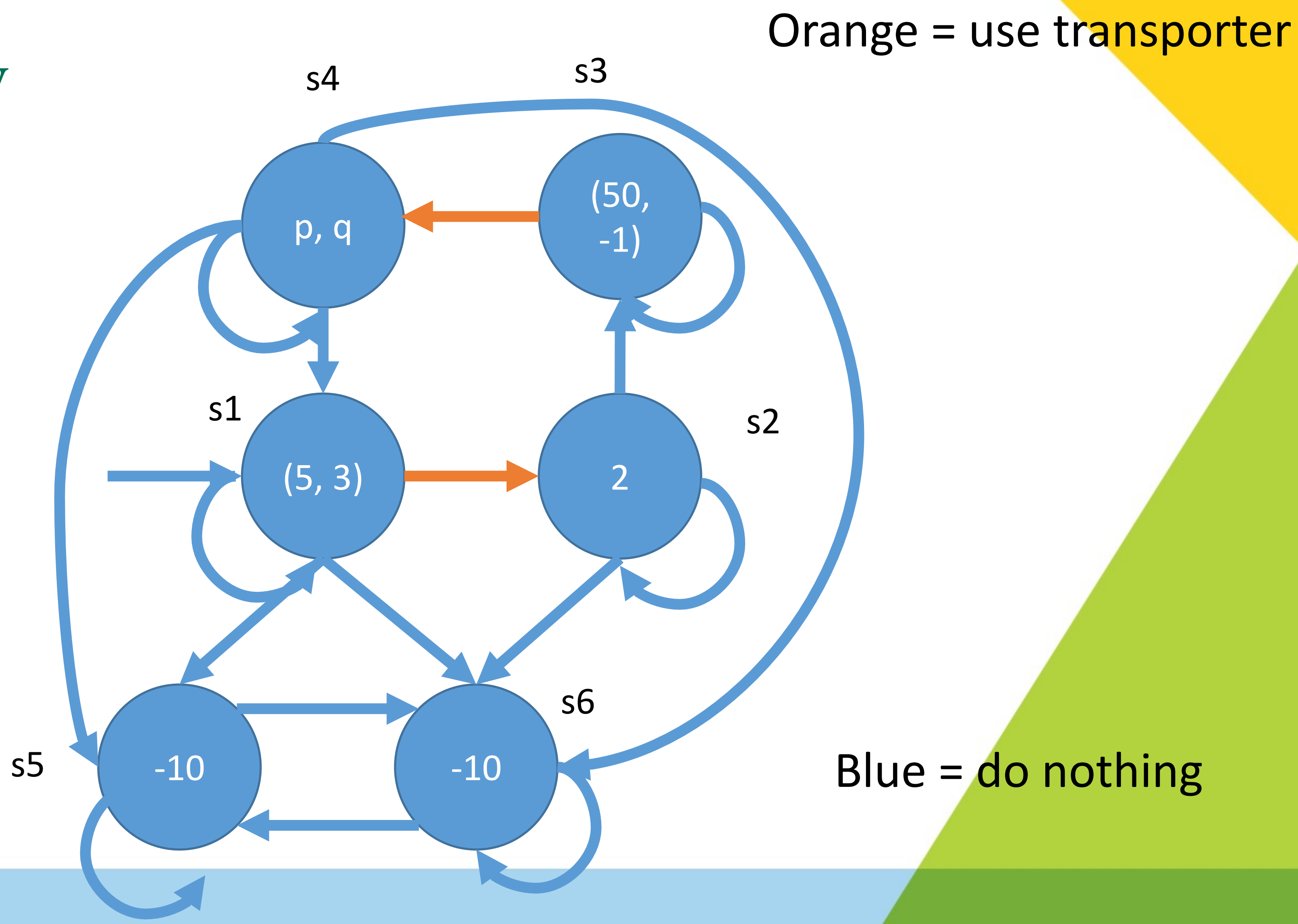
$\pi(a_1 \mid s_1) = 0.5$

$\pi(a_2 \mid s_1) = 0.5$

Why might we do this?

What if

$P(s_6 \mid s_4) = 0.99$?



Transporter Policy

Can express ϕ as π:

