Monitoring Algorithmic Fairness

Thomas A. Henzinger | Mahyar Karimi | Konstantin Kueffner | Kaushik Mallik

Institute of Science and Technology Austria

Monitoring?

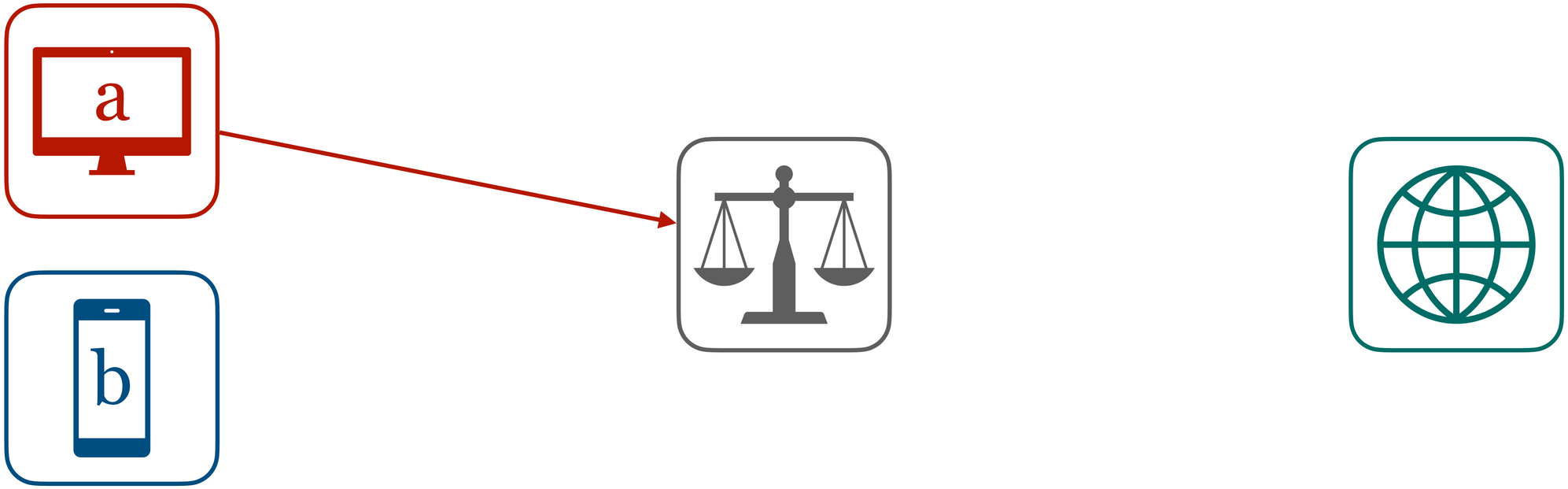
Monitor watches a black-box system and produces verdicts in real-time.

Example.

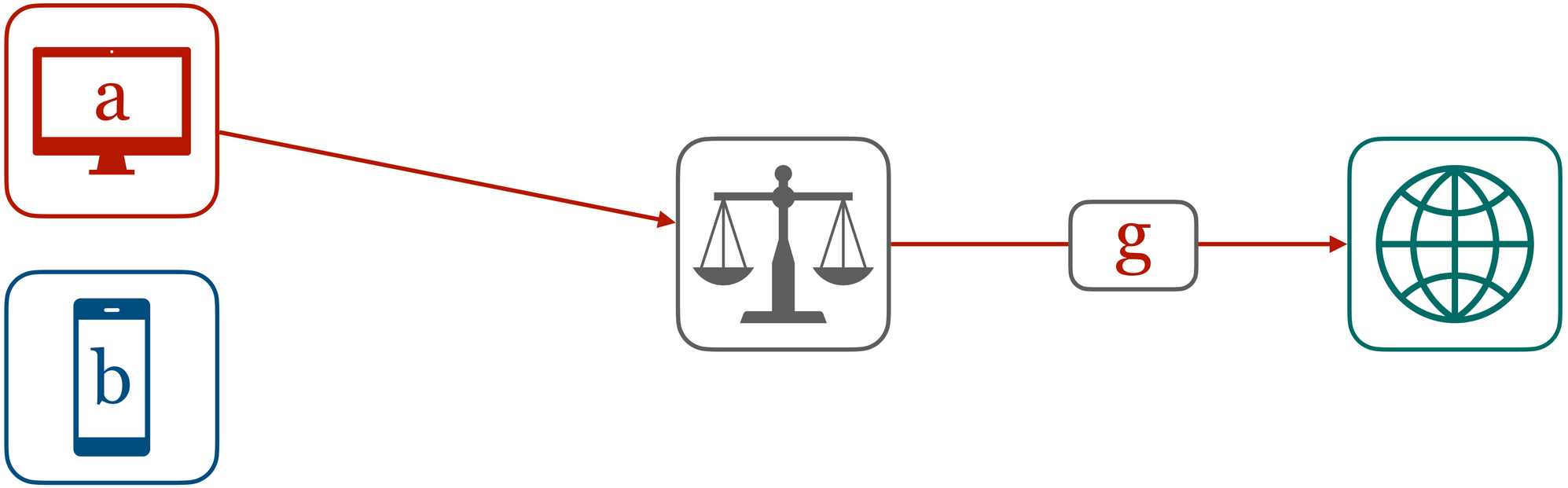
A simple resource allocation problem.


