Making and Influencing Irreversible Policy Decisions under Preference Heterogeneity

Thomas Houlden* and Maximilian Negele

*University of Oxford, Global Priorities Institute

December 4 2023, *13th GPR Workshop*

Irreversible decisions

Focus: How should we (society) influence policy-makers to implement (irreversible) policies aligned with our interests?

Irreversible decisions

Focus: How should we (society) influence policy-makers to implement (irreversible) policies aligned with our interests?

Motivation: Preference heterogeneity amongst policy-makers seems particularly worrying when it comes to irreversible decisions.

Irreversible decisions

Focus: How should we (society) influence policy-makers to implement (irreversible) policies aligned with our interests?

Motivation: Preference heterogeneity amongst policy-makers seems particularly worrying when it comes to irreversible decisions.

- Irreversible decisions: later policy-makers are 'locked-in'
  - Permitting the release of a risky technology
  - Acts of belligerence

Overview

I am going to do two things today:

1. When will incentives for policy-makers be aligned with decisions that will increase social welfare?

2. How does political efficacy impact social welfare? (a taster)

Environment

Environment

- Policy-makers: $\mathcal{A} = \{A_1, A_2, \ldots, A_T\}$
  - Idiosyncratic bias for $A_t \in \mathcal{A}$: $B_t \sim N(0, \sigma_B^2)$ (public information)

Environment

- Policy-makers: $\mathcal{A} = \{A_1, A_2, \ldots, A_T\}$
  - Idiosyncratic bias for $A_t \in \mathcal{A}$: $B_t \sim N(0, \sigma_B^2)$ (public information)
  - Action spaces for $A_t \in \{A_1, \ldots, A_{N-1}\}$ and A_T respectively are implement policy (y) or delegate:

$$a_{t \neq T} = \begin{cases} \{\text{Delegate}\} \cup \{y \in \mathbb{R}\} & \text{if no } y \\ \{\} & \text{if } y \end{cases} \quad ; \quad a_T = \begin{cases} y \in \mathbb{R} & \text{if no } y \\ \{\} & \text{if } y \end{cases}$$

Environment

- Policy-makers: $\mathcal{A} = \{A_1, A_2, \dots, A_T\}$
  - Idiosyncratic bias for $A_t \in \mathcal{A}$: $B_t \sim N(0, \sigma_B^2)$ (public information)
  - Action spaces for $A_t \in \{A_1, \dots, A_{N-1}\}$ and A_T respectively are implement policy (y) or delegate:

$$a_{t \neq T} = \begin{cases} \{\text{Delegate}\} \cup \{y \in \mathbb{R}\} & \text{if no } y \\ \{\} \quad \text{if } y \end{cases} \quad ; \quad a_T = \begin{cases} y \in \mathbb{R} & \text{if no } y \\ \{\} \quad \text{if } y \end{cases}$$

- An unknown state, S, can take any value in the state space $\Omega = \mathbb{R}$.
  - S is revealed over time through a filtration, $\mathcal{F}$ (public information)
  - $\{\mathcal{F}_t\}_{t \geq 1} \subseteq \mathcal{F}$ is information available to each $A_t \in \mathcal{A}$.

Information and preferences

Preferences over outcomes:

- Policy-makers: policy should reflect the state of the world and their own bias; $u_t(y - (S + B_t)) = -(y - (S + B_t))^2$

Information and preferences

Preferences over outcomes:

- Policy-makers: policy should reflect the state of the world and their own bias; $u_t(y-(S+B_t)) = -(y-(S+B_t))^2$
- Society: policy should reflect the state of the world; $w(y-S) = -(y-S)^2$

Information and preferences

Preferences over outcomes:

- Policy-makers: policy should reflect the state of the world and their own bias; $u_t(y - (S + B_t)) = -(y - (S + B_t))^2$
- Society: policy should reflect the state of the world; $w(y - S) = -(y - S)^2$

Define $\hat{S}_t := \mathbb{E}[S|\mathcal{F}_t]$. For quadratic loss utility we have

$$\hat{S}_t + B_t = \arg\max_y \mathbb{E}[-(y - (S + B_t))^2|\mathcal{F}_t]$$

Information and preferences

Preferences over outcomes:

- Policy-makers: policy should reflect the state of the world and their own bias; $u_t(y - (S + B_t)) = -(y - (S + B_t))^2$
- Society: policy should reflect the state of the world; $w(y - S) = -(y - S)^2$

Define $\hat{S}_t := \mathbb{E}[S|\mathcal{F}_t]$. For quadratic loss utility we have

$$\hat{S}_t + B_t = \arg\max_y \mathbb{E}[-(y - (S + B_t))^2|\mathcal{F}_t]$$

We define $y_t := \hat{S}_t + B_t$, i.e., A_t's policy, conditional on implementing.

We also have

- $\mathbb{E}[|\hat{S}_t - S|] \geq \mathbb{E}[|\hat{S}_{t'} - S|]$ for all $t' > t$; and
- $\mathbb{E}[\hat{S}_t - S] = 0$ for all $A_t \in \mathcal{A}$

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,imp.}} \geq \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,del.}}$$

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,imp.}} \geq \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,del.}}$$

$$\implies \underbrace{\mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2]}_{\text{Info gains}} \leq \underbrace{\mathbb{E}[(B_2 - B_1)^2]}_{\text{Pref. cost}}$$

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,imp.}} \geq \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,del.}}$$

$$\implies \underbrace{\mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2]}_{\text{Info gains}} \leq \underbrace{\mathbb{E}[(B_2 - B_1)^2]}_{\text{Pref. cost}}$$

Result (Bias thresholds)

For each $A_t \in \mathcal{A}$, there exist a $\overline{B}_t$ such that if $|B_t| \geq \overline{B}_t$, A_t implements; if $|B_t| < \overline{B}_t$, A_t delegates.

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}\left[\left((\hat{S}_1 + B_1) - (S + B_1)\right)^2\right]}_{\mathbb{E}u_{1,imp.}} \geq \underbrace{-\mathbb{E}\left[\left((\hat{S}_2 + B_2) - (S + B_1)\right)^2\right]}_{\mathbb{E}u_{1,del.}}$$

$$\implies \underbrace{\mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2]}_{\text{Info gains}} \leq \underbrace{\mathbb{E}[(B_2 - B_1)^2]}_{\text{Pref. cost}}$$

Result (Bias thresholds)

For each $A_t \in \mathcal{A}$, there exist a $\overline{B}_t$ such that if $|B_t| \geq \overline{B}_t$, A_t implements; if $|B_t| < \overline{B}_t$, A_t delegates.

Therefore:

$$\overline{B}_1 := \sqrt{\max\{z, 0\}} \quad \text{where } z = \mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2] - \sigma_B^2$$

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,imp.}} \;\geq\; \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,del.}}$$

$$\implies \underbrace{\mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2]}_{\text{Info gains}} \;\leq\; \underbrace{\mathbb{E}[(B_2 - B_1)^2]}_{\text{Pref. cost}}$$

Result (Bias thresholds)

For each $A_t \in \mathcal{A}$, there exist a $\overline{B}_t$ such that if $|B_t| \geq \overline{B}_t$, A_t implements; if $|B_t| < \overline{B}_t$, A_t delegates.

Therefore:

$$\overline{B}_1 := \sqrt{\max\{z, 0\}} \quad \text{where } z = \mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2] - \sigma_B^2$$

This seems concerning!

Policy-maker incentives

Suppose $T = 2$. Decision for A_1 is implement (y_1) when

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,imp.}} \geq \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - (S + B_1)\big)^2]}_{\mathbb{E}u_{1,del.}}$$

$$\implies \underbrace{\mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2]}_{\text{Info gains}} \leq \underbrace{\mathbb{E}[(B_2 - B_1)^2]}_{\text{Pref. cost}}$$

Result (Bias thresholds)

For each $A_t \in \mathcal{A}$, there exist a $\overline{B}_t$ such that if $|B_t| \geq \overline{B}_t$, A_t implements; if $|B_t| < \overline{B}_t$, A_t delegates.

