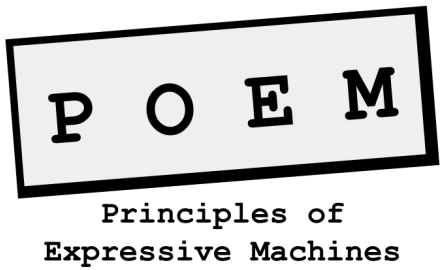
Keeping the Story Straight: A Comparison of Commitment Strategies for a Social Deduction Game

Markus Eger
meger@ncsu.edu

Chris Martens
crmarten@ncsu.edu

Principles of Expressive Machines Lab
Department of Computer Science
North Carolina State University



One Night Ultimate Werewolf

(Source: whatsericplaying.com)

One Night Ultimate Werewolf

- Competitive
- Players are secretly assigned roles
- Factions: Werewolves, Villagers
- Goal of the Villagers: Find the Werewolves
- Goal of the Werewolves: Avoid detection
- Night phase for actions, day phase for communication
- Players can lie
- Players' roles can change without their knowledge

One Night Ultimate Werewolf - Roles

- Seer: Look at another player's card

One Night Ultimate Werewolf - Roles

- Seer: Look at another player's card
- Robber: Exchange cards with another player, look at the new card

One Night Ultimate Werewolf - Roles

- Seer: Look at another player's card
- Robber: Exchange cards with another player, look at the new card
- Rascal: May exchange your two neighbors' cards

One Night Ultimate Werewolf - Roles

- Seer: Look at another player's card
- Robber: Exchange cards with another player, look at the new card
- Rascal: May exchange your two neighbors' cards
- Insomniac: Look at your own card

Challenges

- Hidden information
- How do we determine what to say?
- Real-time, arbitrary statements

Challenges

- Hidden information
- How do we determine what to say?
- ~~Real-time, arbitrary statements~~

Challenges

- Hidden information
- How do we determine what to say?
- ~~Real-time, arbitrary statements~~
- Intentional behavior

Intentionality

- Goal directed behavior
- "Choice with commitment" (Cohen and Levesque, 1990)
- Choose a story to tell
- Commit to telling that story until circumstances change

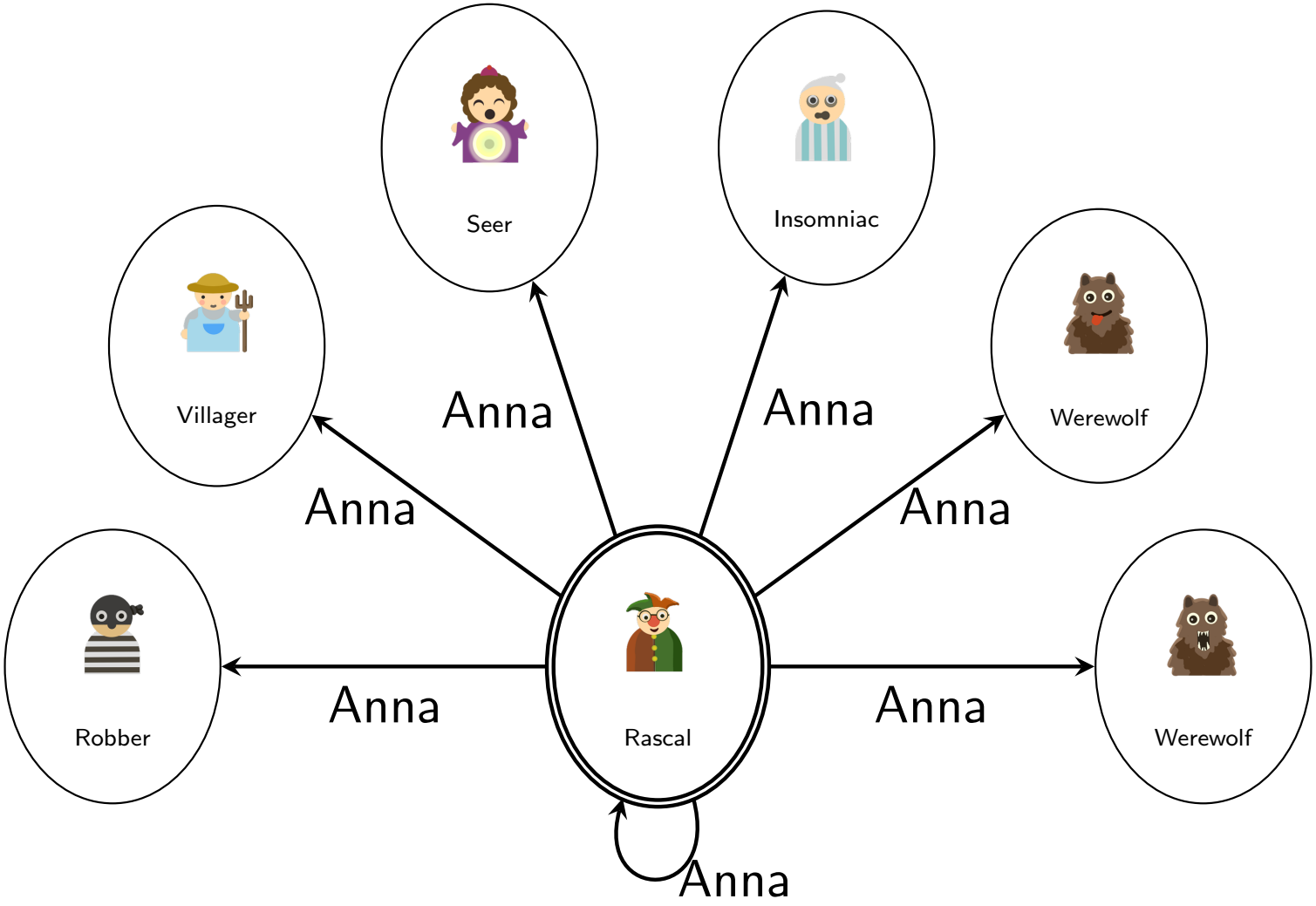
Intentionality

- Goal directed behavior
- "Choice with commitment" (Cohen and Levesque, 1990)
- Choose a story to tell
- Commit to telling that story until circumstances change (enough)