a

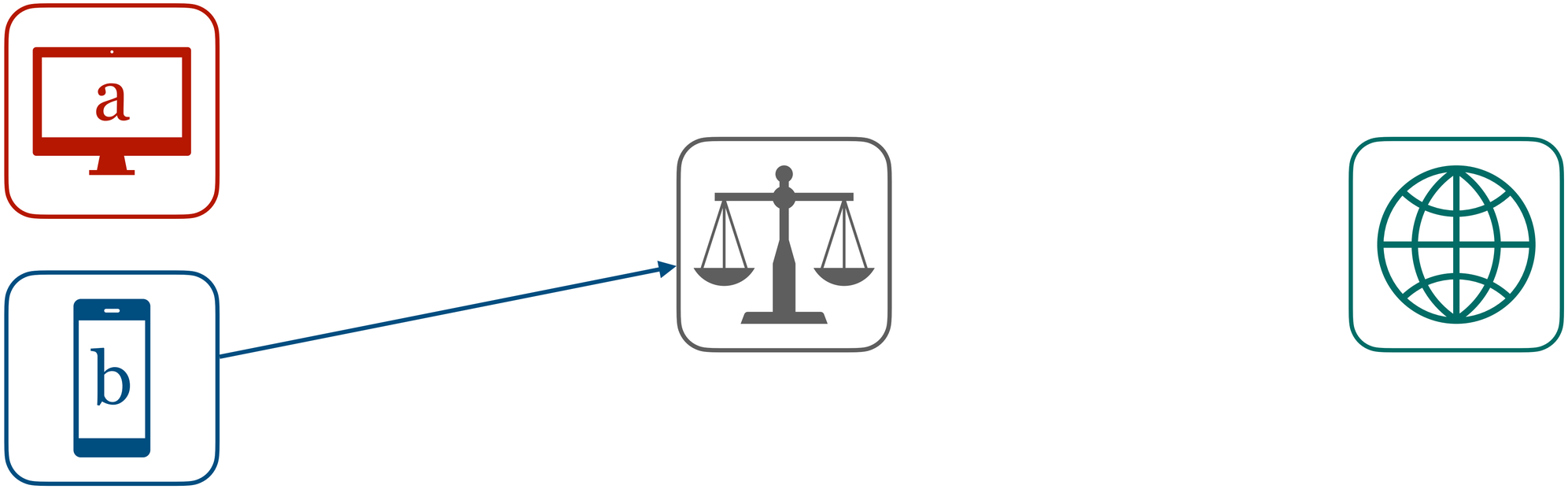

b



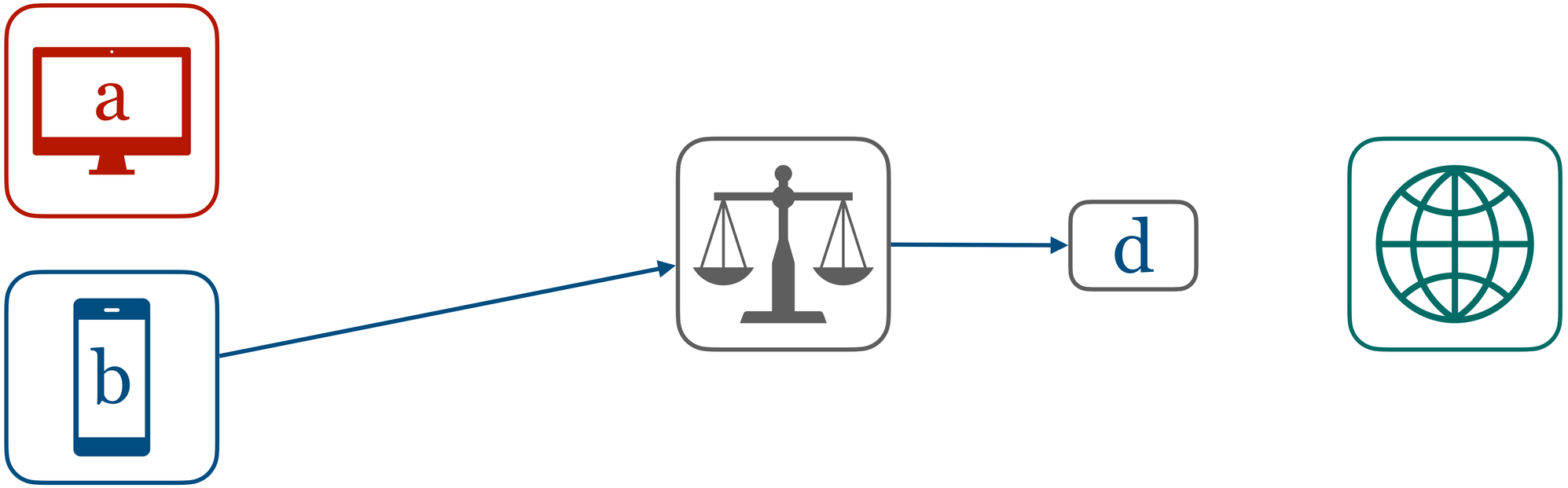
a



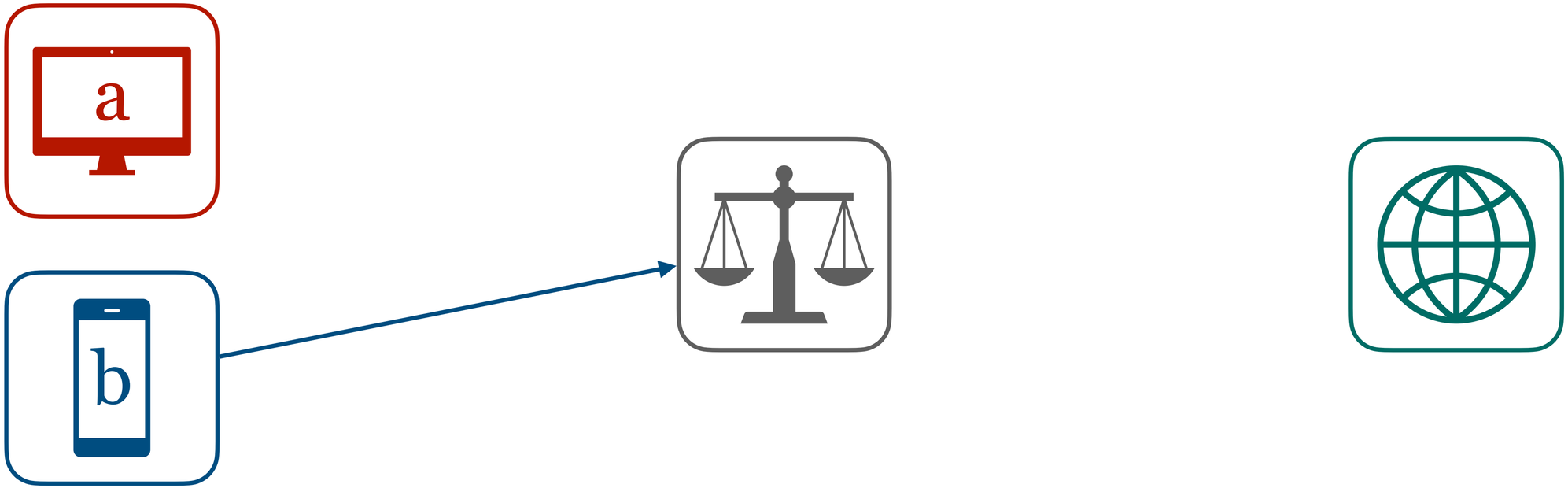
a g



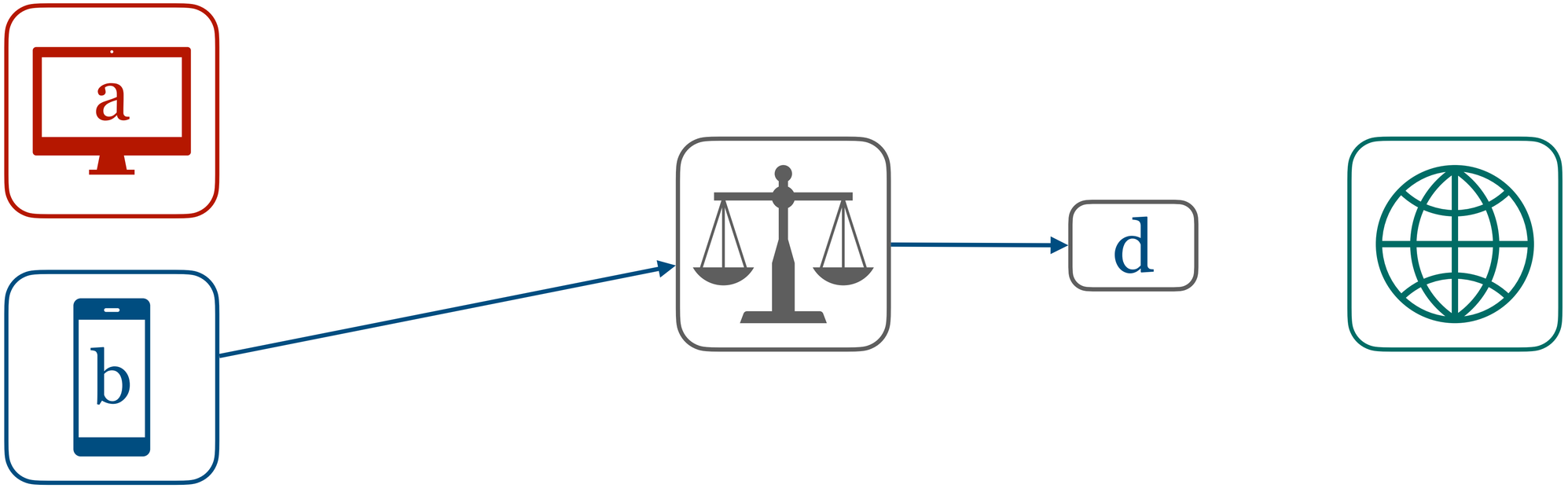
a g b



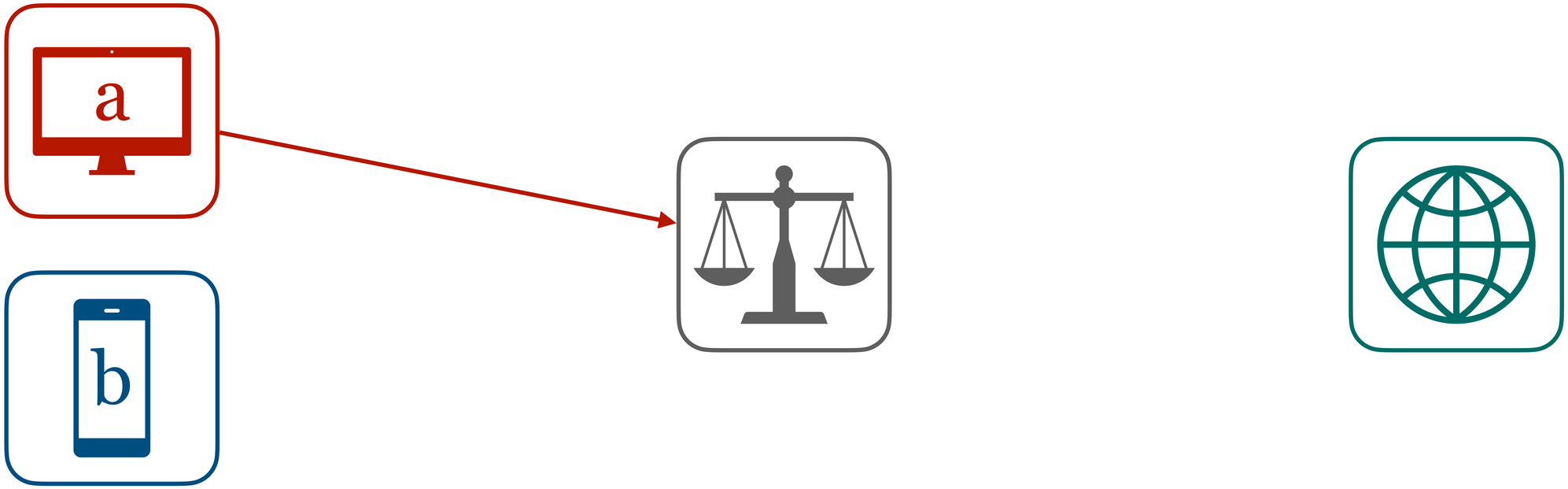
