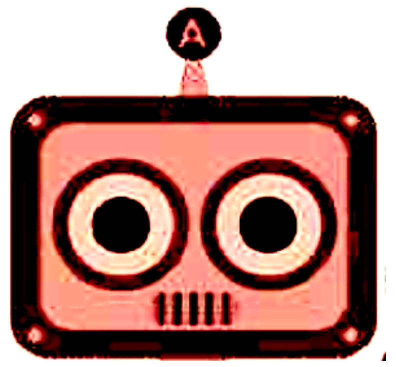
Shapley Values, Attention Flows, and *Faithful Explanations*

"Tony" Runzhe Yang

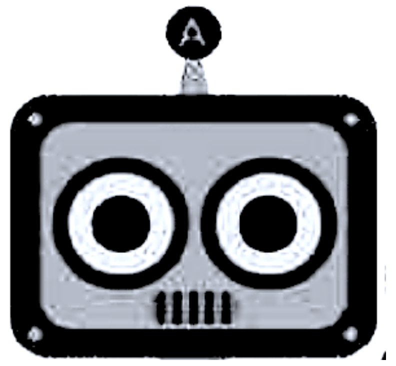
https://runzhe-yang.science

Shapley Values: Fair Division in Cooperative Game



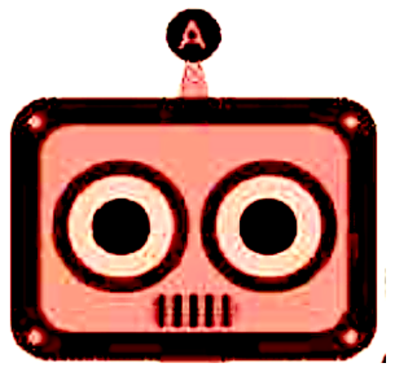
Alice (alone)

5 cookies 🍪 / hour

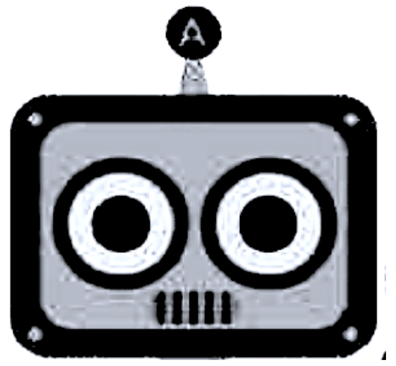


Bob (alone)

