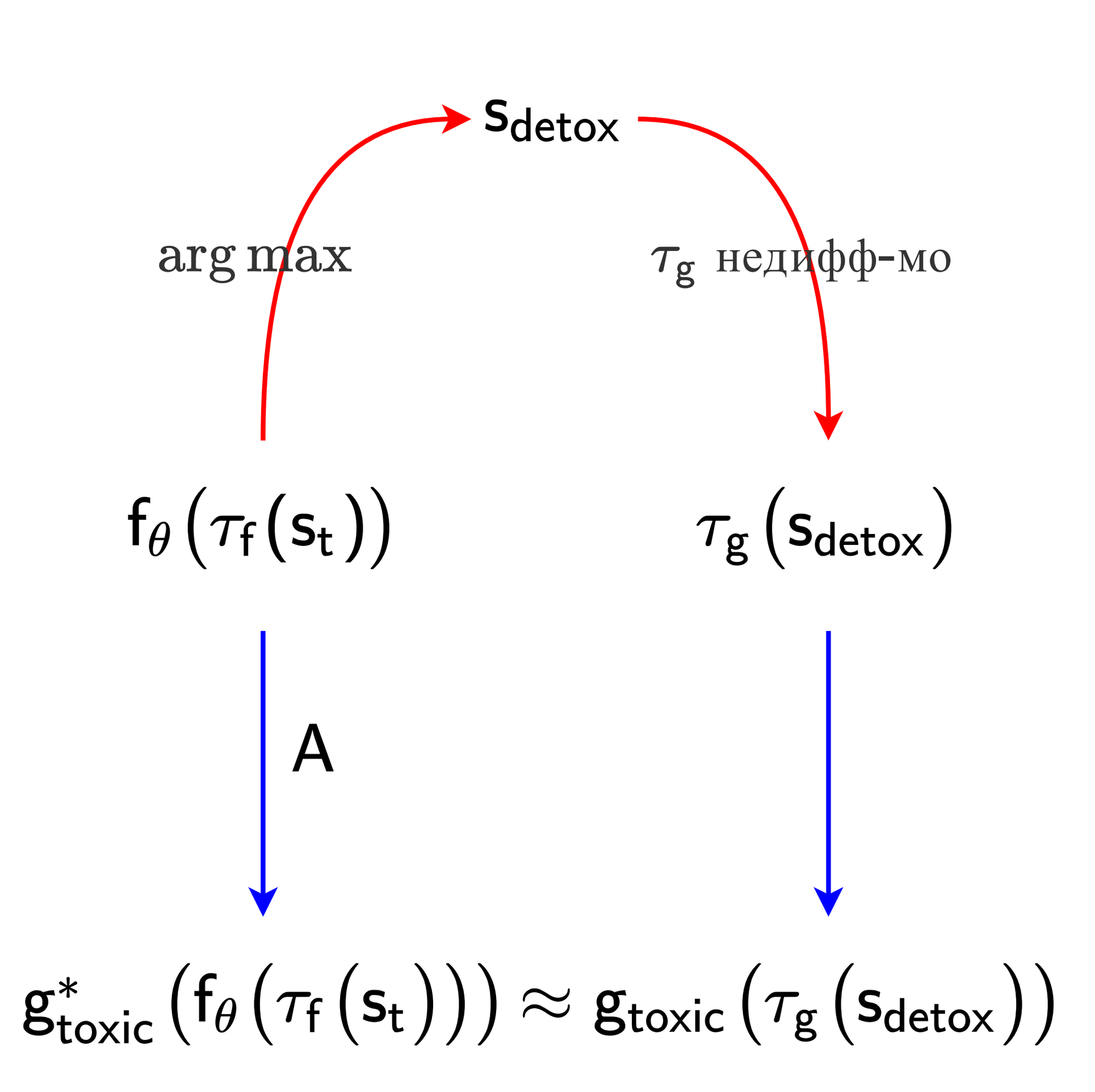
$\mathbf{s}_{\text{detox}}$
$\arg\max$
τ_{g} недифф-мо
$\mathbf{f}_\theta\left(\tau_{\text{f}}\left(\mathbf{s}_{\text{t}}\right)\right)$
$\tau_{\text{g}}\left(\mathbf{s}_{\text{detox}}\right)$
A
$\mathbf{g}^*_{\text{toxic}}\left(\mathbf{f}_\theta\left(\tau_{\text{f}}\left(\mathbf{s}_{\text{t}}\right)\right)\right) \approx \mathbf{g}_{\text{toxic}}\left(\tau_{\text{g}}\left(\mathbf{s}_{\text{detox}}\right)\right)$