a g b d



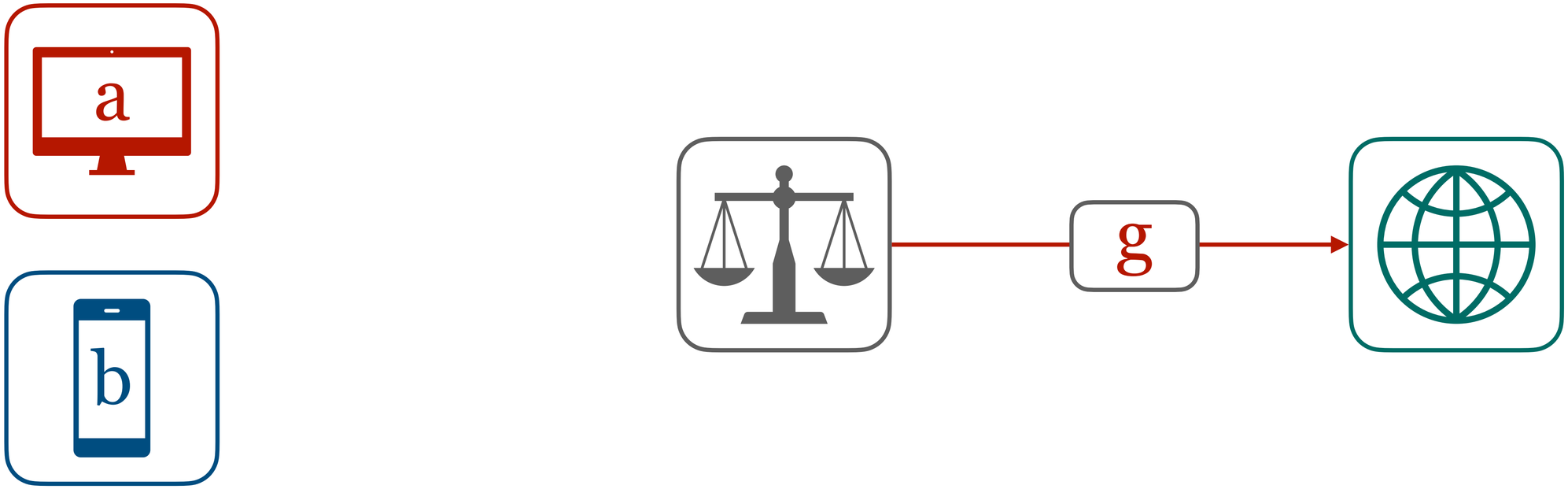
a g b d b



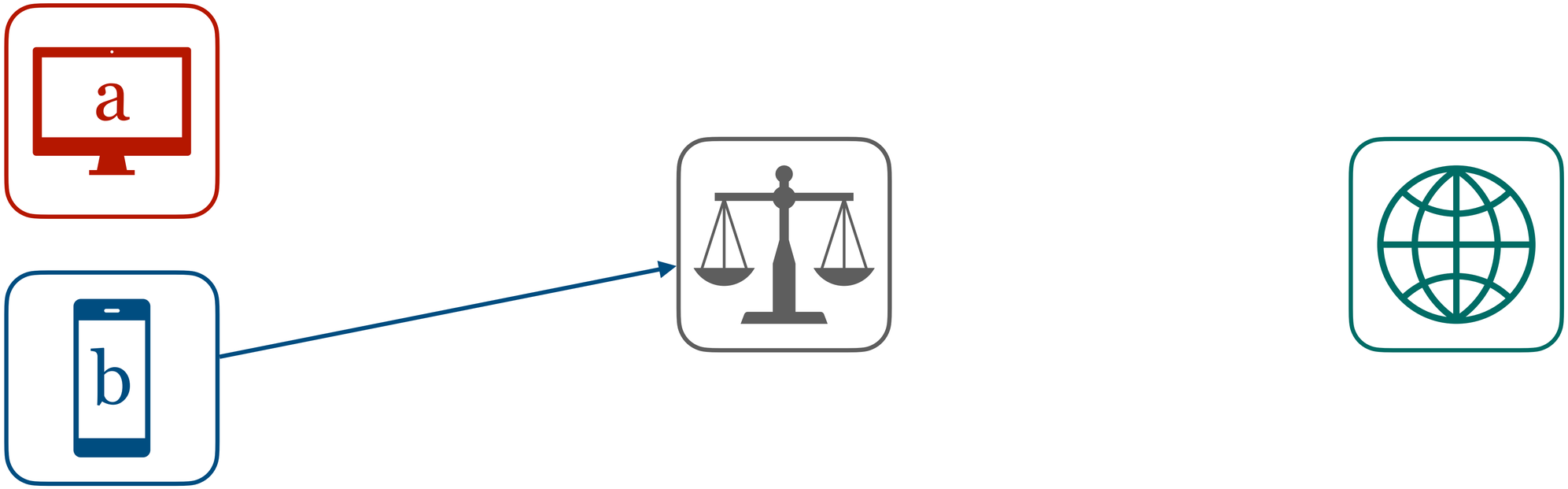
a g b d b d



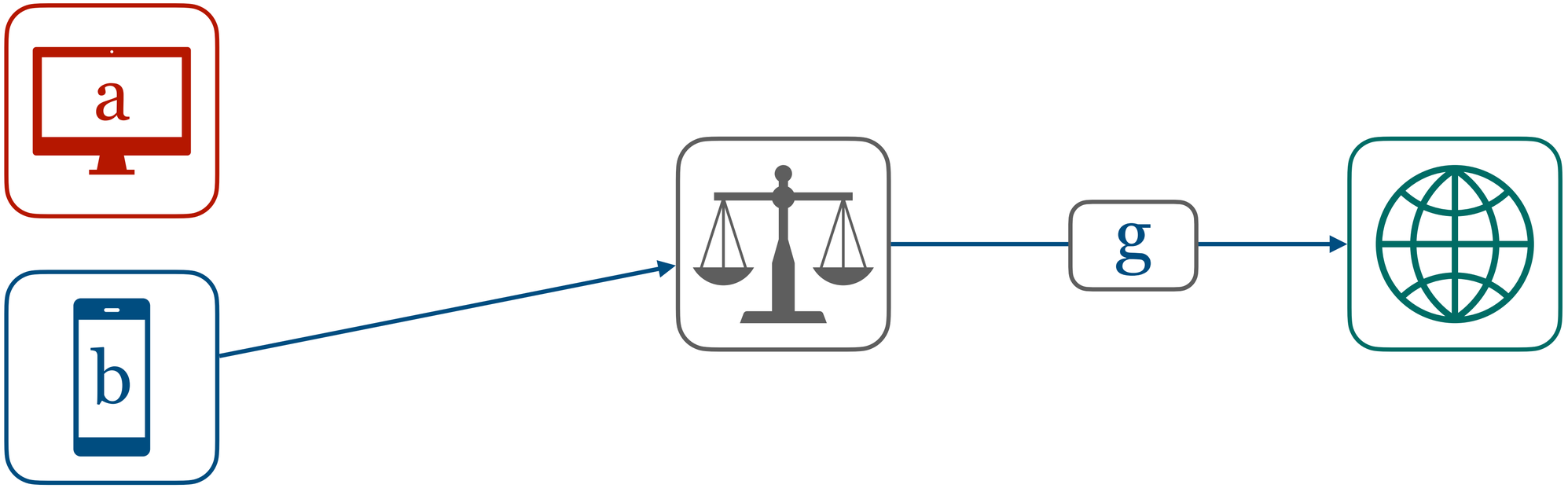
a g b d b d a



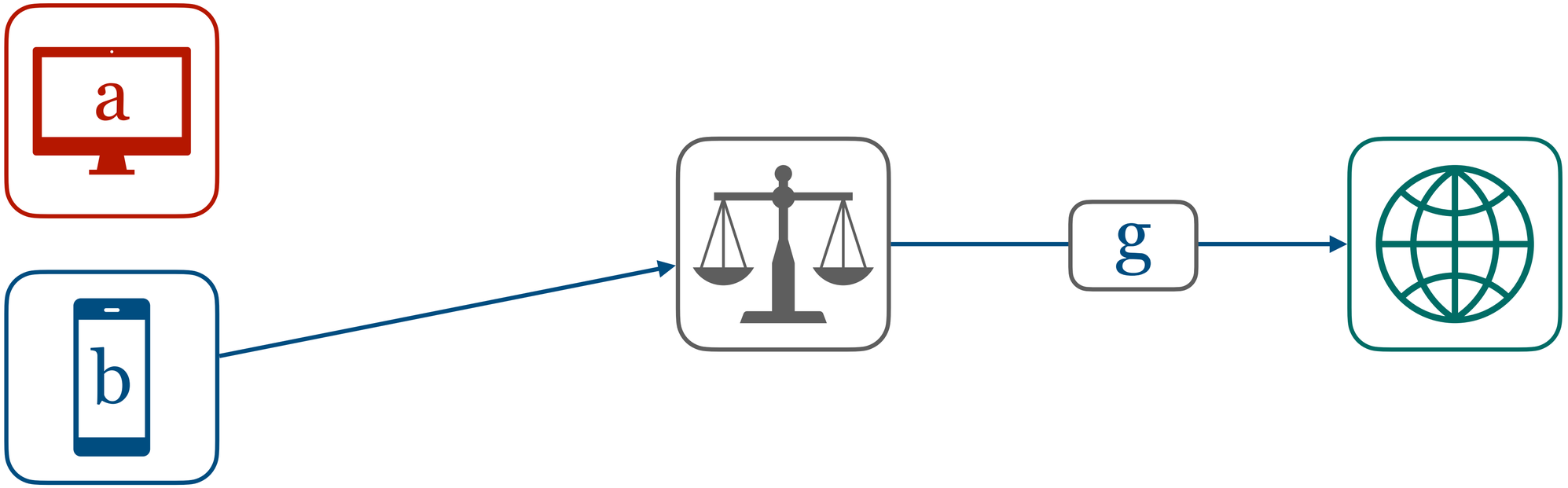