Therefore:

$$\overline{B}_1 := \sqrt{\max\{z, 0\}} \quad \text{where } z = \mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2] - \sigma_B^2$$

This seems concerning! ... but

Policy-maker incentives

How does $\overline{B}$ change over time?

$$\begin{tikzcd}
A_1 \arrow[rr, "\text{Delegate}"] \arrow[d, "\text{Imp.}"] & & A_2 \arrow[rr, "\text{Delegate}"] \arrow[d, "\text{Imp.}"] & & A_3 \arrow[d, "\text{Imp.}"] \\
-\mathbb{E}[(\hat{S}_1 - S)^2] & \leq & -\mathbb{E}[(\hat{S}_2 - S)^2] & \leq & -\mathbb{E}[(\hat{S}_3 - S)^2]
\end{tikzcd}$$

Policy-maker incentives

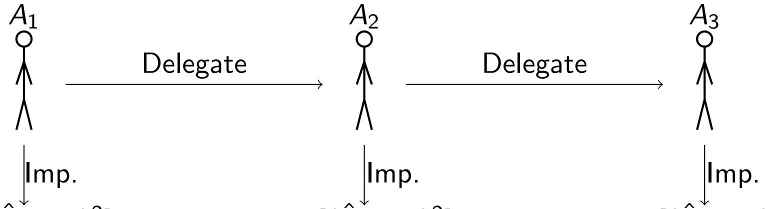
How does $\overline{B}$ change over time?

$$
\begin{tikzcd}
A_1 \arrow[r, "\text{Delegate}"] \arrow[d, "\text{Imp.}"] & A_2 \arrow[r, "\text{Delegate}"] \arrow[d, "\text{Imp.}"] & A_3 \arrow[d, "\text{Imp.}"] \\
-\mathbb{E}[(\hat{S}_1 - S)^2] \arrow[r, phantom, "\leq"] & -\mathbb{E}[(\hat{S}_2 - S)^2] \arrow[r, phantom, "\leq"] & -\mathbb{E}[(\hat{S}_3 - S)^2]
\end{tikzcd}
$$

Since policy-makers' 'implementation option' becomes better over time, then $\implies \overline{B}_1 \geq \overline{B}_2 \geq \overline{B}_3$.

Policy-maker incentives

How does $\overline{B}$ change over time?



$$-\mathbb{E}[(\hat{S}_1 - S)^2] \quad \leq \quad -\mathbb{E}[(\hat{S}_2 - S)^2] \quad \leq \quad -\mathbb{E}[(\hat{S}_3 - S)^2]$$

Since policy-makers' 'implementation option' becomes better over time, then $\implies \overline{B}_1 \geq \overline{B}_2 \geq \overline{B}_3$.

More generally

Proposition (Thresholds weakly decline)

For each $A_t \in \mathcal{A}$*, each decision threshold* $\overline{B}_t \in \{\overline{B}_1, \ldots, \overline{B}_T\}$ *we have* $\overline{B}_t \geq \overline{B}_{t'}$ *if* $t' > t$.

Social welfare

In the $T = 2$ example, society prefers A_1 to implement if

$$\underbrace{-\mathbb{E}[((\hat{S}_1 + B_1) - S)^2]}_{w \text{ under } A_1 \text{ implement}} \geq \underbrace{-\mathbb{E}[((\hat{S}_2 + B_2) - S)^2]}_{w \text{ under } A_2 \text{ implement}}$$

Just as we had $\overline{B}_t$, we have $\overline{B}_t^S$ where society prefers A_t to implement a policy if $|B_t| < \overline{B}_t^S$

Social welfare

In the $T = 2$ example, society prefers A_1 to implement if

$$\underbrace{-\mathbb{E}[\big((\hat{S}_1 + B_1) - S\big)^2]}_{w \text{ under } A_1 \text{ implement}} \geq \underbrace{-\mathbb{E}[\big((\hat{S}_2 + B_2) - S\big)^2]}_{w \text{ under } A_2 \text{ implement}}$$

Just as we had $\overline{B}_t$, we have $\overline{B}_t^S$ where society prefers A_t to implement a policy if $|B_t| < \overline{B}_t^S$