A simple scenario

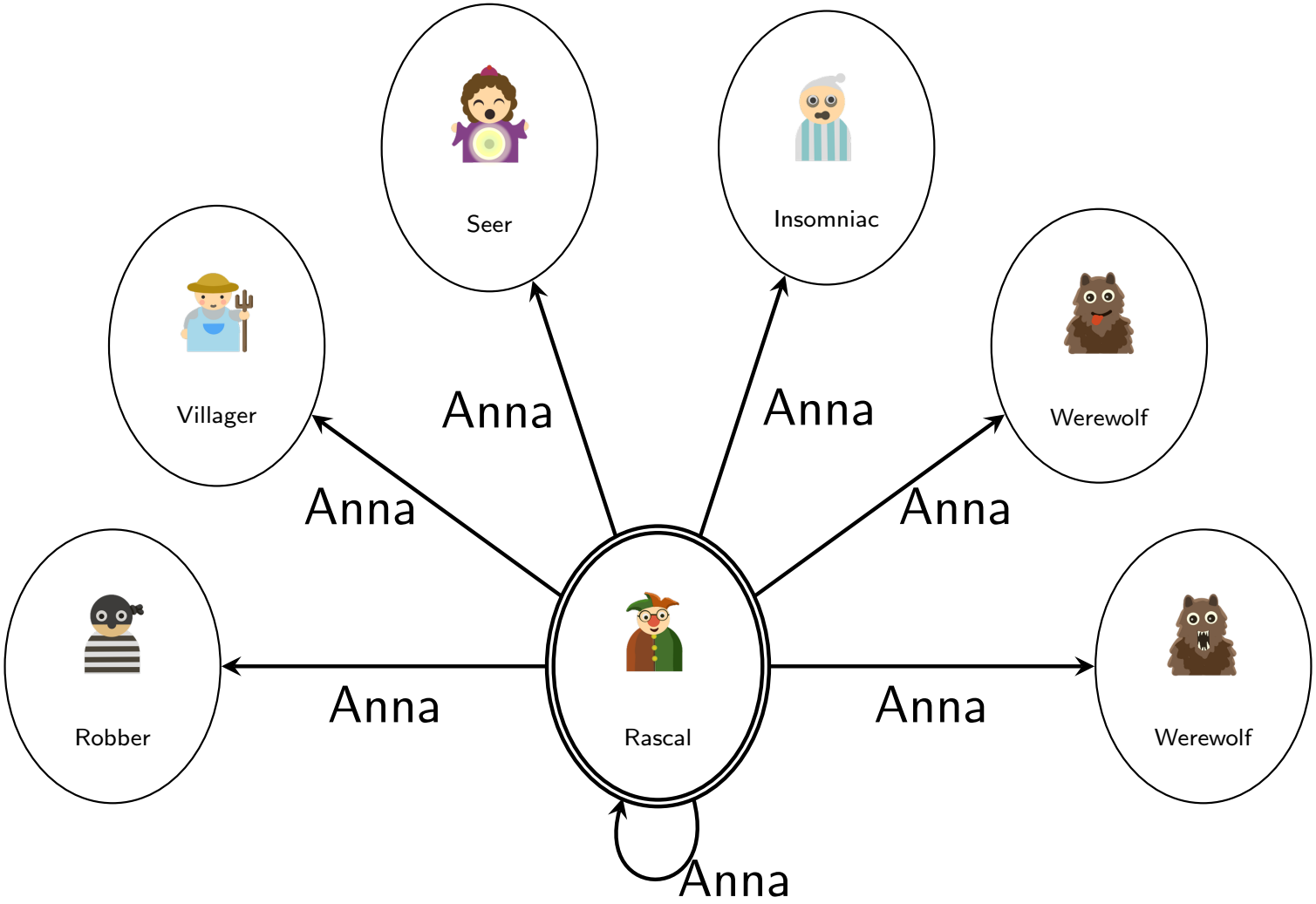
- Anna and Brian play One Night Ultimate Werewolf
- Anna got a Villager card
- Brian got the Rascal card
- But Anna does not know which card Brian has