a g b d b d a g



a g b d b d a g b



a g b d b d a g b g

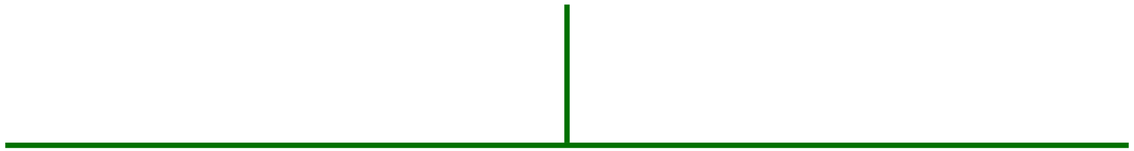


a g b d b d a g b g . . . . . .

Is the arbiter fair?

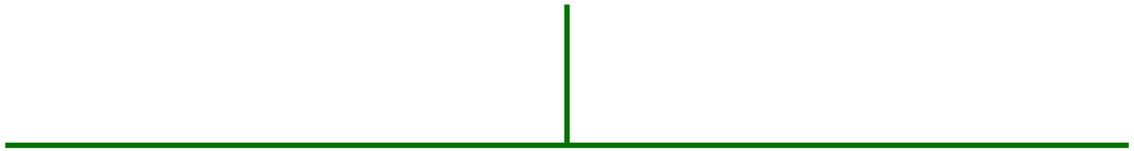
Quantitative notions of fairness.