3 cookies 🍪 / hour



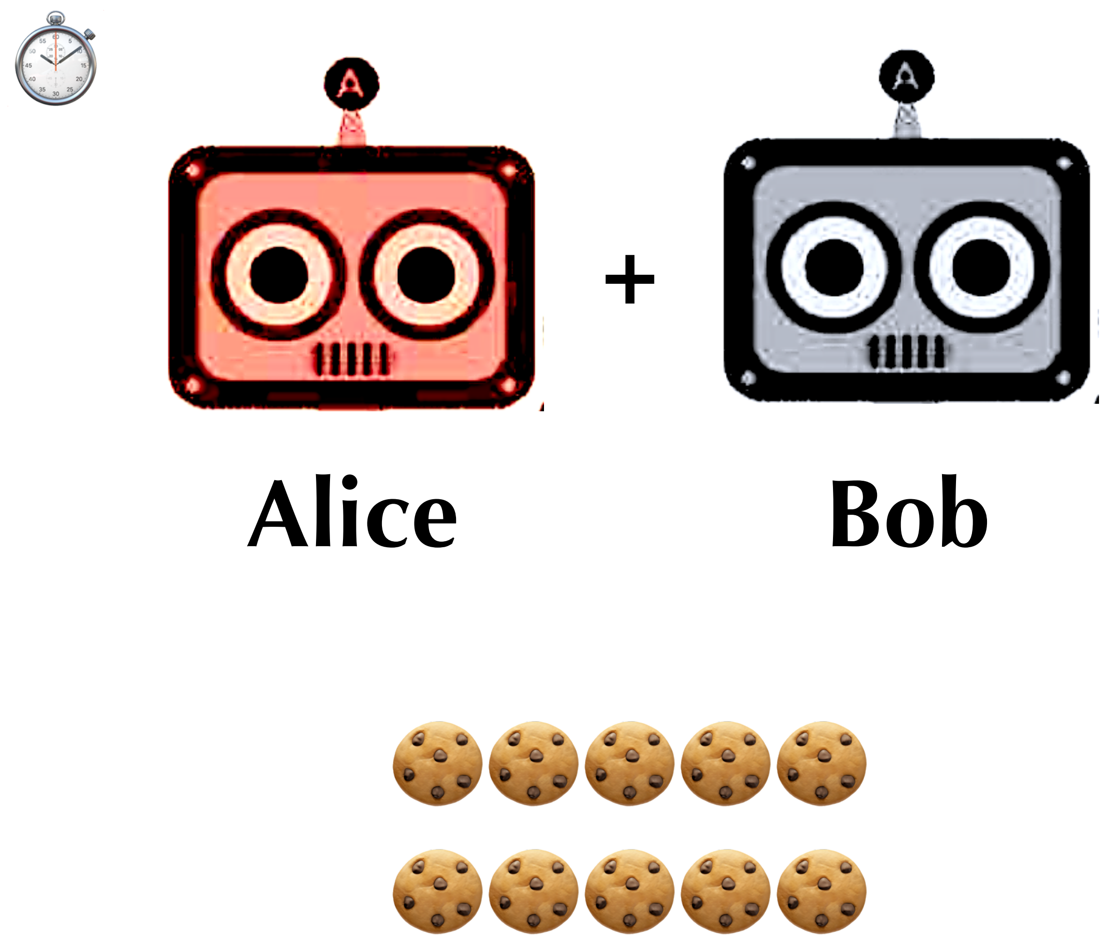
+



Alice **Bob**

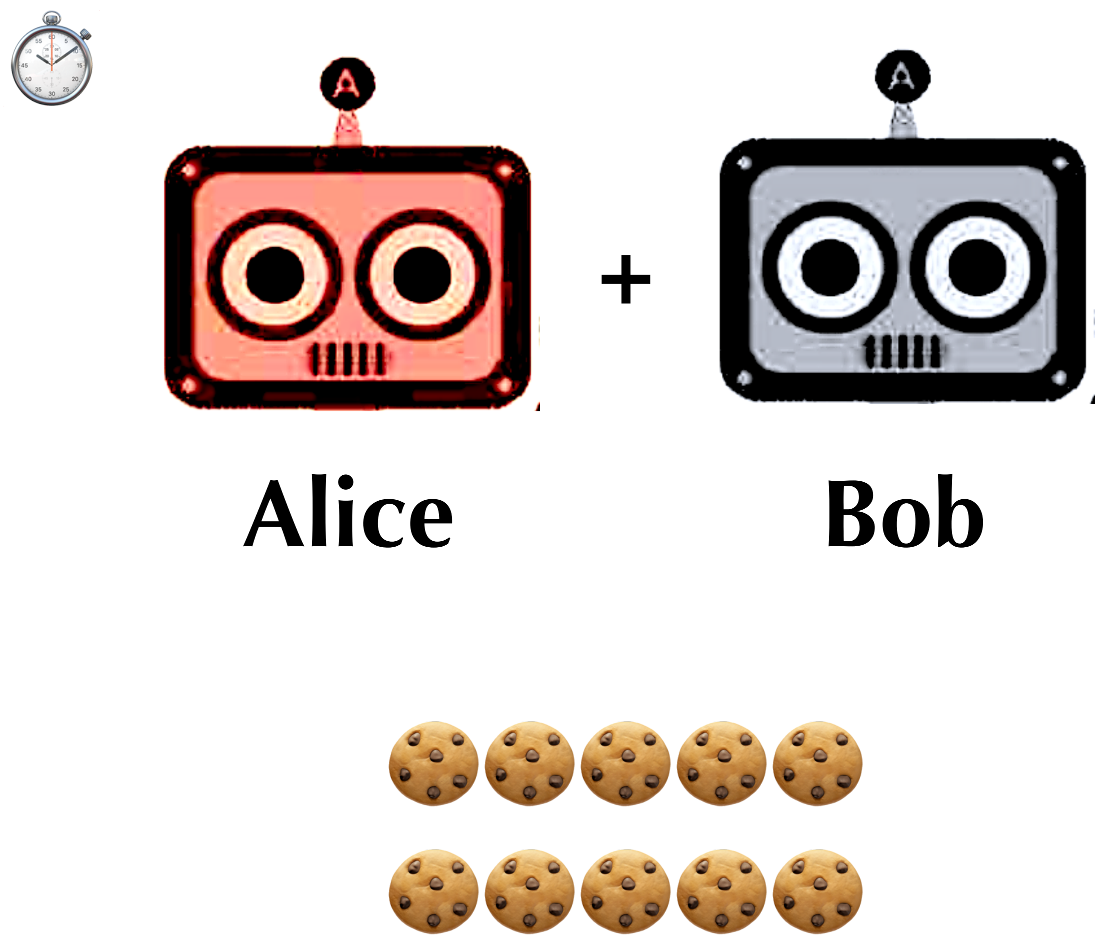
10 cookies 🍪 / hour

Example: Alice and Bob Making Cookies



What's the best way
to distribute cookies?

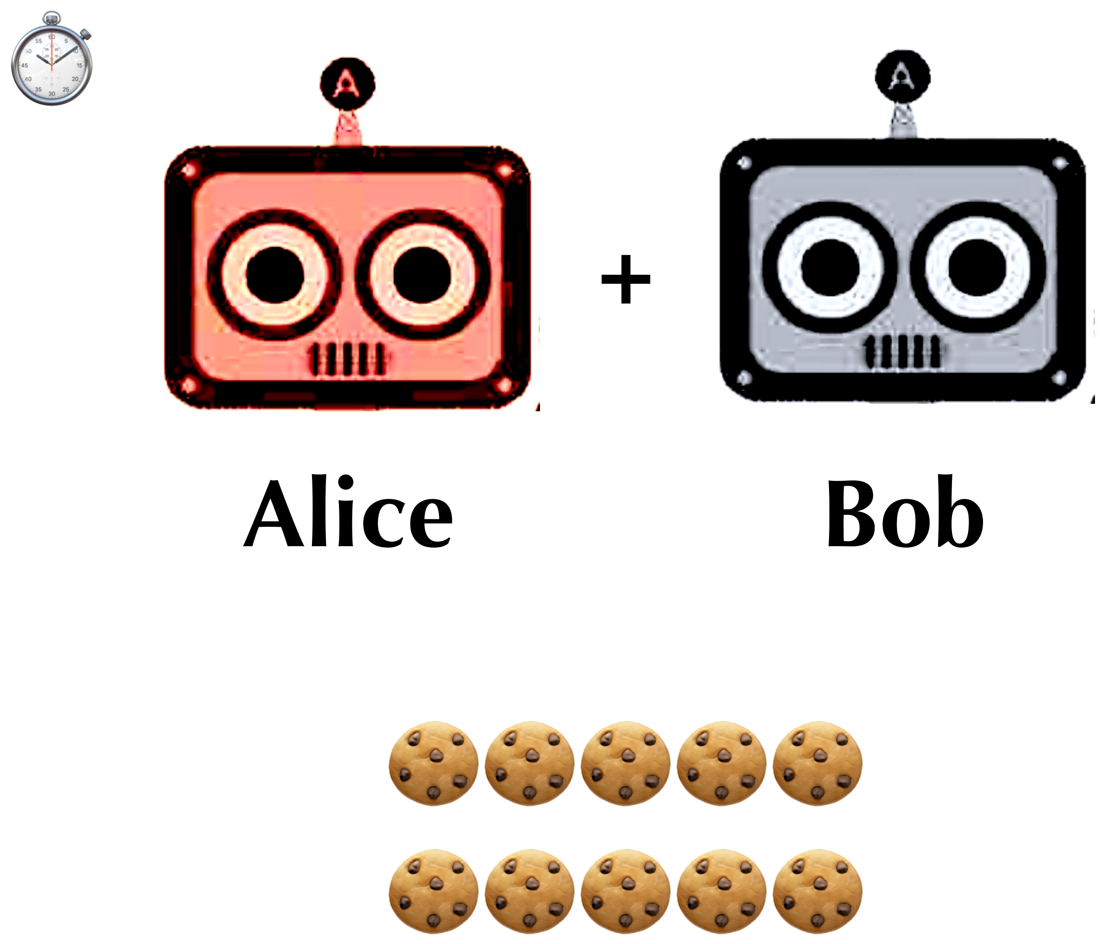
4 Axioms of Fair Division



Define "the best division":

1. **[Efficiency]** We don't want to waste cookies… all cookies should belong to either Alice or Bob.