$\phi = \ G[(r \wedge q \rightarrow F \neg q$

$\vee (r \wedge \neg q \rightarrow F\ q)]$

$\pi(\, a_1 \mid s_1\,) = 0.5$

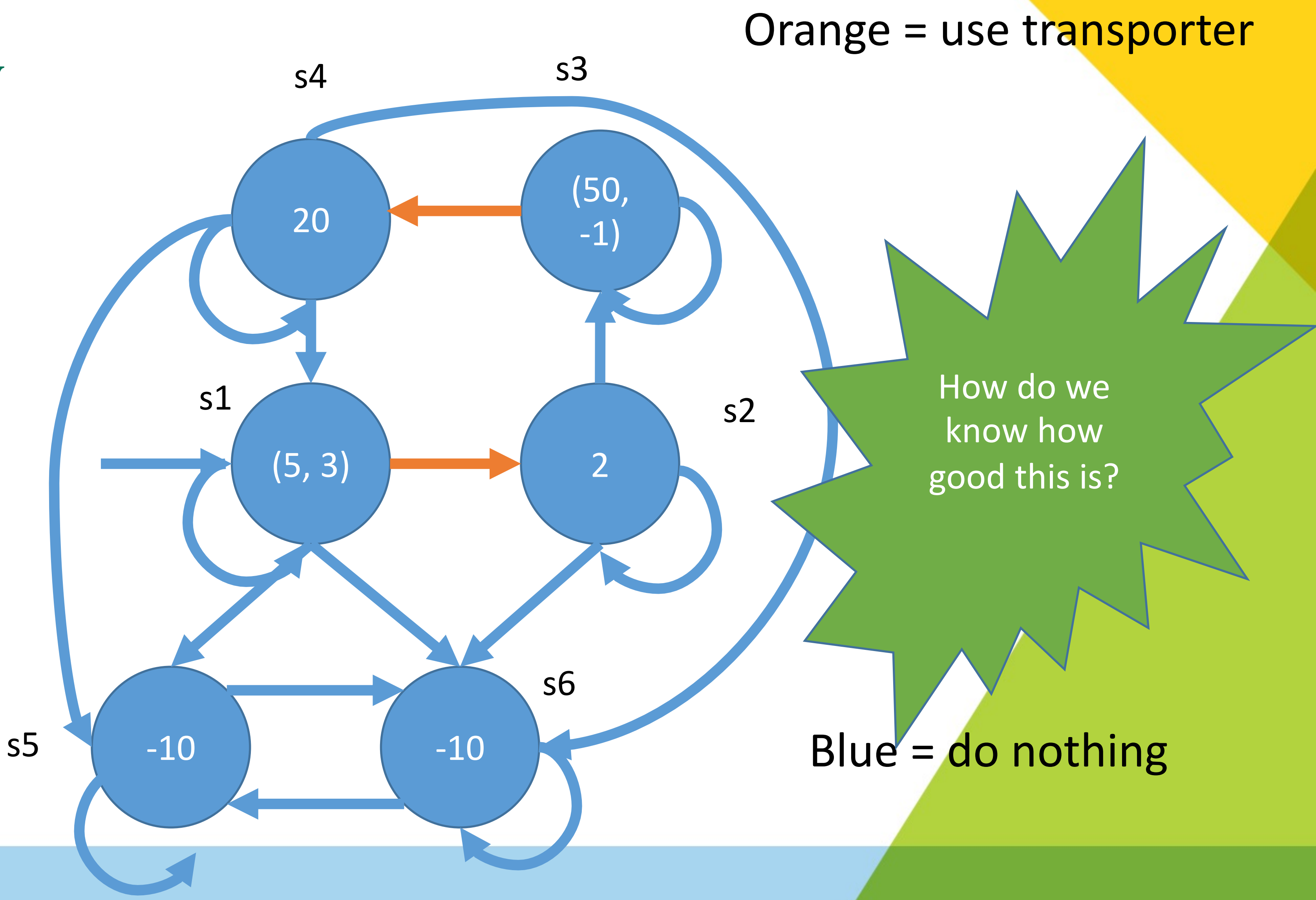
$\pi(\, a_2 \mid s_1\,) = 0.5$

Deterministic o/w

Why might we do this?

What if

$P(\, s_6 \mid s_4\,) = 0.99$?



Recall: decision theory computed expected utility over the tree

What is the equivalence construct for MDPs?

Value Functions

A *value function* $V^\pi : S \to \Re$ is a utility function *specific to a given policy* that assigns real numbers to each state in S.

This function is designed to be the expected "return":

$$V^\pi(s) = \sum_{t=0}^{\infty} \gamma^t E[R_t] = \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} \sum_{a \in A} R(s,a,s') P(s' \mid s,a) \pi(a \mid s)$$

$$= \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} \sum_{t=0}^{\infty} \gamma^t P(s' \mid s,a) R(s,a,s') = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P(s' \mid s,a)[R(s,a,s') + \gamma V^\pi(s')]$$

“Policy Evaluation” : Inferring the value function

Initialize $V_0^\pi(s) \coloneqq 0$ for all *s*

For *t* until convergence:

For each state s_{from}:

For each state s_{to}:

$$V_{t+1}^\pi(s_{from}) \coloneqq \sum_{s_{to}} \sum_a P(s_{to} \mid s_{from}, a)\pi(a \mid s_{from})[R(s_{from}, a, s_{to}) + \gamma V_t(s_{to})]$$

Exercise at home:

Finish assigning transition probabilities and compute the value of s1 after 3 iterations for the transporter problem

Given two policies, how do we compare?

- The value function defines a *partial order* over states.
- One policy is better than another iff, the values of all states of one policy are strictly better than another
- Can define strategies for *constructing better policies*