$$\mathbb{P}(g \mid a) - \mathbb{P}(g \mid b)$$



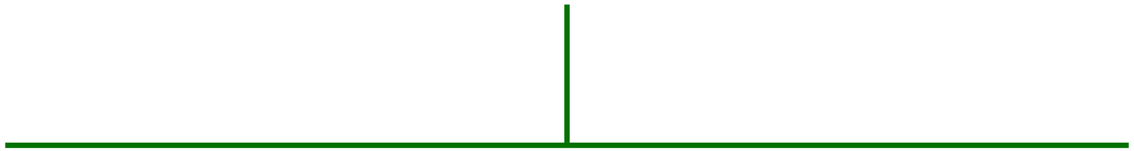
Demographic Parity

$$\mathbb{P}(g \mid a) \div \mathbb{P}(g \mid b)$$



Disparate Impact

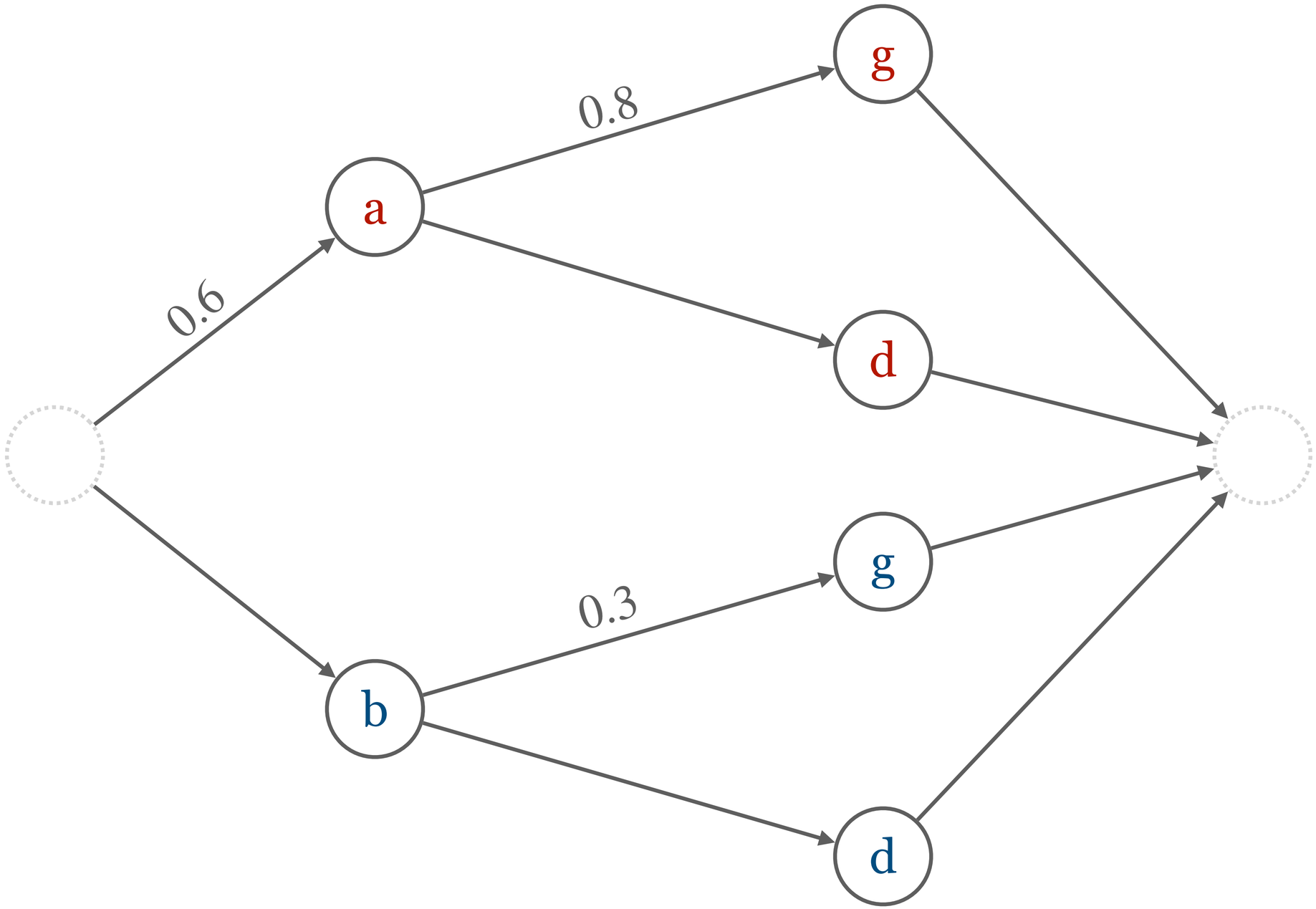
$$\Big(c_g\mathbb{P}(g\,|\,a)+c_d\mathbb{P}(d\,|\,a)\Big)-\Big(c_g\mathbb{P}(g\,|\,b)+c_d\mathbb{P}(d\,|\,b)\Big)$$

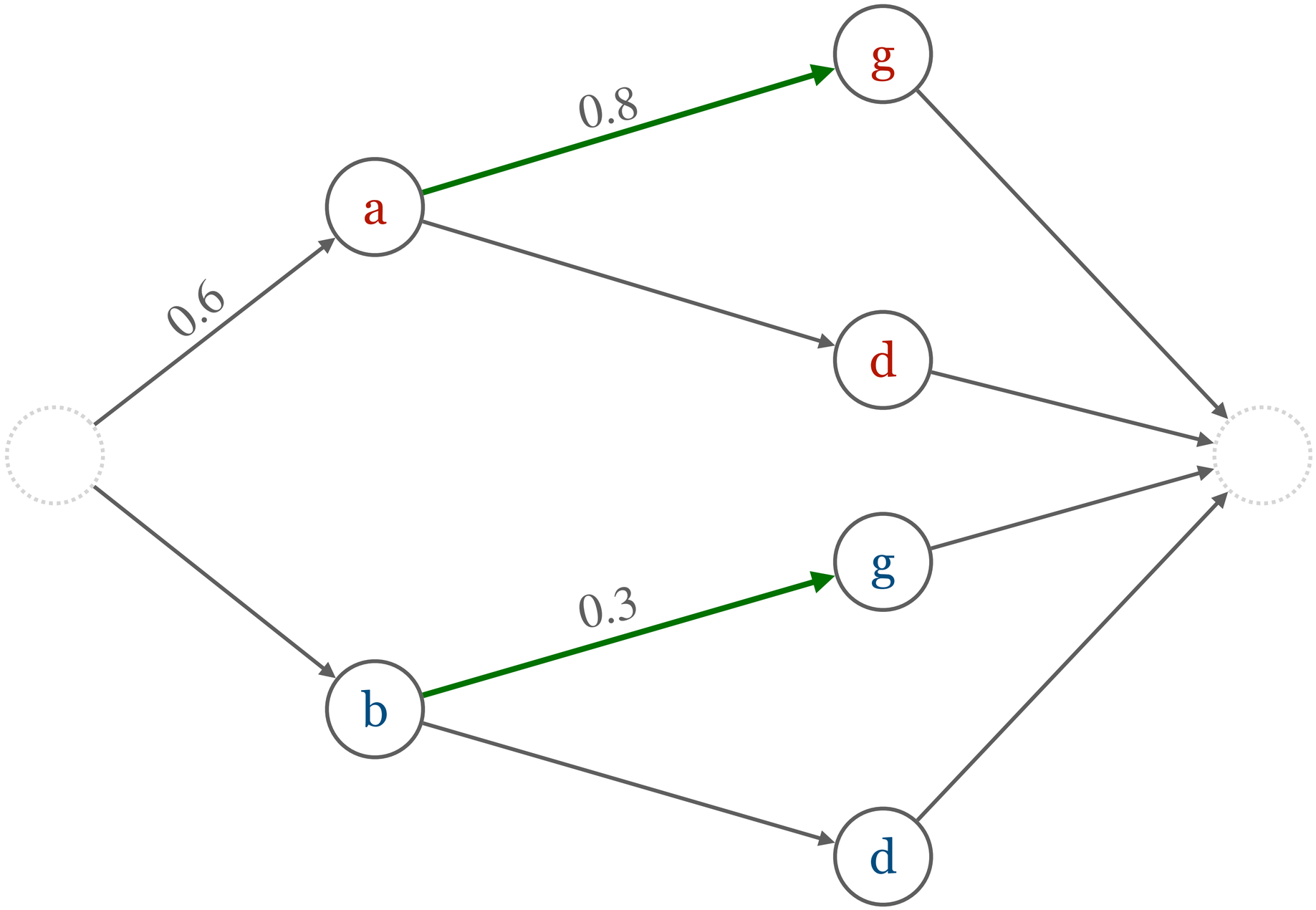


Social Burden

Monitorability.

We need to make assumptions about the system.