Recalling policy-maker incentives:

$$-\mathbb{E}[(\hat{S}_1 - S)^2] \geq -\mathbb{E}[(\hat{S}_2 + B_2 - (S + B_1))^2]$$

Social welfare

In the $T = 2$ example, society prefers A_1 to implement if

$$\underbrace{-\mathbb{E}[((\hat{S}_1 + B_1) - S)^2]}_{w \text{ under } A_1 \text{ implement}} \geq \underbrace{-\mathbb{E}[((\hat{S}_2 + B_2) - S)^2]}_{w \text{ under } A_2 \text{ implement}}$$

Just as we had $\overline{B}_t$, we have $\overline{B}_t^S$ where society prefers A_t to implement a policy if $|B_t| < \overline{B}_t^S$

Recalling policy-maker incentives:

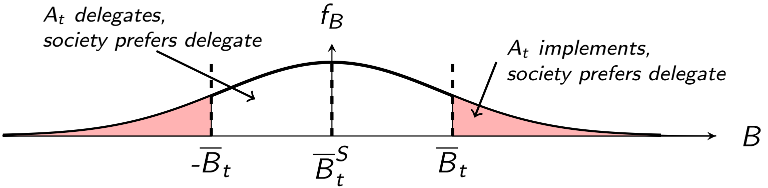
$$-\mathbb{E}[(\hat{S}_1 - S)^2] \geq -\mathbb{E}[(\hat{S}_2 + B_2 - (S + B_1))^2]$$

$\overline{B}_1^S := \sqrt{\max\{-z, 0\}}$ (where $z = \mathbb{E}[(\hat{S}_1 - S)^2] - \mathbb{E}[(\hat{S}_2 - S)^2] - \sigma_B^2$)

Society and policy-makers preference alignment

Result (Unaligned for $|B_t| > \max\{\overline{B}_t, \overline{B}_t^S\}$)

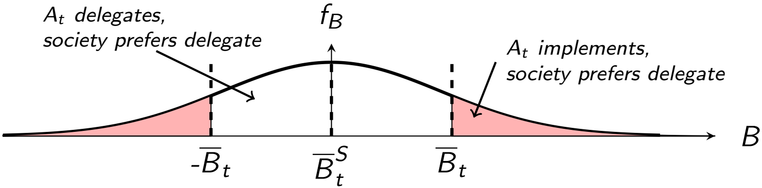
- *If* $\overline{B}_t > 0$, *then* $\overline{B}_t^S = 0$ *(society always prefers delegate); and*



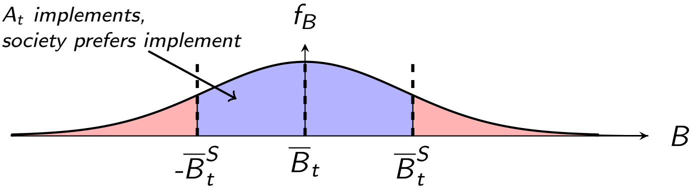
Society and policy-makers preference alignment

Result (Unaligned for $|B_t| > \max\{\overline{B}_t, \overline{B}^S_t\}$)

- *If $\overline{B}_t > 0$, then $\overline{B}^S_t = 0$ (society always prefers delegate); and*



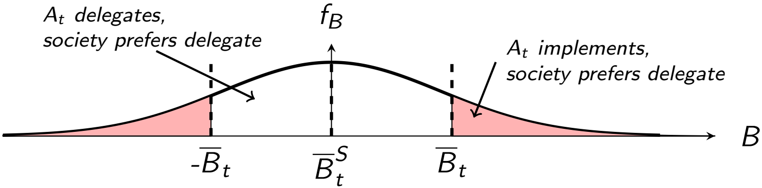
- *if $\overline{B}^S_t > 0$, then $\overline{B}_t = 0$ (PMs always prefer implement)*



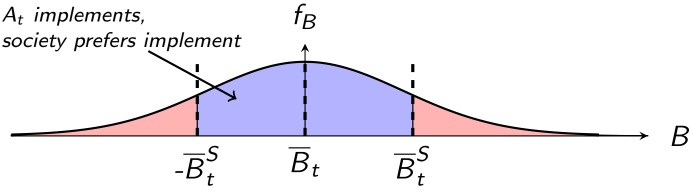
Society and policy-makers preference alignment

Result (Unaligned for $|B_t| > \max\{\overline{B}_t, \overline{B}_t^S\}$)

- *If $\overline{B}_t > 0$, then $\overline{B}_t^S = 0$ (society always prefers delegate); and*



- *if $\overline{B}_t^S > 0$, then $\overline{B}_t = 0$ (PMs always prefer implement)*



Society should be concerned about policy-makers being overly eager to implement irreversible policies.