Possible Worlds Quality



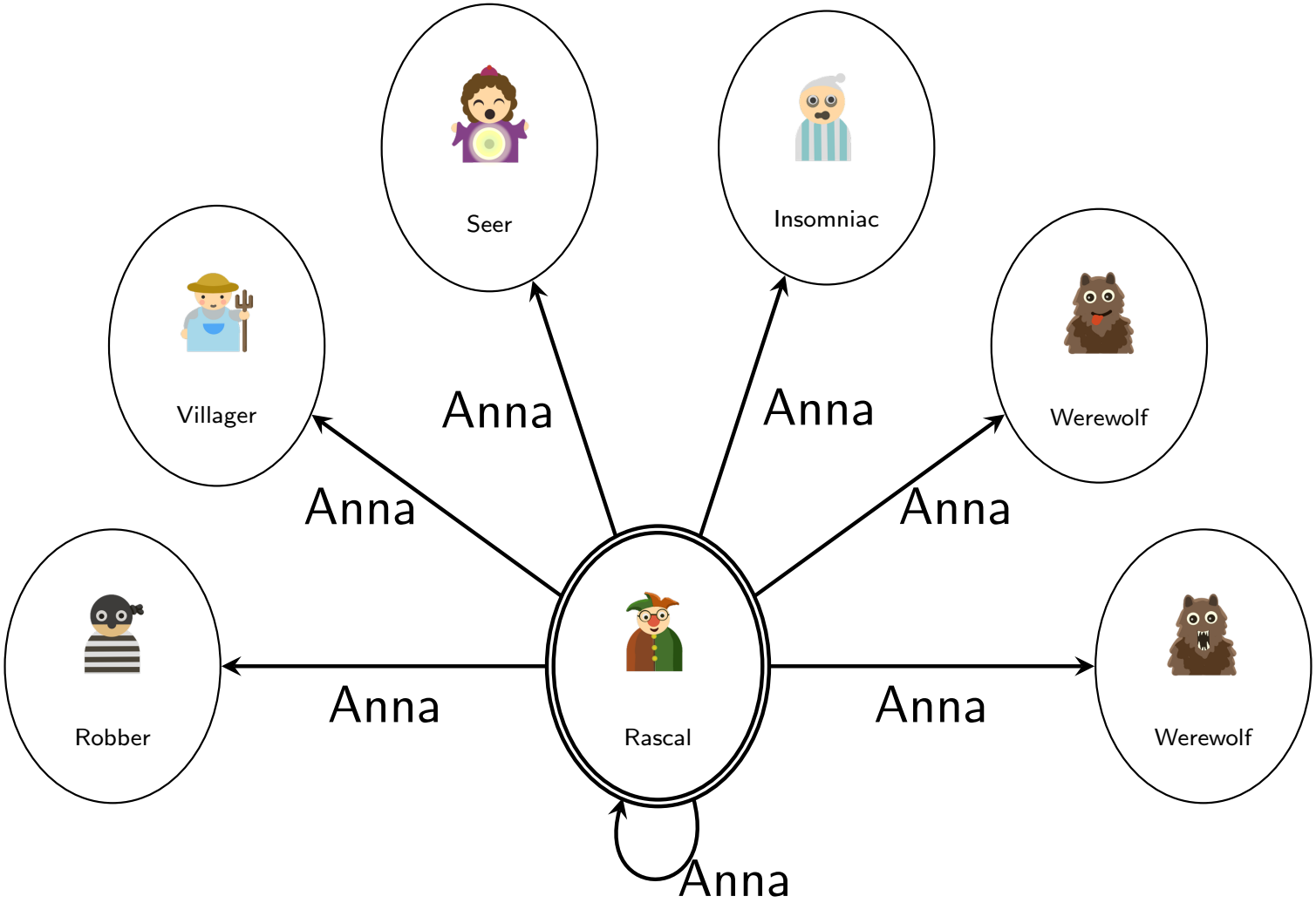
Worlds Anna considers possible

Possible Worlds Quality



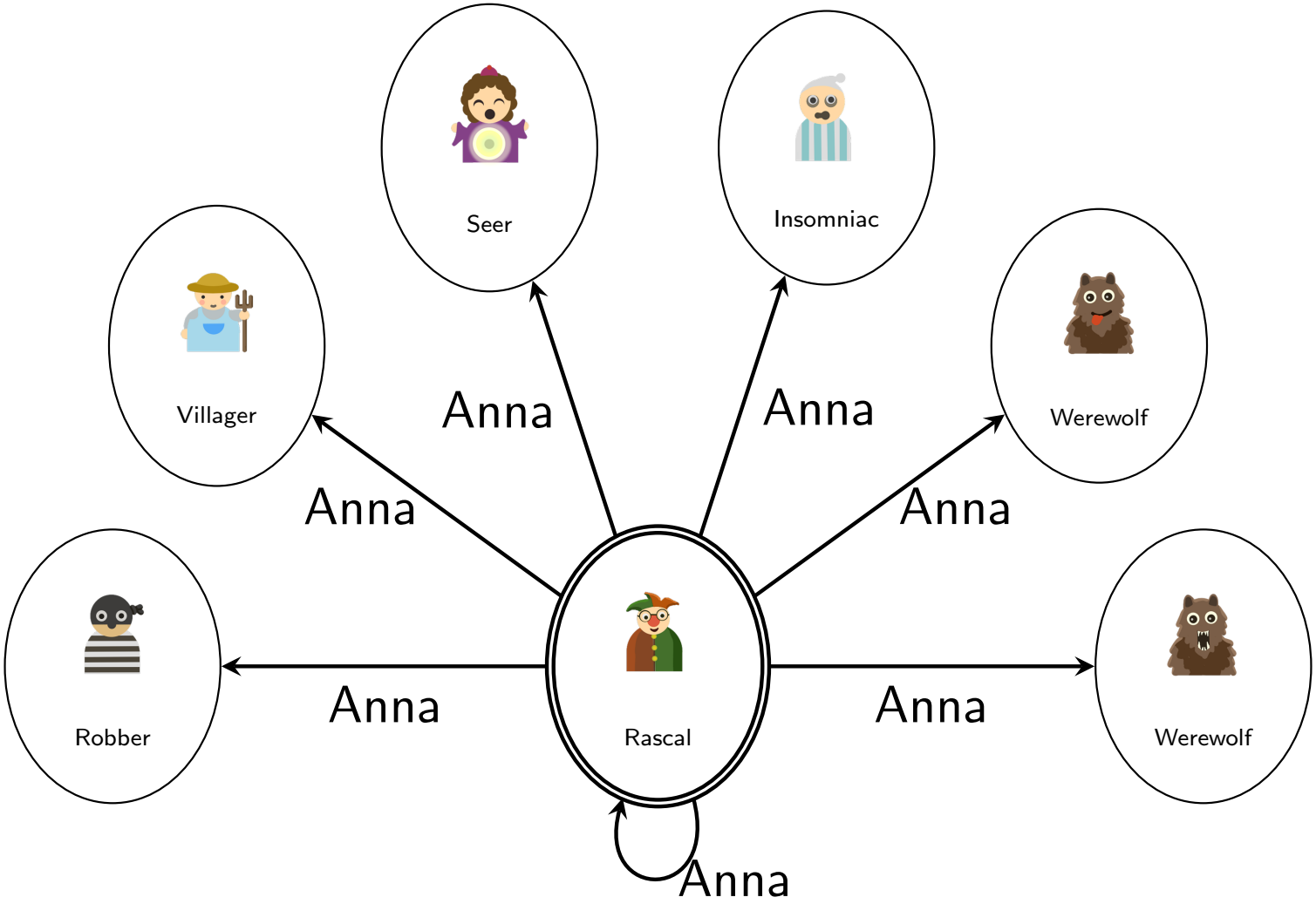
Anna does not believe that Brian is a Werewolf

Possible Worlds Quality



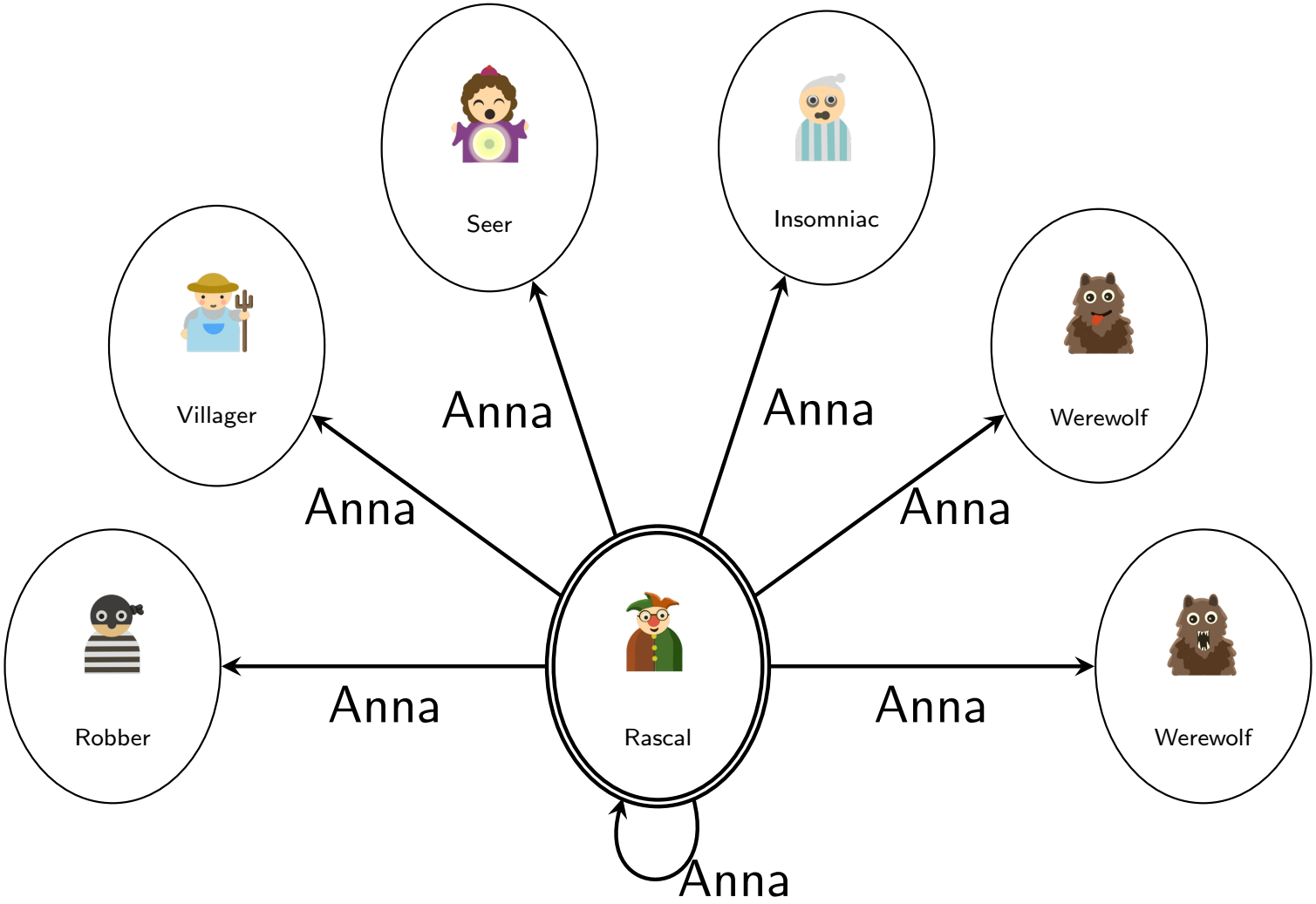
Anna does not believe that Brian is not a Werewolf either