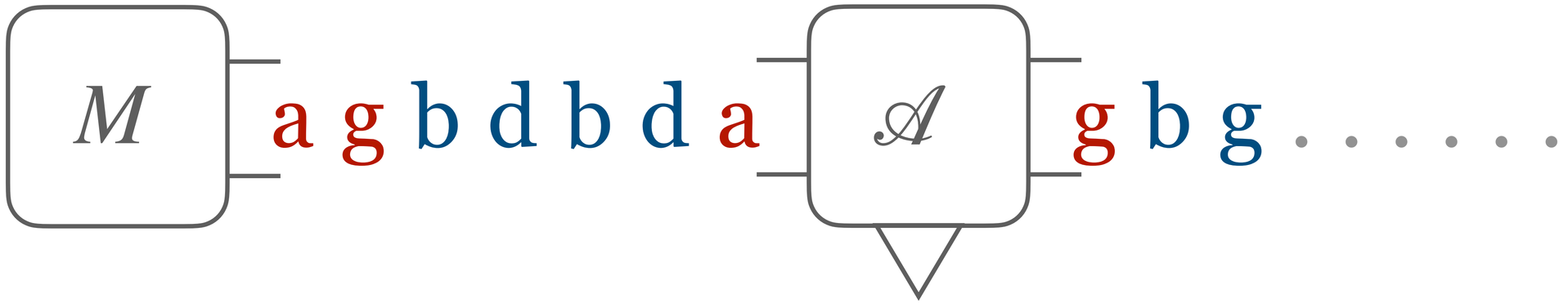

g
0.8
a
0.6
d
g
0.3
b
d

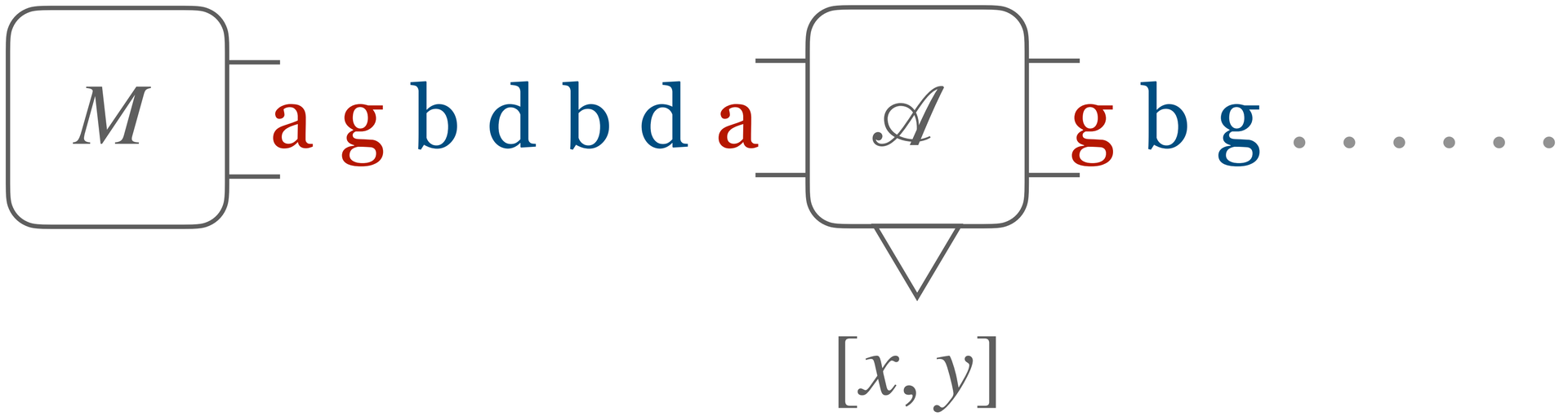

g
0.8
a
0.6
d
g
0.3
b
d

Intuition.

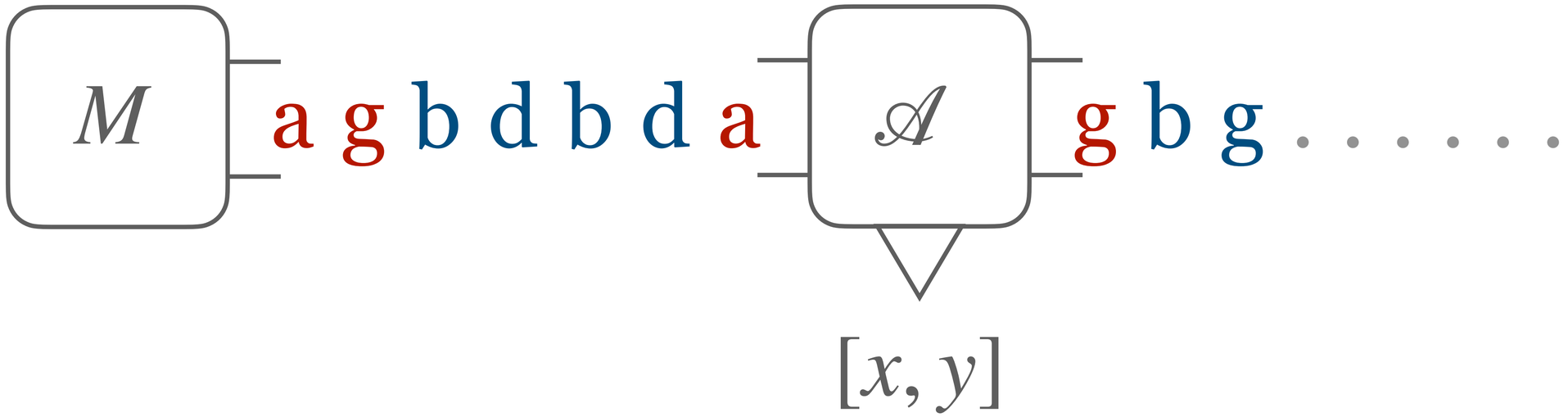
We observe a Markov chain and at every time step the monitor provides a PAC-style estimate for some fairness property.


M
a g b d b d a g b g . . . . . .


M
a g b d b d a
A
g b g . . . . . .


M
a g b d b d a
𝒜
g b g . . . . . .
[x, y]

$$\mathbb{P}(g \mid a) - \mathbb{P}(g \mid b) \in [x, y] \text{ with probability } 1 - \delta$$



Formal Problem.

Let's be a bit more general.

Language.

How to specify fairness?

Probabilistic Specification Expression

$$\xi ::= v \in \{v_{ij}\}_{i,j \in Q} \mid \xi \cdot \xi \mid 1 \div \xi$$

$$\varphi := \kappa \in \mathbb{R} \mid \xi \mid \varphi + \varphi \mid \varphi - \varphi \mid \varphi \cdot \varphi \mid (\varphi)$$

Probabilistic Specification Expression

$$\xi ::= v \in \{v_{ij}\}_{i,j \in Q} \mid \xi \cdot \xi \mid 1 \div \xi$$

$$\varphi := \kappa \in \mathbb{R} \mid \xi \mid \varphi + \varphi \mid \varphi - \varphi \mid \varphi \cdot \varphi \mid (\varphi)$$

Special Forms:

Polynomial: $\Sigma_i \kappa_i \cdot \xi_i$

Single-Division: $\varphi_1 + \varphi_2 \div \varphi_3$

What can we express?

Arithmetic expressions over*

$\mathbb{P}(r \mid q)$

**with limited division*

What can we express?

Arithmetic expressions over*

$\mathbb{P}(r \mid q)$

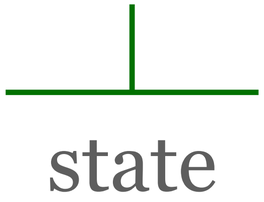
star-free regular expression

**with limited division*

What can we express?