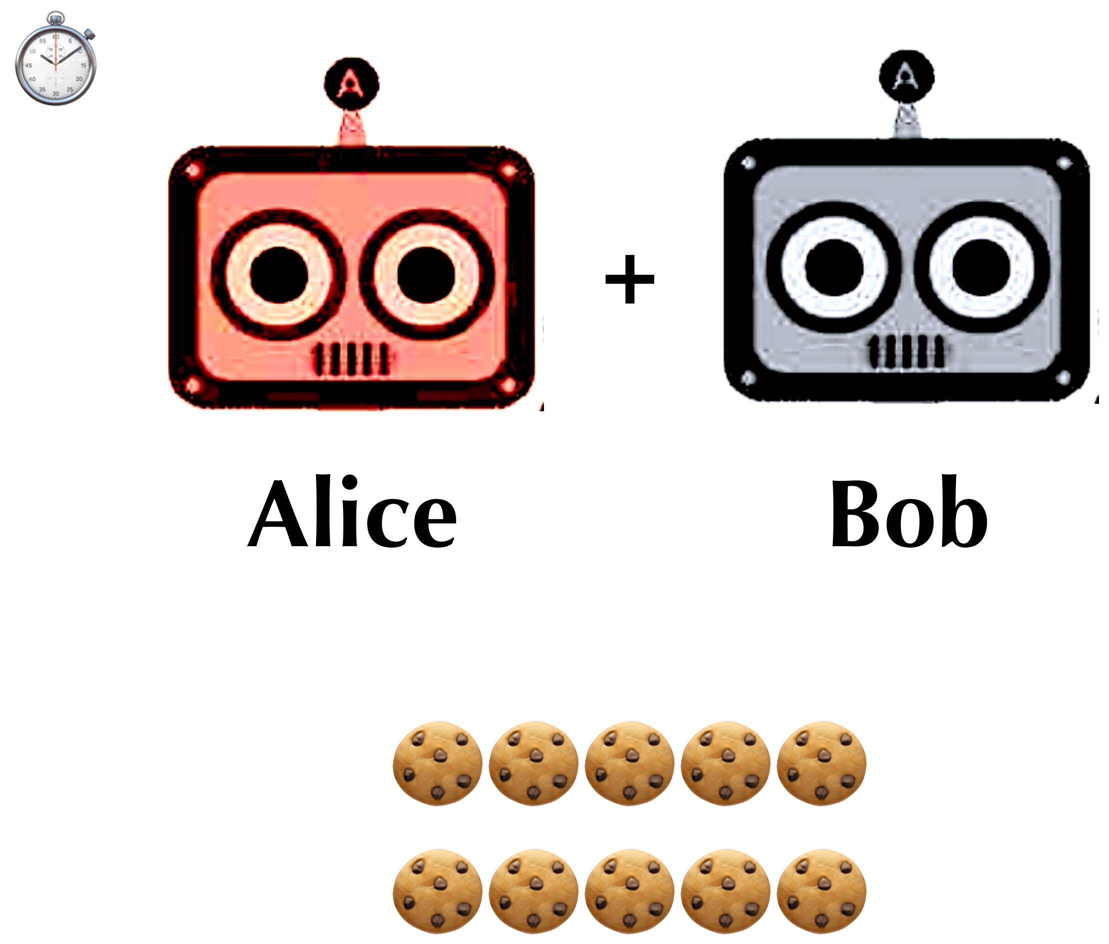
4 Axioms of Fair Division



Define "the best division":

1. **[Efficiency]** We don't want to waste cookies… all cookies should belong to either Alice or Bob.
2. **[Null Player]** If someone has no contribution at all, he should get nothing.

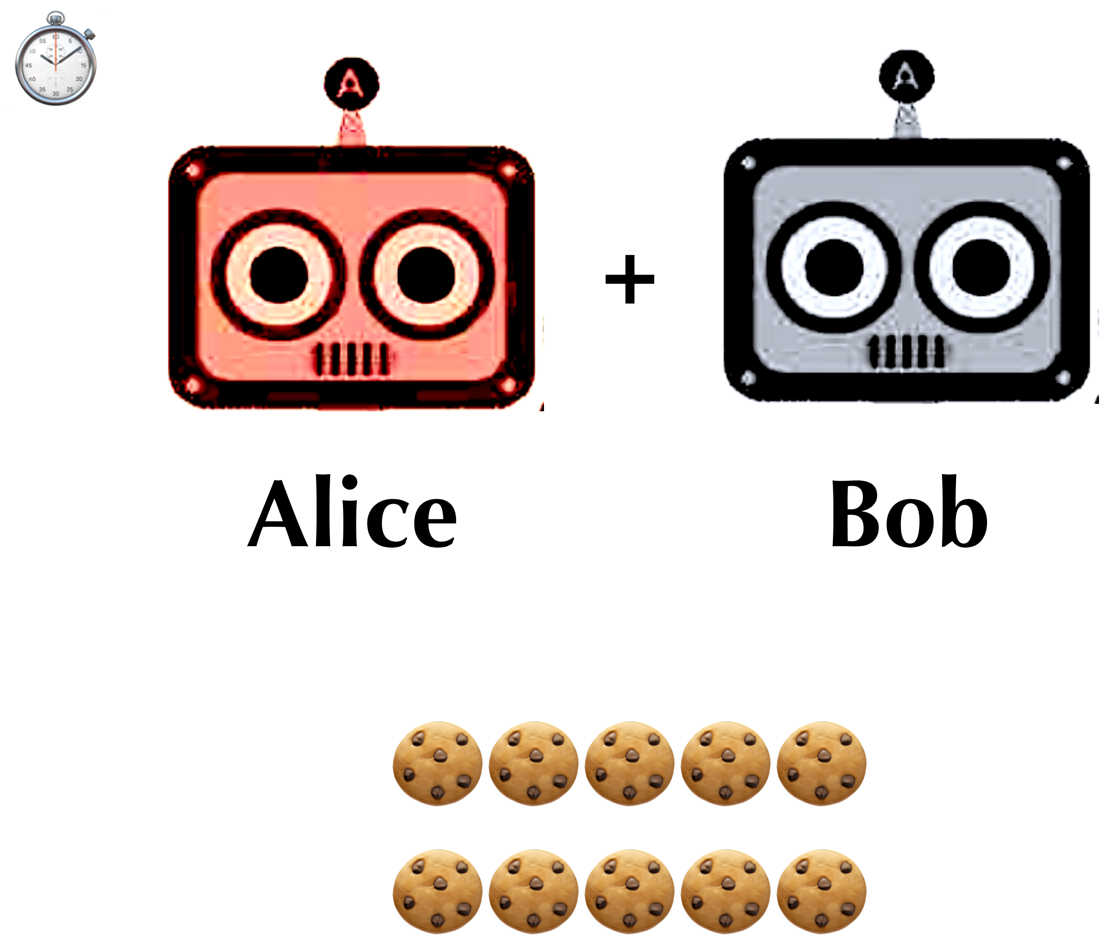
4 Axioms of Fair Division



Define "the best division":

1. **[Efficiency]** We don't want to waste cookies… all cookies should belong to either Alice or Bob.
2. **[Null Player]** If someone has no contribution at all, he should get nothing.
3. **[Symmetry]** If two people have exact same contribution, they should get the same number of cookies.