Society with political efficacy

In reality, society can influence policy-makers.

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

We suppose that society can impose a punishment C (with $C > 0$ and finite) on policy-makers.

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

We suppose that society can impose a punishment C (with $C > 0$ and finite) on policy-makers.

Punishment can make decisions more informed by incentivising delegation

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

We suppose that society can impose a punishment C (with $C > 0$ and finite) on policy-makers.

Punishment can make decisions more informed by incentivising delegation

... but it may also increase the bias of the ultimate policy-maker.

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

We suppose that society can impose a punishment C (with $C > 0$ and finite) on policy-makers.

Punishment can make decisions more informed by incentivising delegation

... but it may also increase the bias of the ultimate policy-maker.

Analogy: partial dose of antibiotics.

Society with political efficacy

In reality, society can influence policy-makers.

But it seems like this power is often limited.

We suppose that society can impose a punishment C (with $C > 0$ and finite) on policy-makers.

Punishment can make decisions more informed by incentivising delegation

... but it may also increase the bias of the ultimate policy-maker.

Analogy: partial dose of antibiotics.

Claim

Political efficacy may reduce social welfare (early stage results).

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

But allowing society to impose costs impacts both $\mathbf{p}_i$ and $\mathbb{E}y_i|\text{Imp.}$:

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

But allowing society to impose costs impacts both $\mathbf{p}_i$ and $\mathbb{E}y_i|\text{Imp.}$:

- $\mathbb{E}[B_i(C)\big||B_i| \geq \overline{B}_i(C)]$: require extreme preferences

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

But allowing society to impose costs impacts both $\mathbf{p}_i$ and $\mathbb{E}y_i|\text{Imp.}$:

- $\mathbb{E}[B_i(C)\big||B_i| \geq \overline{B}_i(C)]$: require extreme preferences
- $\mathbf{p}_i(C)$: probability density shifts to later entries in the vector

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

But allowing society to impose costs impacts both $\mathbf{p}_i$ and $\mathbb{E}y_i|\text{Imp.}$:

- $\mathbb{E}[B_i(C)\big||B_i| \geq \overline{B}_i(C)]$: require extreme preferences
- $\mathbf{p}_i(C)$: probability density shifts to later entries in the vector

Decomposing effect of C:

- Biasing effect

$$\mathbf{p}_i(C) \cdot \frac{d}{dC}[\mathbb{E}\mathbf{w}(y_i(C) - S)] \leq 0$$

Evaluating social outcomes: hollowing out

Society prefers A_t to delegate (so some $A_i \in \{A_{t+1}, \ldots, A_T\}$ implements) when

$$\mathbb{E}[w(y_t - S)] \leq \underbrace{\mathbf{p}_i}_{\text{prob. vector}} \cdot \underbrace{\mathbb{E}[\mathbf{w}(y_i|\text{Imp.} - S)]}_{\text{welfare vector}}$$

But allowing society to impose costs impacts both $\mathbf{p}_i$ and $\mathbb{E}y_i|\text{Imp.}$:

- $\mathbb{E}[B_i(C)| |B_i| \geq \overline{B}_i(C)]$: require extreme preferences
- $\mathbf{p}_i(C)$: probability density shifts to later entries in the vector

Decomposing effect of C:

- Biasing effect

$$\mathbf{p}_i(C) \cdot \frac{d}{dC}[\mathbb{E}\mathbf{w}(y_i(C) - S)] \leq 0$$

- Delaying effect

$$\mathbb{E}[\mathbf{w}(y_i(C) - S)] \cdot \frac{d}{dC}\mathbf{p}_i(C) \geq 0$$