Possible Worlds Quality



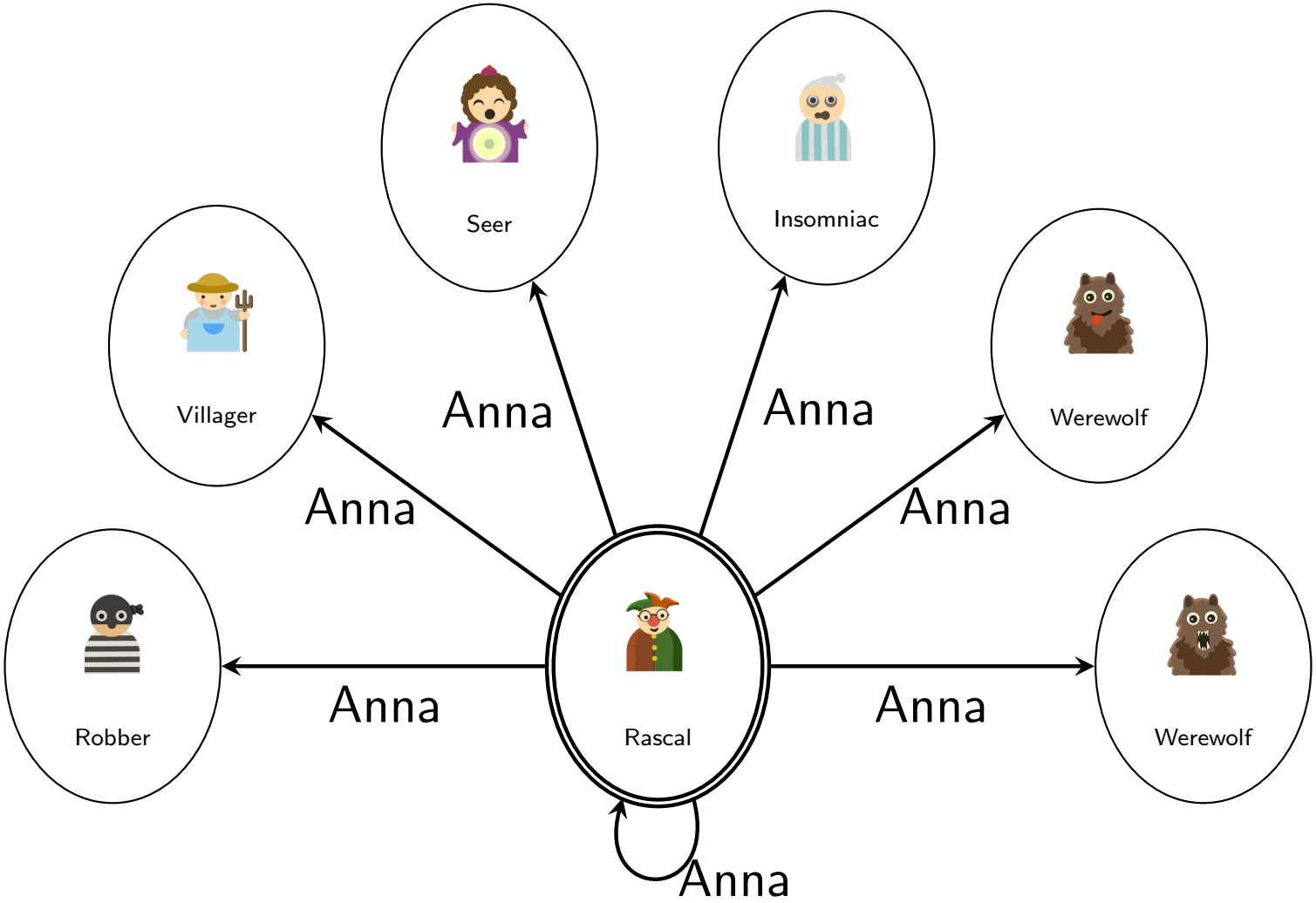
Brian is a Werewolf in 2 out of 7 worlds Anna considers possible!