4 Axioms of Fair Division



Define "the best division":

1. **[Efficiency]** We don't want to waste cookies… all cookies should belong to either Alice or Bob.
2. **[Null Player]** If someone has no contribution at all, he should get nothing.
3. **[Symmetry]** If two people have exact same contribution, they should get the same number of cookies.
4. **[Linearity]** If they collaborate for a longer time, the strategy of division shouldn't change.

Formal Descriptions

Players: $N = \{1, \ldots, n\}$

Coalitions: $S \subseteq N$ $\{\}$ $\{\ \}$ $\{\ \}$

$\{\ ,\ \}$

Payoff Function: $v : 2^N \mapsto \mathbb{R}$

$v(\{\}) = 0$

$v(\{\ \}) =$ **5x**🍪

$v(\{\ \}) =$ **3x**🍪

$v(\{\ ,\ \}) =$ **10x**🍪

Define "the best division":

1. **[Efficiency]** We don't want to waste cookies… all cookies should belong to either Alice or Bob.
2. **[Null Player]** If someone has no contribution at all, he should get nothing.
3. **[Symmetry]** If two people have exact same contribution, they should get the same number of cookies.
4. **[Linearity]** If they collaborate for a longer time, the strategy of division shouldn't change.

Formal Descriptions

Players: $N = \{1, \ldots, n\}$

Coalitions: $S \subseteq N$ $\quad \{\}\ \{\ \}\ \{\ \}$

$\{\ ,\ \}$

Payoff Function: $v : 2^N \mapsto \mathbb{R}$

$v(\{\}) = 0$

$v(\{\ \}) =$ **5x**

$v(\{\ \}) =$ **3x**

$v(\{\ ,\ \}) =$ **10x**

Division: $\{\phi_i(v)\}$

4 axioms of Shapley values:

1. **[Efficiency]** $v(N) = \sum_{i \in N} \phi_i(v)$

2. **[Null Player]** $v(S \cup \{i\}) - v(S) = 0, \forall S \subseteq N \backslash \{i\}$

   $\Rightarrow \phi_i(v) = 0$

3. **[Symmetry]** $v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S),$

   $\Rightarrow \phi_i(v) = \phi_j(v) \qquad \forall S \subseteq N \backslash \{i, j\}$

4. **[Linearity]** $\phi_i(v + w) = \phi_i(v + w)$

   $\phi_i(\alpha v) = \alpha \phi_i(v), \forall i \in N$

Shapley value exists and is unique

$$\phi_i(v) = \frac{1}{n!} \sum_{R} \left[v(P_{R[:i]} \cup \{i\}) - v(P_{R[:i]}) \right]$$

Shapley, 1953

Shapley value exists and is unique

subset of players that
precede the player *i*

$$\phi_i(v) = \frac{1}{n!} \sum_{R} \left[v(P_{R[:i]} \cup \{i\}) - v(P_{R[:i]}) \right]$$

all possible permutations
of *n* players

Shapley value exists and is unique

marginal contribution of the
player i to the coalition $P_{R[:i]} \cup \{i\}$

$$\phi_i(v) = \frac{1}{n!} \sum_{R} \left[v(P_{R[:i]} \cup \{i\}) - v(P_{R[:i]}) \right]$$

Shapley value exists and is unique

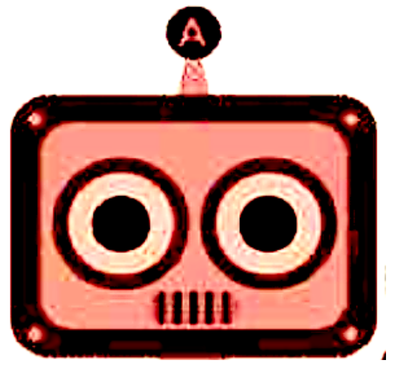
marginal contribution of the player *i* to the coalition $P_{R[:i]} \cup \{i\}$

$$\phi_i(v) = \frac{1}{n!} \sum_R \left[v(P_{R[:i]} \cup \{i\}) - v(P_{R[:i]}) \right]$$

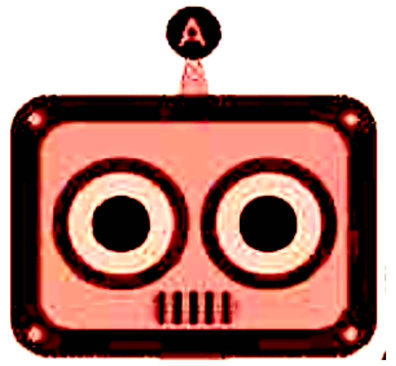
Shapley value is the **average marginal contribution** to all ordered coalitions.

Shapley Values: Fair Division in Cooperative Game

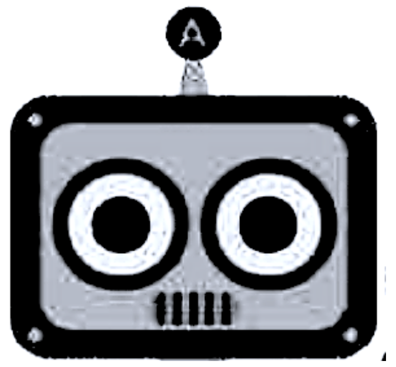
Alice’s marginal contribution


Alice

5x🍪

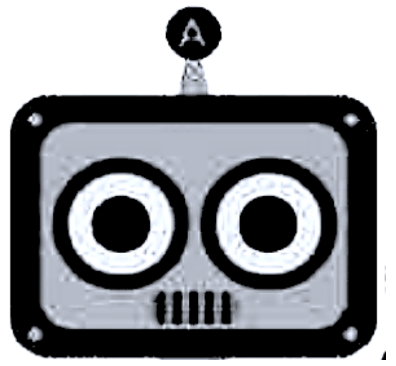

Alice

+

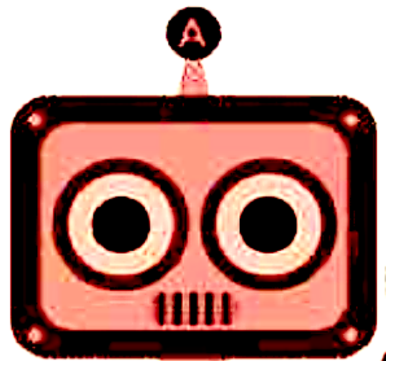

Bob

10x🍪 - 3x🍪 = 7x🍪

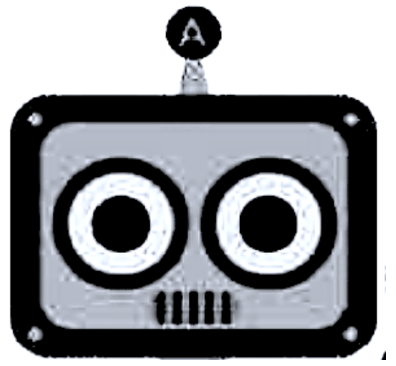
Bob’s marginal contribution


Bob

3x🍪


Alice

+

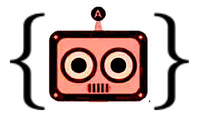
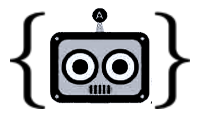

Bob

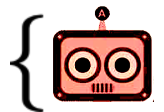
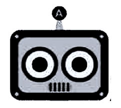
10x🍪 - 5x🍪 = 5x🍪

Alice should get 6x🍪, and Bob should get 4x🍪.