Arithmetic expressions over*

$\mathbb{P}(r \mid q)$

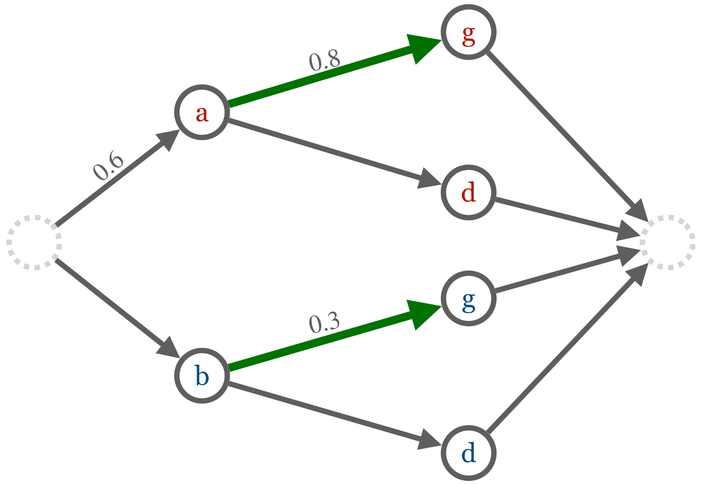


state

**with limited division*

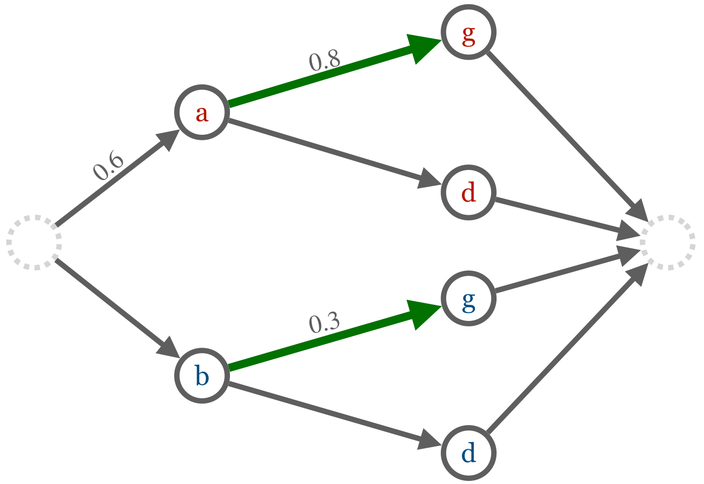
$$\varphi := \nu_{ag} - \nu_{bg}$$

$$M :=$$



$$\varphi := v_{ag} - v_{bg}$$

$$M :=$$



$$\varphi(M) = 0.8 - 0.3 = 0.5$$

Problem Statement.

Two interpretations.

Frequentist:

Let $M \in \Delta(N-1)^N$,

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$ and $U \prec W$.

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$ and $U \prec W$. Given a PSE φ,

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$ and $U \prec W$. Given a PSE φ and $\delta > 0$

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$ and $U \prec W$. Given a PSE φ and $\delta > 0$ find a monitor $\mathscr{A} : [N]^* \to \mathbb{R}^2$ such that:

Frequentist:

Let $M \in \Delta(N-1)^N$, $W \sim (M, q_0)$ and $U \prec W$. Given a PSE φ and $\delta > 0$ find a monitor $\mathscr{A} : [N]^* \to \mathbb{R}^2$ such that:

$$\mathbb{P}(\varphi(M) \in \mathscr{A}(U)) \geq 1 - \delta$$

Bayesian:

Let $M \in \Delta(N-1)^N$, $M \sim \mathscr{D}$ and $u \in \{q_0\} \times [N]^*$. Given a PSE φ and $\delta > 0$ find a monitor $\mathscr{A} : [N]^* \to \mathbb{R}^2$ such that:

$$\mathbb{P}(\varphi(M) \in \mathscr{A}(u) \mid u) \geq 1 - \delta$$

Frequentist.

Input has to be in Single-Division form.

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

For each v_{ij} in φ create sequence $X_{ij} = X_{ij}^1, X_{ij}^2 \ldots$ s.t.:

$X_{ij}^k = 1$ iff the k^{th} visit to state i transitions to j.

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

$$X_{ag} = 1,1,1,0,0,1$$

$$X_{bg} = 0,0,1,0$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Combine X_{ij}'s element wise according to φ into X_φ
such that $\mathbb{E}(X_\varphi^k) = \varphi(M)$.

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

$$X_{ag} = 1,1,1,0,0,1$$

$$X_{bg} = 0,0,1,0$$

$$X_{\varphi} = 1,1,0,0$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Compute a.s.-bounds $X_\varphi^k \in [l_\varphi, u_\varphi]$ and empirical estimate $\hat{X}_\varphi$.

Use Hoeffding's inequality to compute output interval around $\hat{X}_\varphi$.

$$X_\varphi = 1,1,0,0$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

$$\mathbb{P}\left(\varphi(M) \in \hat{X}_\varphi \pm \underbrace{\sqrt{-\frac{(u_\varphi - l_\varphi)^2}{2n} \cdot \ln\frac{\delta}{2}}}_{\text{Monitor Output}}\right) \geq 1 - \delta$$

Comments.

Dependent multiplication and division.
Memory due to imbalance in observations.

Bayesian.

Matrix Beta distribution as prior and input has to be in polynomial form.

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Compute φ^2. Split φ and φ^2 into its monomials.

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

$$\mathbb{E}(\varphi) = \mathbb{E}(v_{ag}) - \mathbb{E}(v_{bg})$$

$$\mathbb{E}(\varphi^2) = \mathbb{E}(v_{ag}^2) - 2\mathbb{E}(v_{ag}v_{bg}) + \mathbb{E}(v_{bg}^2)$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Incrementally compute $\mathbb{E}(\xi_i)$ for each monomial.
Aggregate them together according to φ and φ^2.

$$\mathbb{E}(\varphi^2) = \mathbb{E}(v_{ag}^2) - 2\mathbb{E}(v_{ag}v_{bg}) + \mathbb{E}(v_{bg}^2)$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

Use Chebyshev's inequality to compute output interval around $\mathbb{E}(\varphi)$.

$$\mathbb{E}(\varphi^2) = \mathbb{E}(v_{ag}^2) - 2\mathbb{E}(v_{ag}v_{bg}) + \mathbb{E}(v_{bg}^2)$$

Property: $\varphi := v_{ag} - v_{bg}$

Observations: a g b d b d a g b g a g a d b d a d a g

$$\mathbb{P}\left(\varphi(M) \in \underbrace{\mathbb{E}(\varphi) \pm \sqrt{\frac{\mathbb{E}(\varphi^2) - \mathbb{E}(\varphi)^2}{\delta}}}_{\text{Monitor Output}}\right) \geq 1 - \delta$$

Comments.

Result valid w.r.t. to the prior.
Convergence speed depends on the prior.
Chebyshev's inequality is pessimistic.

Resources.

How efficient are the monitors?

Input.

Probabilistic Specification Expression.

	Update-Time	Counters
Frequentist	$\mathcal{O}(n^4 2^{2n})$	$\mathcal{O}(n^4 2^{2n})$
Bayesian	$\mathcal{O}(n^2 2^n)$	$\mathcal{O}(n^2 2^n)$