Possible Worlds Quality



Quality of Belief that Brian is a Werewolf: $\frac{2}{7}$

Possible Worlds Quality



Quality of Belief that Brian is the Rascal: $\frac{1}{7}$

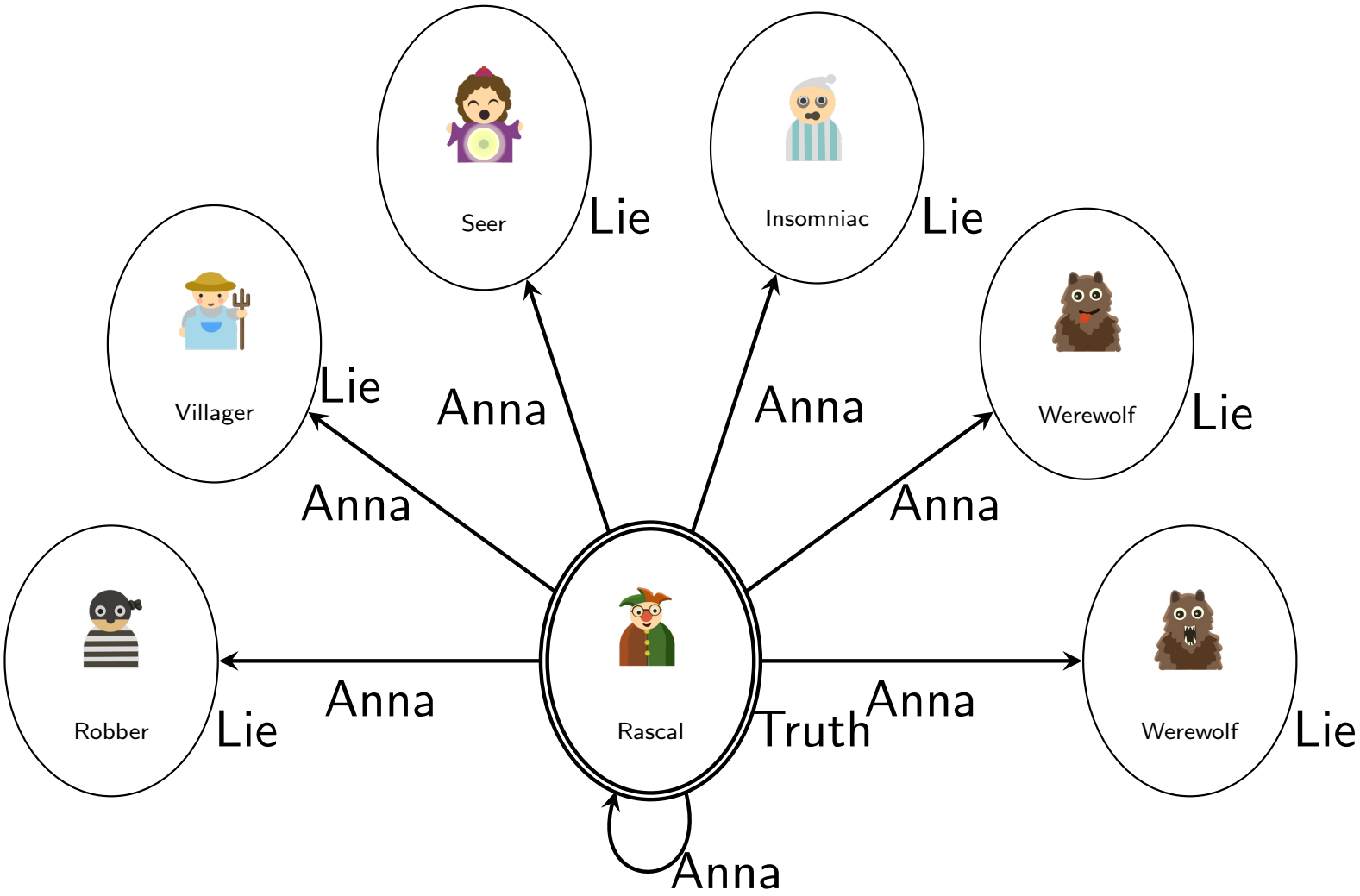
What about communication?

- Brian says (truthfully) that he is the Rascal
- What is Anna supposed to do with this information?
- Anna does not know that Brian is telling the truth

What about communication?

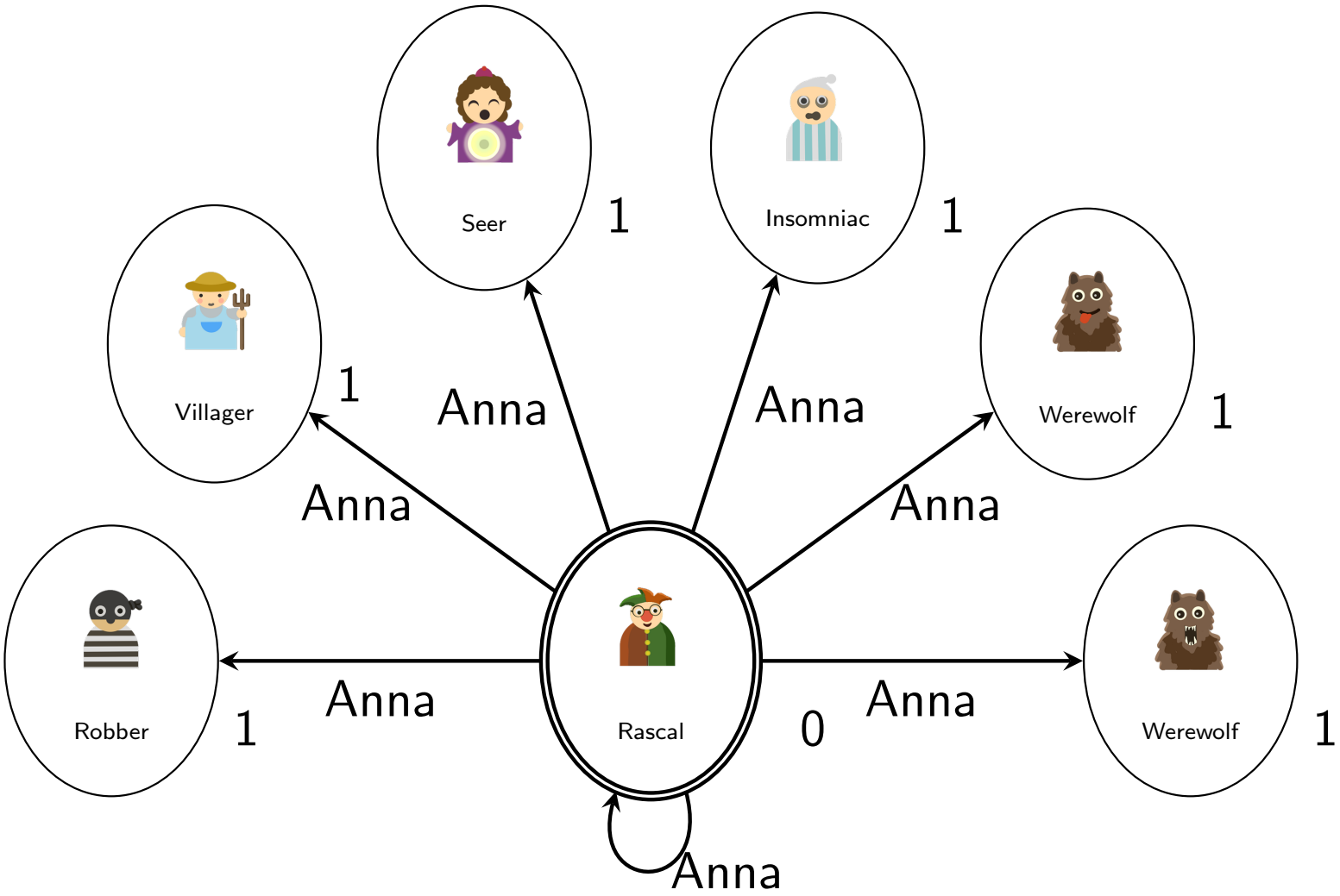
- Brian says (truthfully) that he is the Rascal
- What is Anna supposed to do with this information?
- Anna does not know that Brian is telling the truth
- But what if he is?