Model Prediction as a Cooperative Game

Players: $N = \{1, \ldots, n\}$ **-> possible subjects of the explanation, e.g., tokens**

Coalitions: $S \subseteq N$ $\{\}$ {} {}

{ , } **-> e.g., all non-masked out Tokens**

Payoff Function: $v : 2^N \mapsto \mathbb{R}$

$$v(\{\}) = 0$$

-> e.g., quality of prediction

Credit assignment: $\{\phi_i(v)\}$

-> importance accorded to subjects

Why making this connection?

- We can provide **more specific** interpretations of model behavior, backed by **theoretical guarantees**.
- We can understand the **role of groups of tokens** by treating them as a single player; there's no canonical way to aggregate units in most current methods.
- This will give us explanations that are both **fast** and **faithful**.

Ethayarajh and Jurafsky, 2021

Attention Weights Are Not Faithful Explanations

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore	*after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore*
original α	adversarial $\tilde{\alpha}$
$f(x\|\alpha, \theta) = 0.01$	$f(x\|\tilde{\alpha}, \theta) = 0.01$

Jain and Wallace, 2019

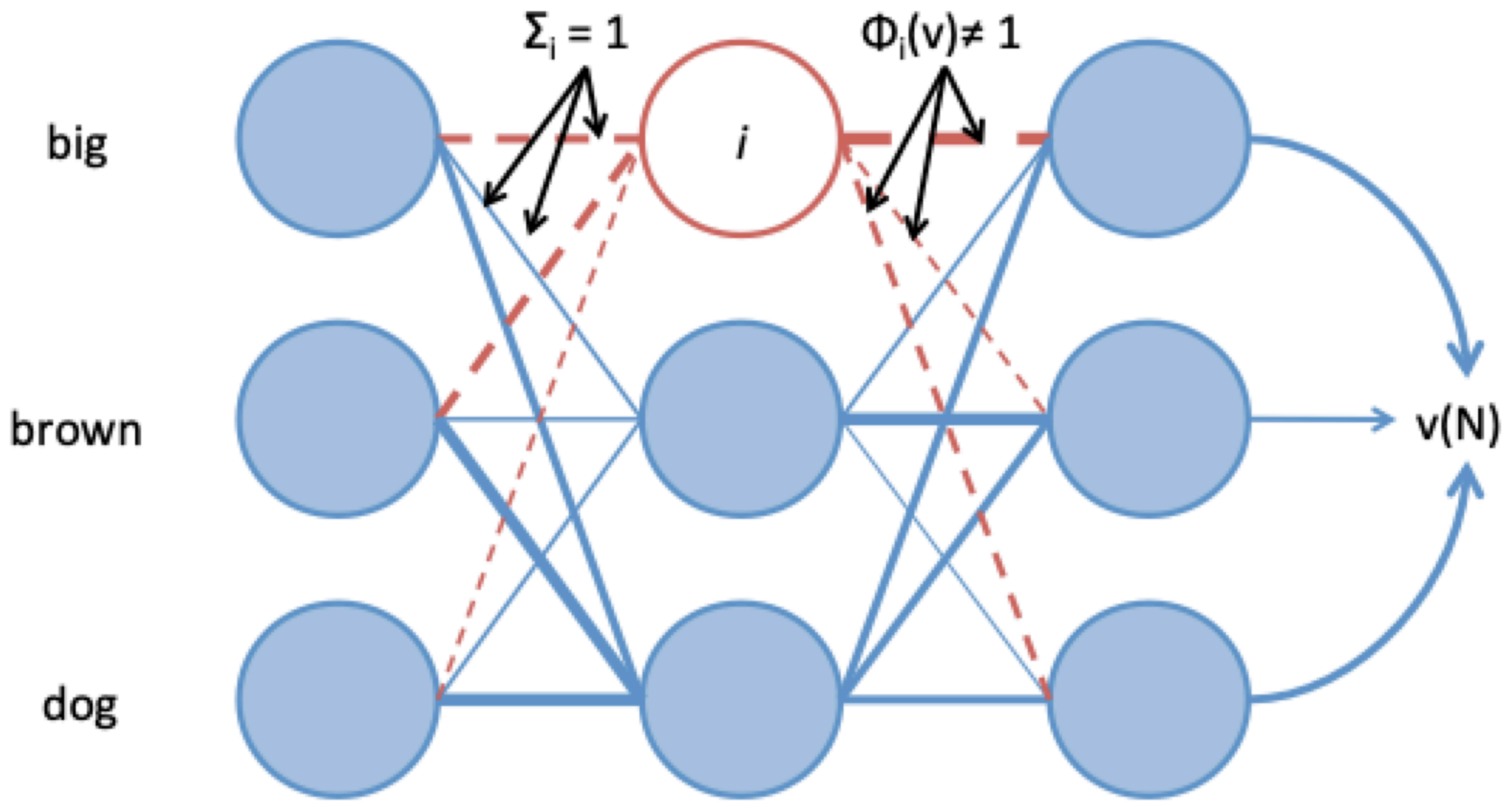
Attention Weights Are Not Shapley Values

Proposition 1. If some player is attended to more than another, there is no TU-game (N, v) for which attention weights are Shapley Values.