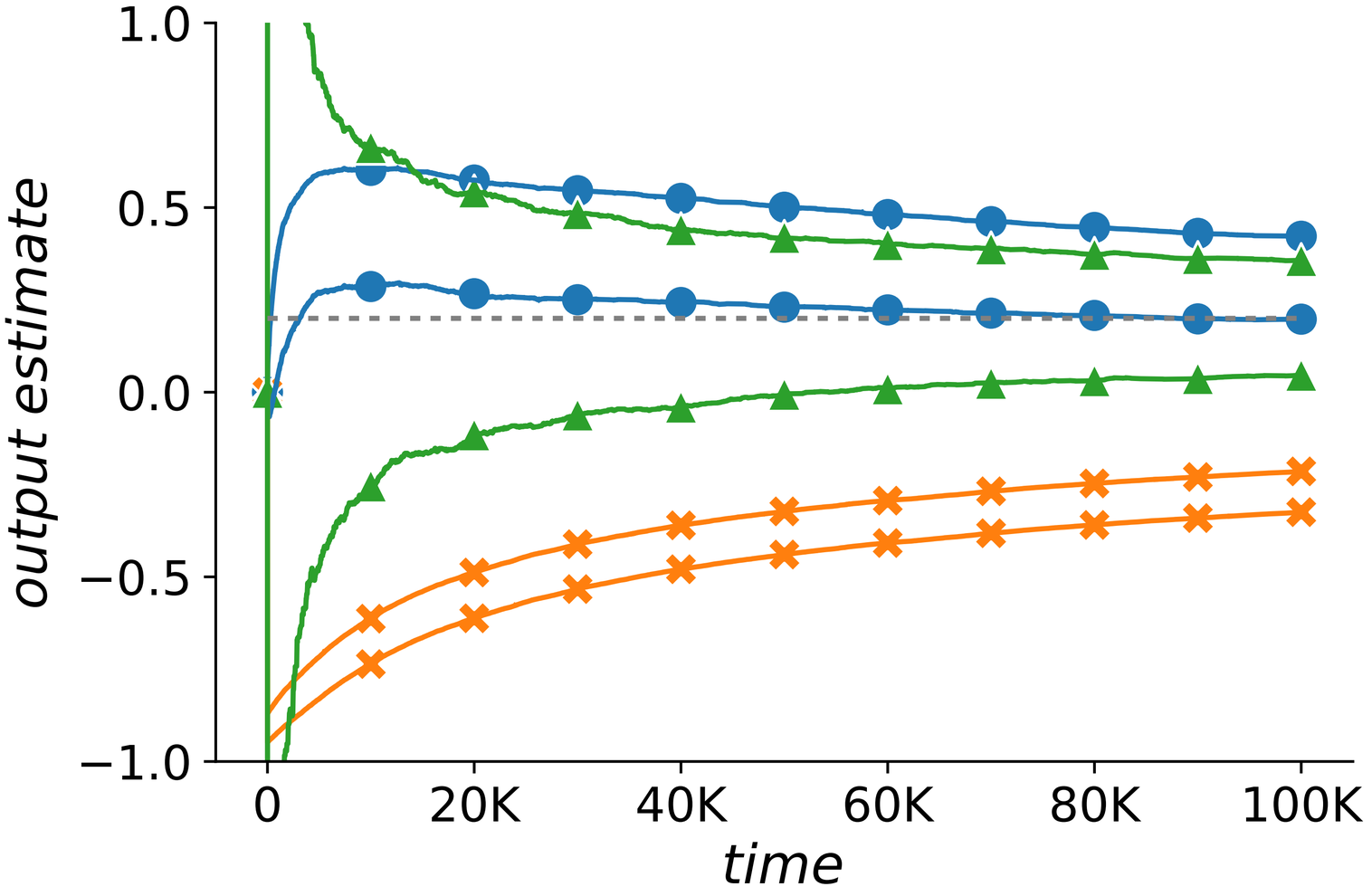
Input.

In special form.

	Update-Time	Counters
Frequentist	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$
Bayesian	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

Experiments.

How do the bounds look in praxis?



Property:

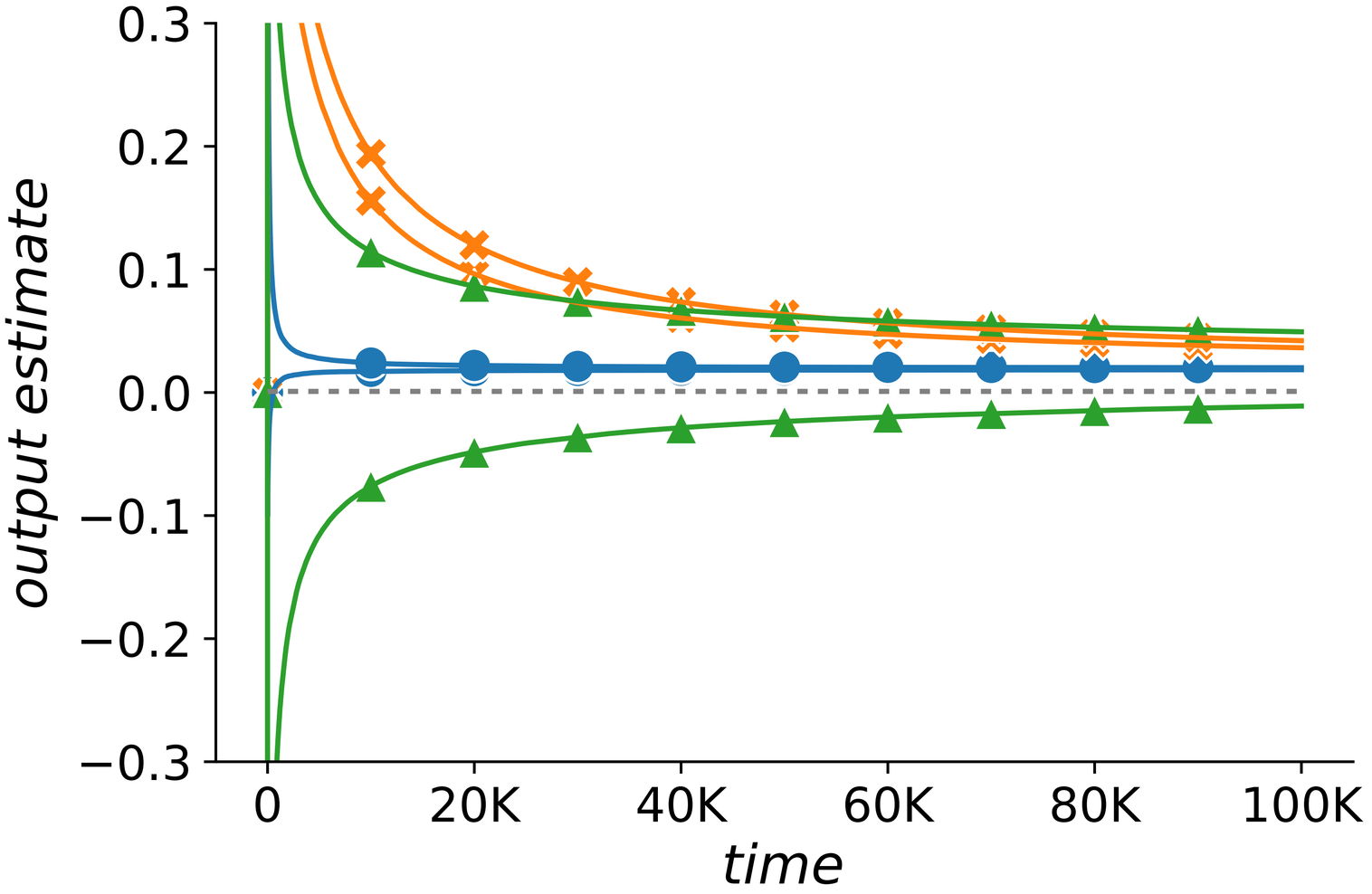
$v_{a1} - v_{b1}$

$n = 1$

Average Execution Time:

Frequentist: $13.0\mu s$

Bayesian: $29.3\mu s$



Property:

$$\Sigma_{i=1}^{10} c_i \cdot v_{ai} - \Sigma_{i=1}^{10} d_i \cdot v_{bi}$$

$$n = 19$$

Average Execution Time:

Frequentist: $53.8\mu s$

Bayesian: $184.6\mu s$

Related Work.

What has been done so far?

Static verification of algorithmic fairness

Albarghouthi, et al. "Fairsquare: probabilistic verification of program fairness.” OOPSLA 2017.
Bastani et al. "Probabilistic verification of fairness properties via concentration." OOPSLA 2019.
Ghosh et al. “Justicia: A stochastic sat approach to formally verify fairness.” AAAI 2021.
Sun, et al. "Probabilistic verification of neural networks against group fairness." FM 2021.
Ghosh, et. al. "Algorithmic fairness verification with graphical models." AAAI 2022.

Monitoring algorithmic fairness

Albarghouthi and Vinitsky. "Fairness-aware programming." FAccT 2019.
Henzinger et al. "Runtime Monitoring of Dynamic Fairness Properties." FAccT 2023.
Henzinger et al. “Monitoring Algorithmic Fairness under Partial Observations.” RV 2023 (to appear).

Summary.

Main points.

Introduced a specification language for fairness properties over Markov chains.

Considered the problem of monitoring fairness from a frequentist and a Bayesian perspective.

Presented two algorithms for monitoring fairness properties on Markov chains.

Institute of Science and Technology Austria