Belief Quality



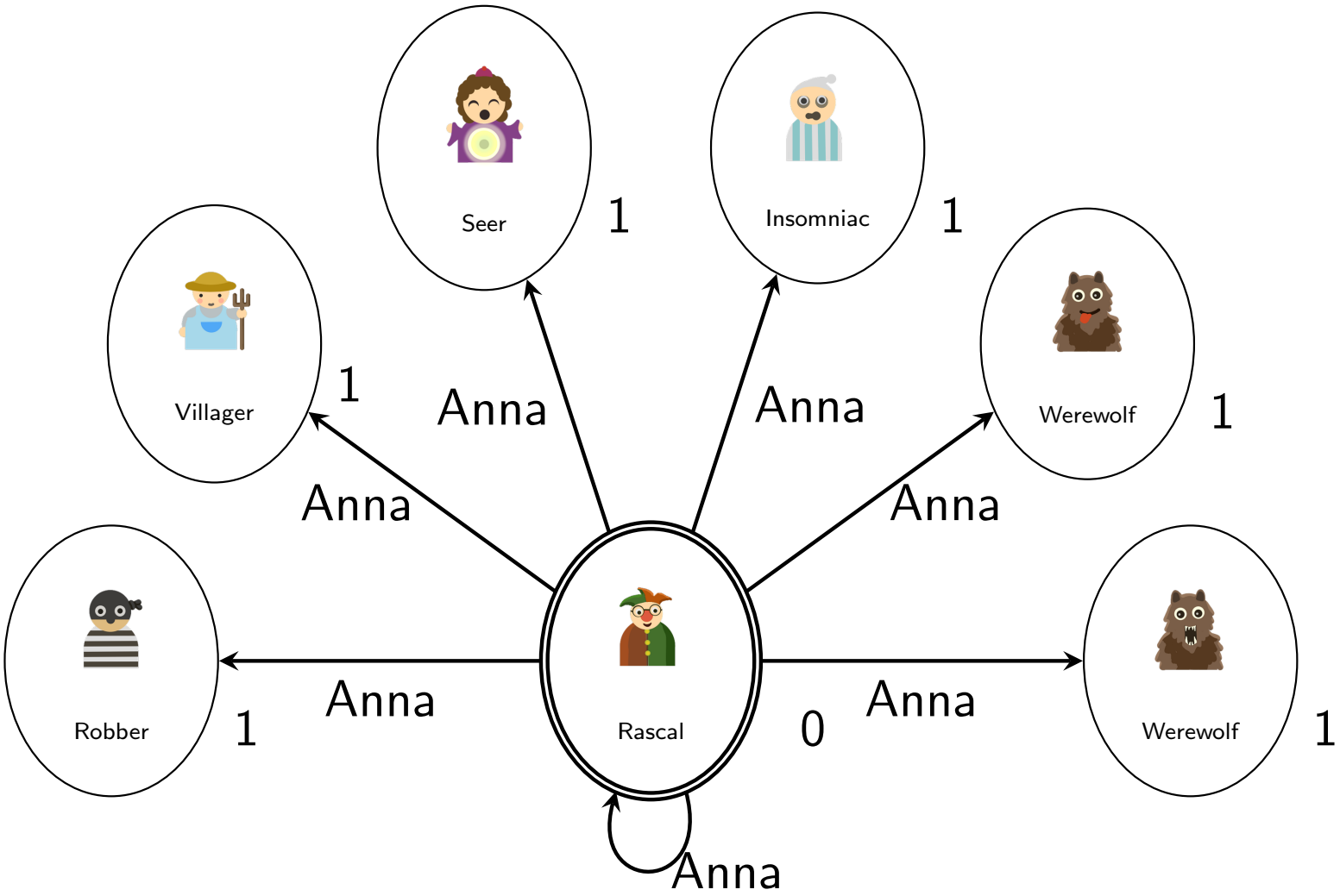
Annotate worlds

Belief Quality



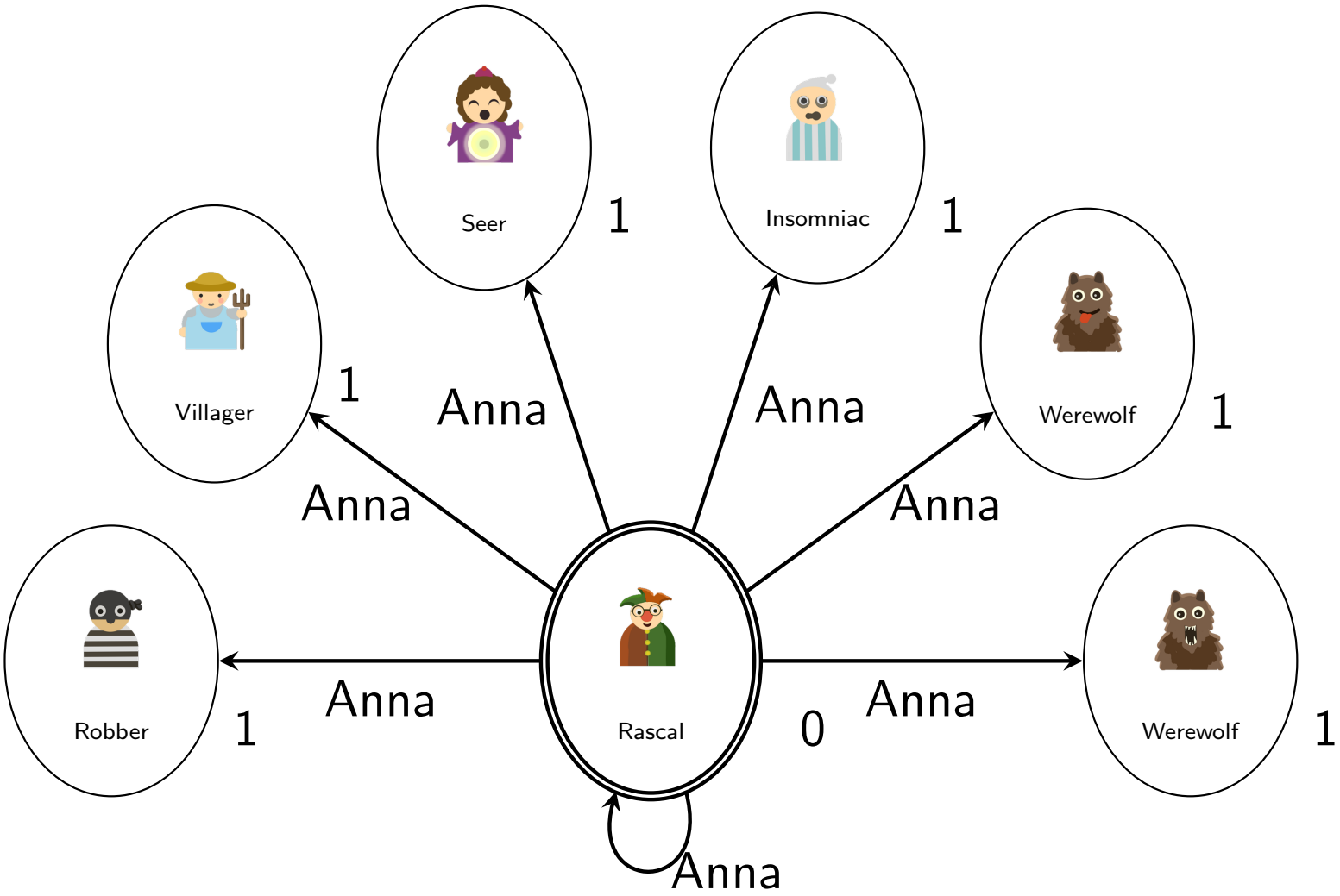
Higher numbers mean more lies, less likely to be true.