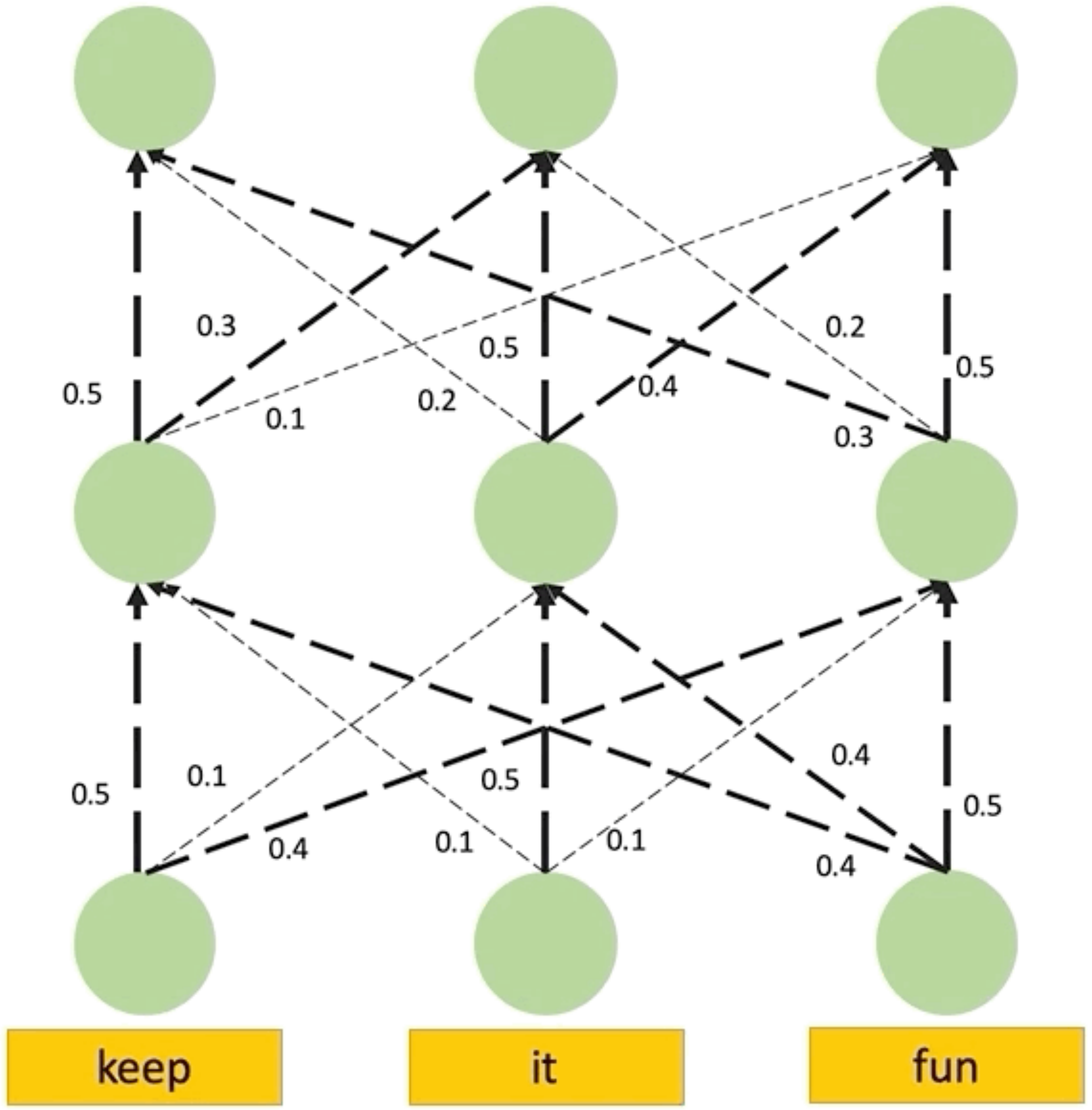
Ethayarajh and Jurafsky, 2021

Attention Weights Are Not Shapley Values

Intuition: A player's contribution to the total payoff ($\Sigma_i = 1$) is rarely equal to the total attention paid to it, so the latter cannot be its Shapley Value (Φ_i)…



Attention Flows



Abnar and Zuidema, 2020

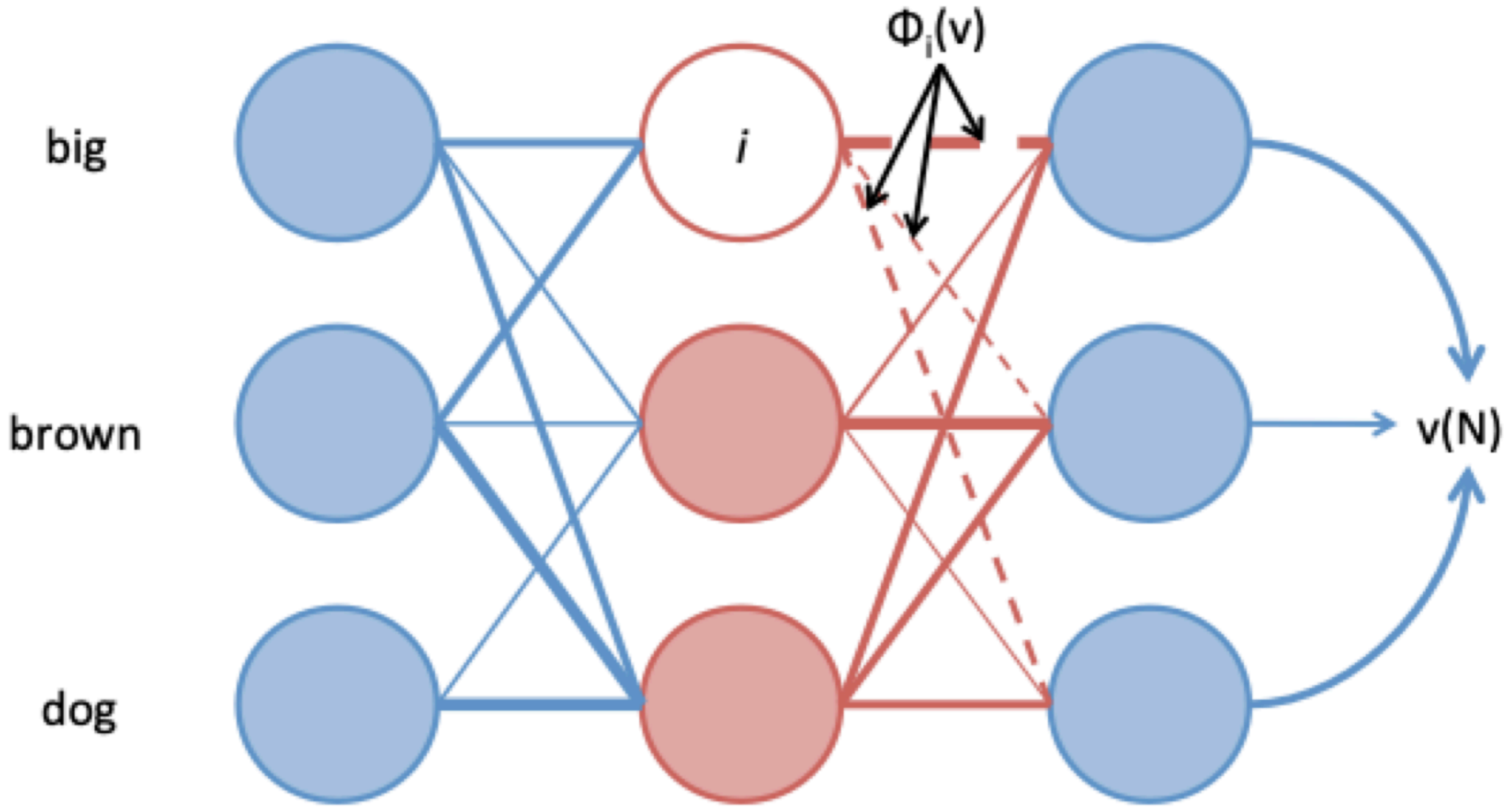
Attention Flows Can be Shapley Value

Proposition 2. Consider a TU-game (N, v), where $N = \{1, ..., n\}$ players are all from the same layer. Let f denote the flow obtained by running a max-flow algorithm on the graph defined by the self-attention matrix, where the capacities are the attention weights. Let $v(S) = |f(S)|$, the *value of the flow* when only permitting flow through players in the coalition $S \subseteq N$. Then for each player i, its total outflow $|f_o(i)|$ is its Shapley Value.

Ethayarajh and Jurafsky, 2021

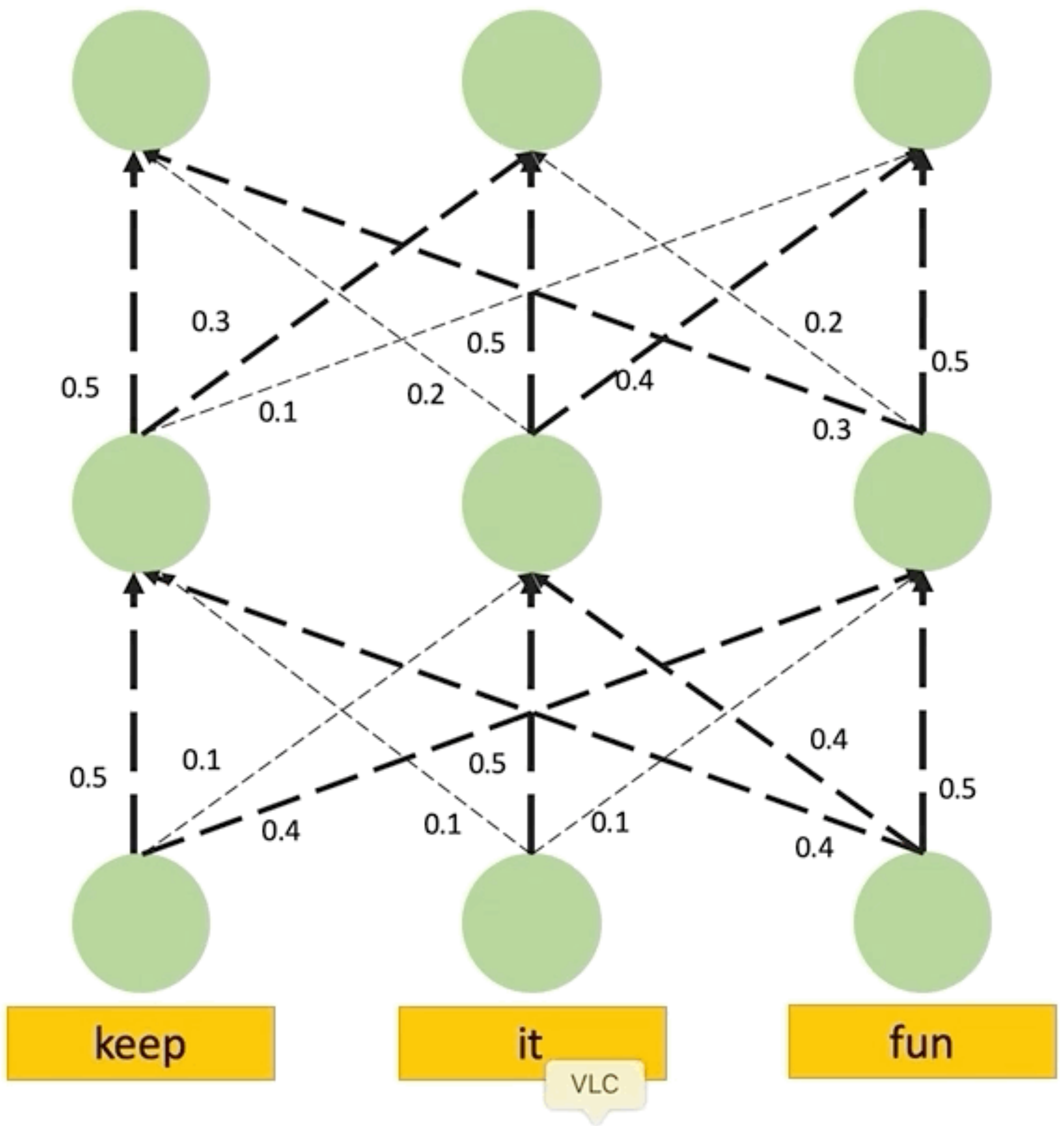
Attention Flows Can be Shapley Value

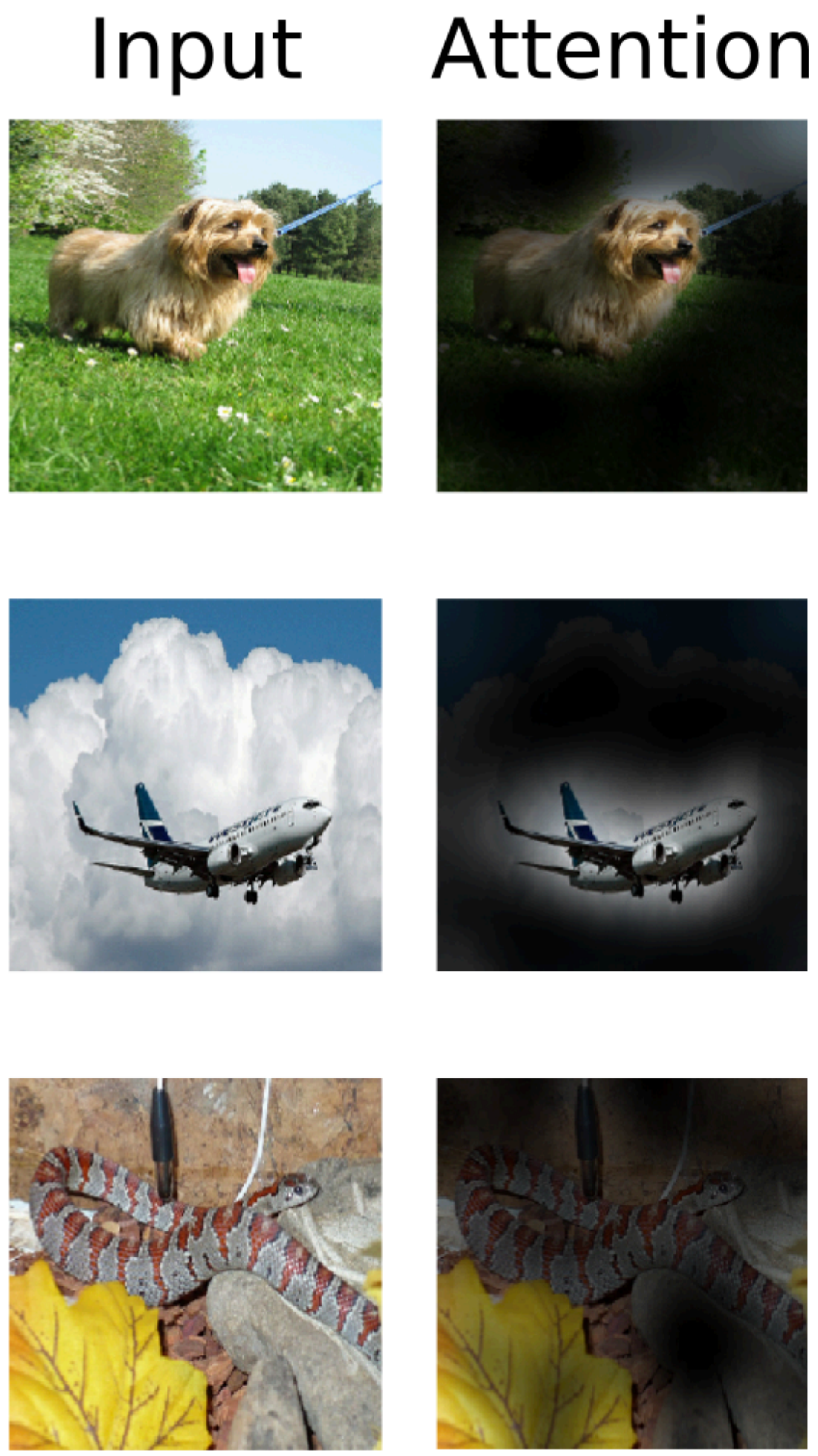
Intuition: when all players are from the same layer of a network, and the payoff is the total flow through the network, a player *i*'s total outflow is independent of others'…



Ethayarajh and Jurafsky, 2021

Attention rollout





Abnar and Zuidema, 2020, Dosovitskiy et al. 2021

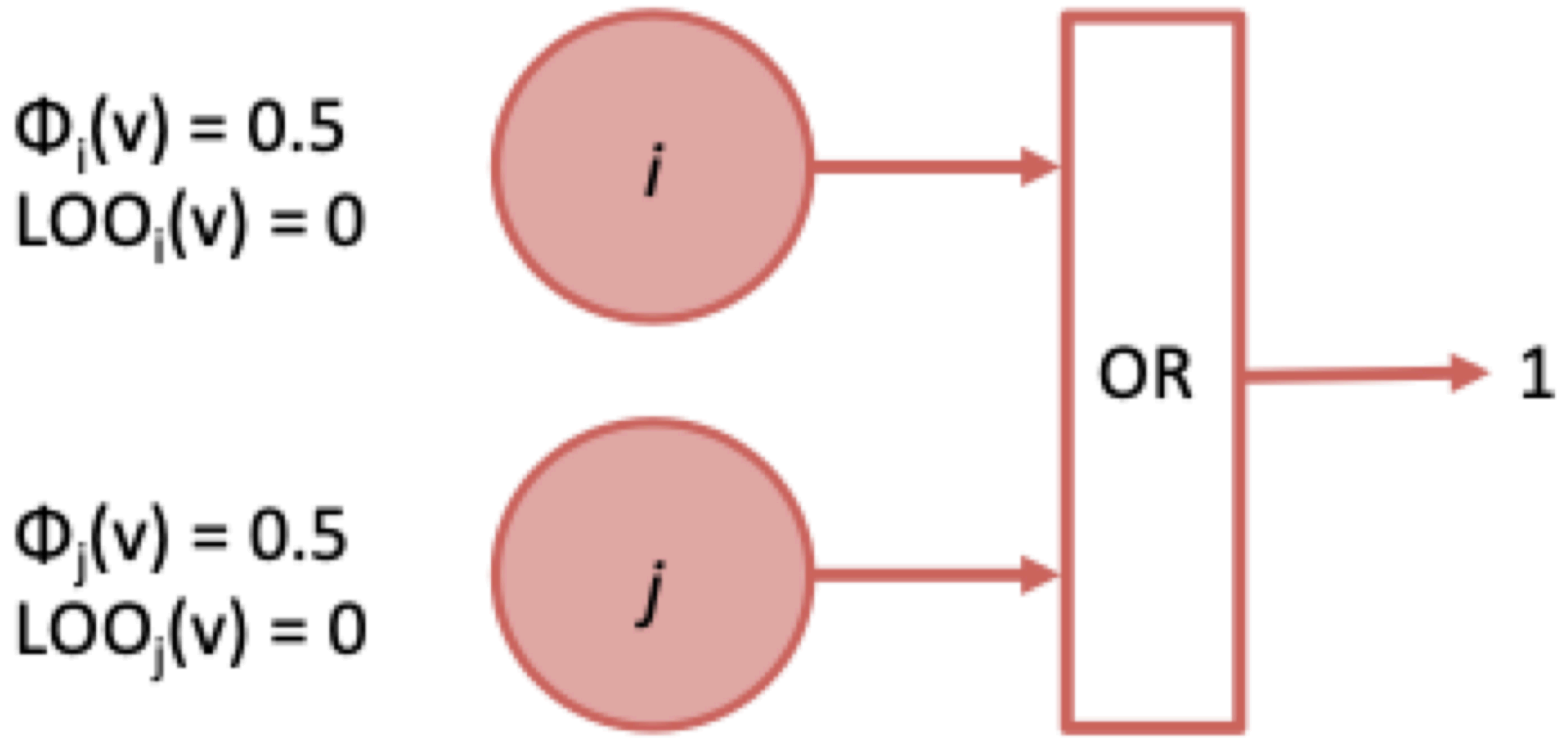
Leave-One-Out Values

Proposition 3. If $\exists\ i \in N$ such that player i is not a null player even when excluding the coalition $N \setminus \{i\}$, then there is no TU-game (N, v) for which leave-one-out values are Shapley Values.

Ethayarajh and Jurafsky, 2021

Leave-One-Out Values Are Not Shapley Values

Intuition: when small coalitions matter more than the largest one… e.g. if two representations played a critical role in a prediction but only one was necessary — then leave-one-out would assign each a value of zero.



Ethayarajh and Jurafsky, 2021

Discussion

Does the theory make sense … ?

***When to use leave-one-out values?**
***Flexibility in the choice of payoff functions.**
***Generalized cooperative game with multiple actions.**