Belief Quality



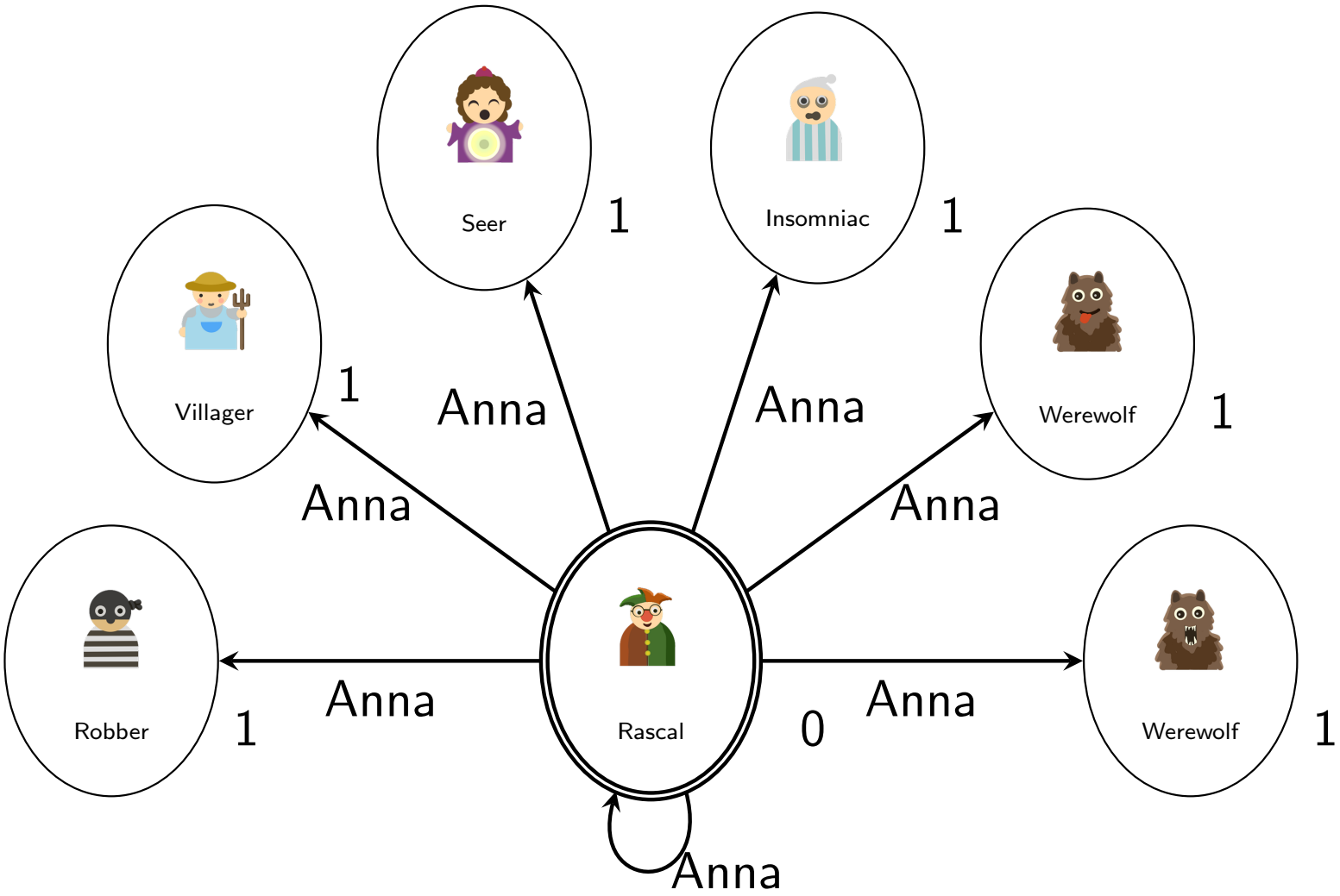
Instead of counting each world equally, use factor $\frac{1}{1+w}$.

Belief Quality



Weighted Quality of the Belief that Brian is a Werewolf: $\frac{1}{4}$

Belief Quality



Weighted Quality of the Belief that Brian is the Rascal: $\frac{1}{4}$

How do the agents use this concept?

- Brian says (truthfully) that he is the Rascal
- The outcome is *not* that Anna now believes him
- Anna thinks it is more likely that he is the Rascal, though

How do the agents use this concept?

- Brian says (truthfully) that he is the Rascal
- The outcome is *not* that Anna now believes him
- Anna thinks it is more likely that he is the Rascal, though
- Instead of planning to reach a goal: plan to maximize weighted quality

Intentional Agents

- Form intentions from candidate goals
- Calculate plan to get close to goal
- Decide when to drop/revise intentions

Intentional Agents - Intention selection and revision

- Pick the plan that reaches its goal most closely
- Record the expected weighted quality w of that plan
- If a new plan reaches a goal with weighted quality w', change plans iff:

$$w' \geq \alpha \cdot w$$

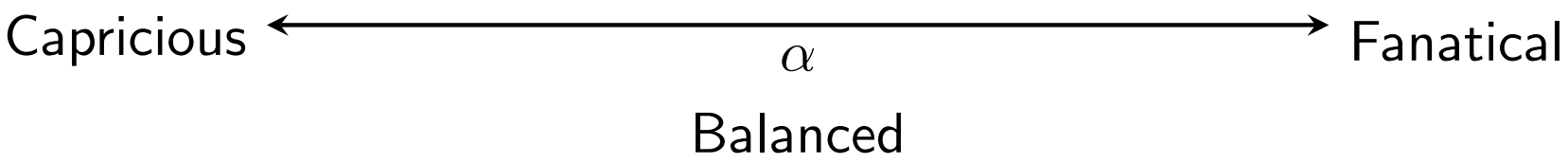
- Changing α leads to different *levels of commitment*
  $\alpha = 0$: capricious
  $\alpha = \infty$: fanatical

Intentional Agents - Intention selection and revision

- Pick the plan that reaches its goal most closely
- Record the expected weighted quality w of that plan
- If a new plan reaches a goal with weighted quality w', change plans iff:

$$w' \geq \alpha \cdot w$$

- Changing α leads to different *levels of commitment*
  $\alpha = 0$: capricious
  $\alpha = \infty$: fanatical



Goals

- Werewolf: Pretend to be someone else
- Knows a Werewolf: Convince other players of suspicion
- Villager: Convince other players of ones innocence

Balanced game: No swapping

- *Anna* starts as the Werewolf
- No one can take her Werewolf card away

Balanced game: No swapping

- *Anna* starts as the Werewolf
- No one can take her Werewolf card away
- New information should not affect plans

Balanced game: No swapping

- *Anna* starts as the Werewolf
- No one can take her Werewolf card away
- New information should not affect plans
- Agents with different levels of commitment

Balanced game: No swapping

- *Anna* starts as the Werewolf
- No one can take her Werewolf card away
- New information should not affect plans
- Agents with different levels of commitment
- Anna's win rate: $50.25\% \pm 1.5\%$

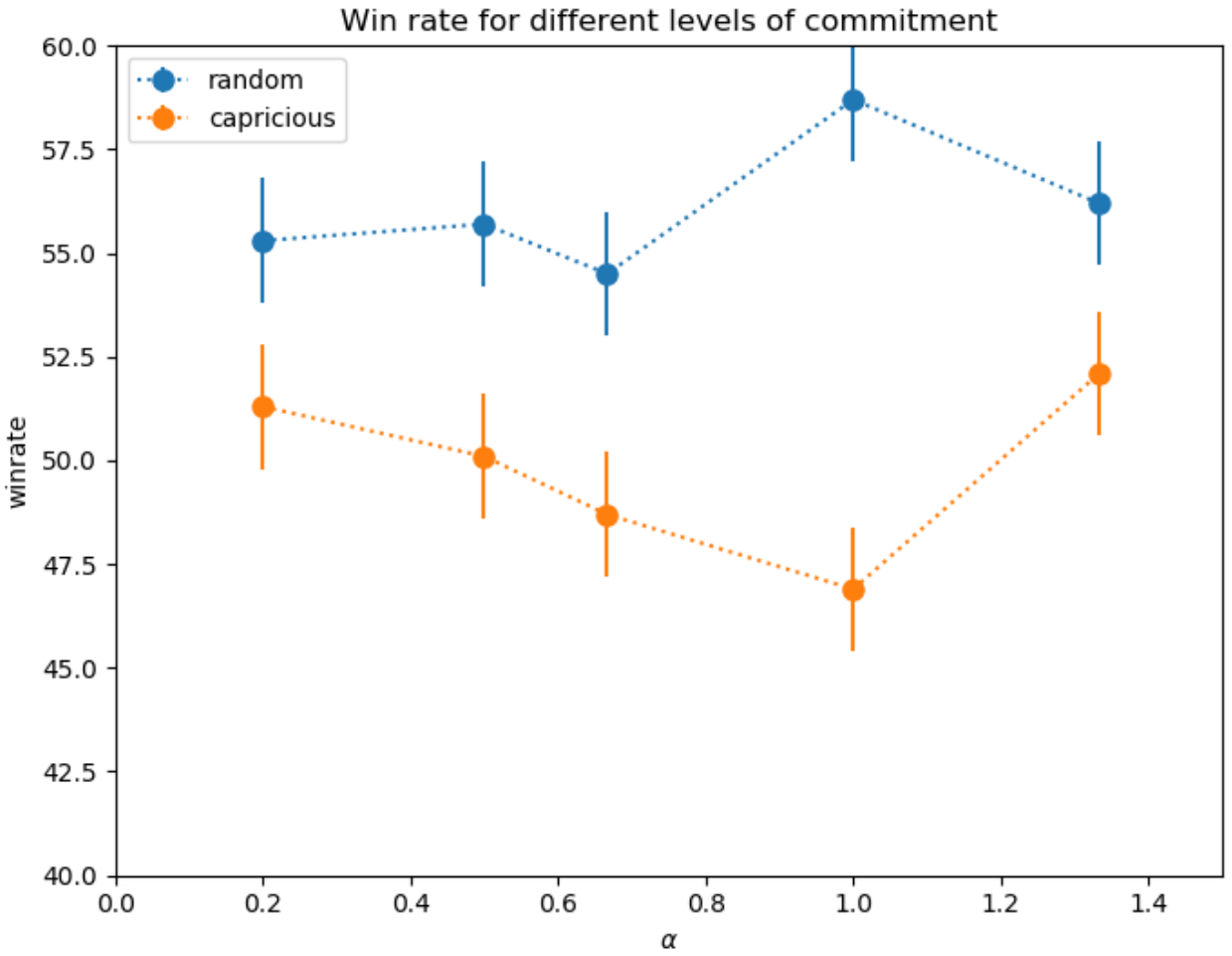
Balanced game: Swapping

- *Anna* starts as the Werewolf
- Her Werewolf card may be swapped

Balanced game: Swapping

- *Anna* starts as the Werewolf
- Her Werewolf card may be swapped
- New information may affect the agents' plans

Balanced game: Swapping



Conclusion

- Agent commitment matters, $\pm 5\%$ win rate
- Direction of change depends on opponent's strategy

Conclusion

- Agent commitment matters, $\pm 5\%$ win rate
- Direction of change depends on opponent's strategy
- Weighted quality of belief
- You may not be able to *reach* your communicative goal

Thank you

Thank you for your attention

meger@ncsu.edu
@yawgmoth46
http://github.com/yawgmoth/ONUW

Thank you

Thank you for your attention

meger@ncsu.edu
@yawgmoth46
http://github.com/yawgmoth/ONUW

Special thanks to Hui-Yin "Helen" Wu for the Werewolf roles art