Reasoning about Causality in Games

Lewis Hammond

Department of Computer Science / Future of Humanity Institute / Hertford College
University of Oxford

This is joint work with several others!

James Fox
(Oxford)

Tom Everitt
(DeepMind)

Ryan Carey
(Oxford / FHI)

Alessandro Abate
(Oxford)

Michael Wooldridge
(Oxford)

Outline

Outline

- Introduction

Outline

- Introduction
  - Motivation

Outline

- Introduction
  - Motivation
  - Background

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games

Outline

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions
- Counterfactuals

Outline

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions
  - Counterfactuals
- Additional Topics
  - Game-Theoretic Reasoning

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions
  - Counterfactuals
- Additional Topics
  - Game-Theoretic Reasoning
  - Other Models

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions
  - Counterfactuals
- Additional Topics
  - Game-Theoretic Reasoning
  - Other Models
  - Applications

Outline

- Introduction
  - Motivation
  - Background
- Representing Strategic Dependencies
  - Extended Models
- Answering Queries
  - A Causal Hierarchy for Games
  - Predictions
  - Interventions
  - Counterfactuals
- Additional Topics
  - Game-Theoretic Reasoning
  - Other Models
  - Applications
- So What?

Outline

Outline

Introduction

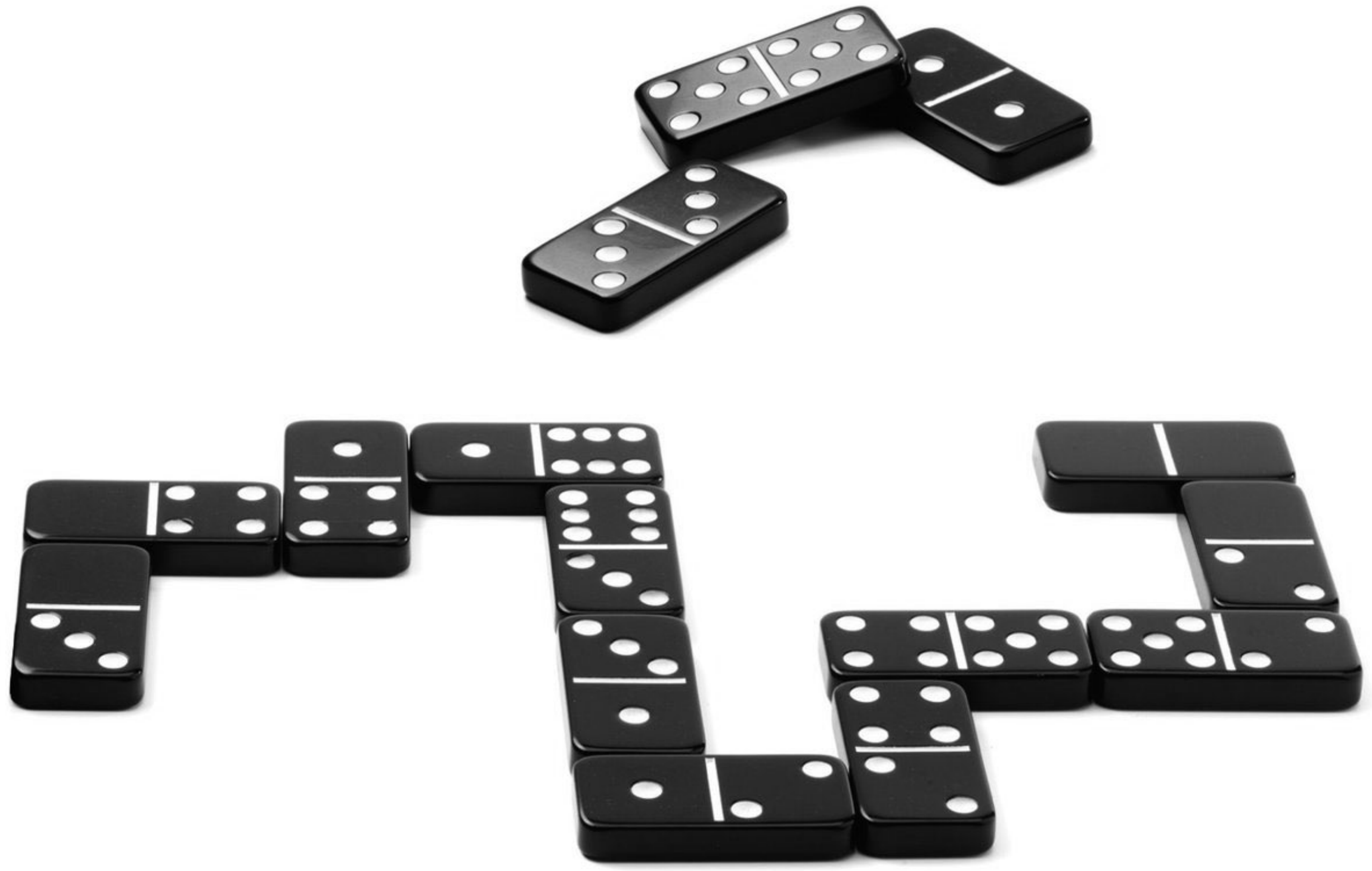
Motivation

Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking

Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking



Motivation

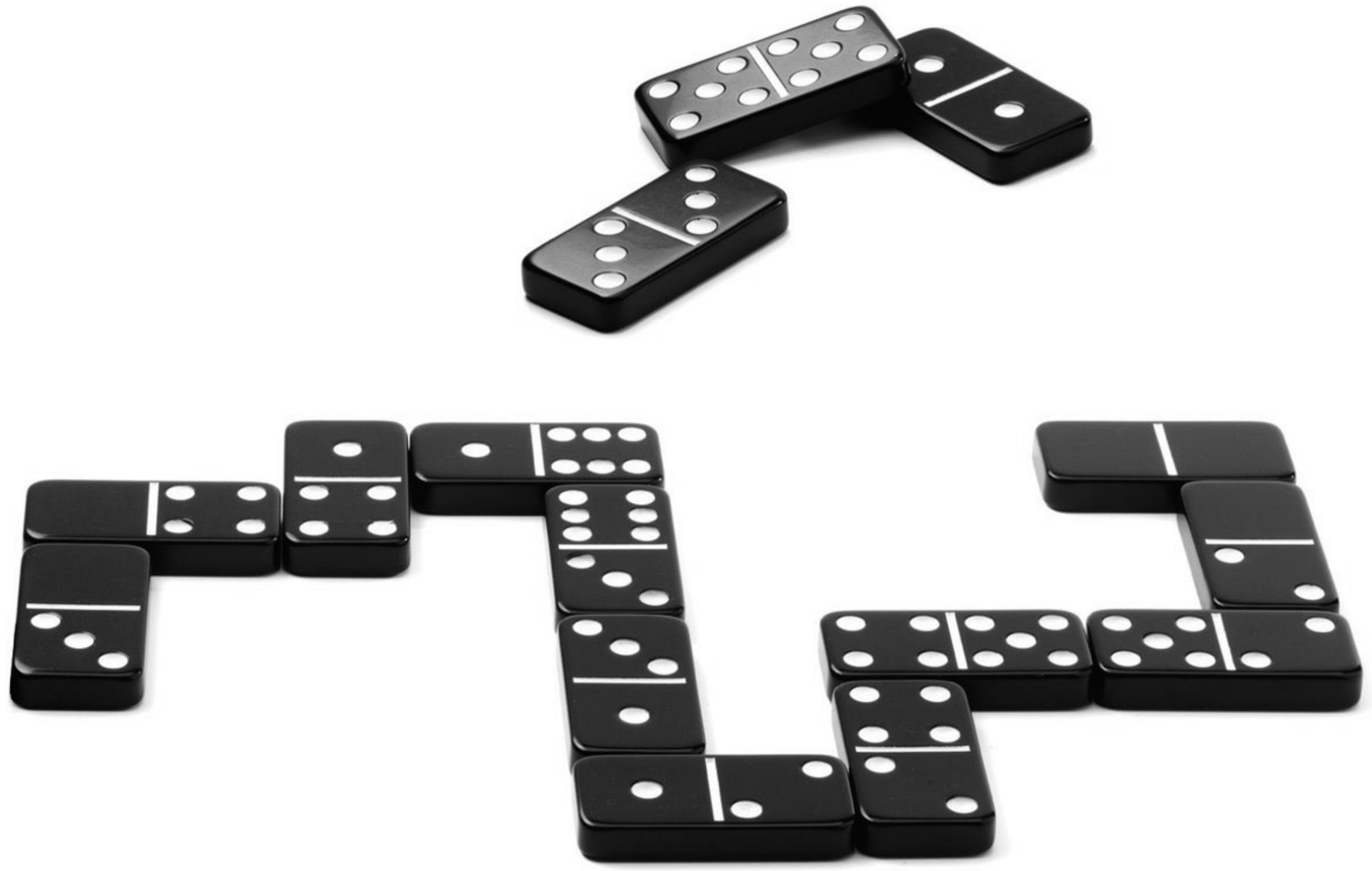
- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking

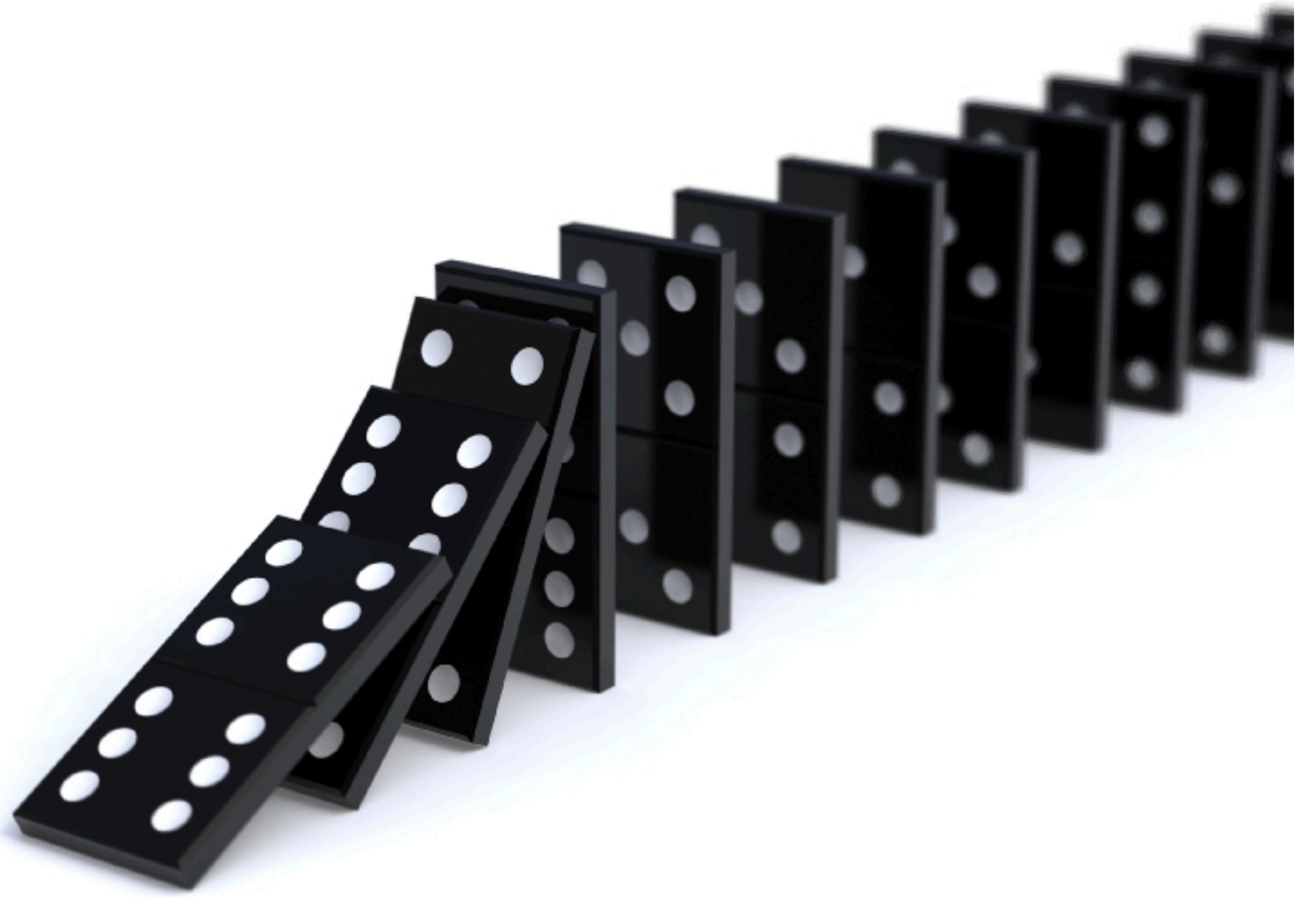
Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking
- Key questions:

Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking
- Key questions:
    1. How should we represent strategic dependencies in games?





Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking
- Key questions:
  1. How should we represent strategic dependencies in games?
  2. How can we answer causal queries in games?

Motivation

- Despite much previous work, a general, principled framework for reasoning about causality in strategic settings is lacking
- Key questions:
  1. How should we represent strategic dependencies in games?
  2. How can we answer causal queries in games?
  3. How does what we propose relate to other formalisms?

Background

Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- Example: Job market signalling [16]

Background

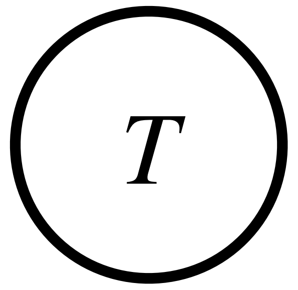
- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)

Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

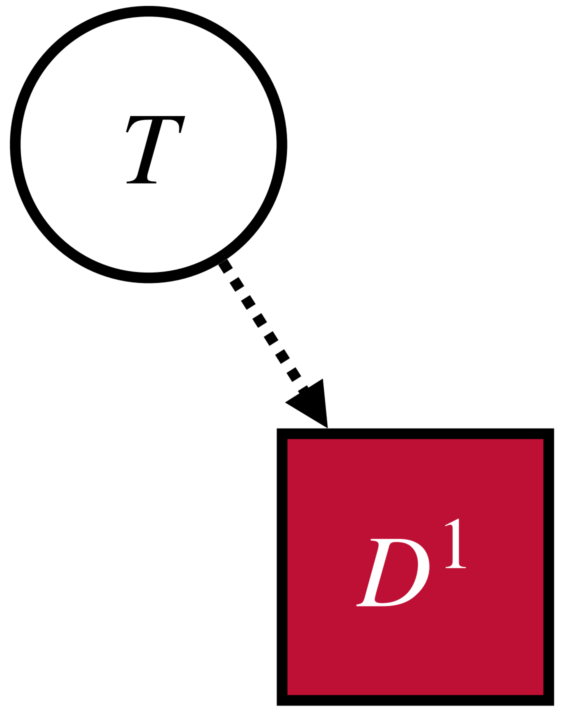
- <u>Example:</u> Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

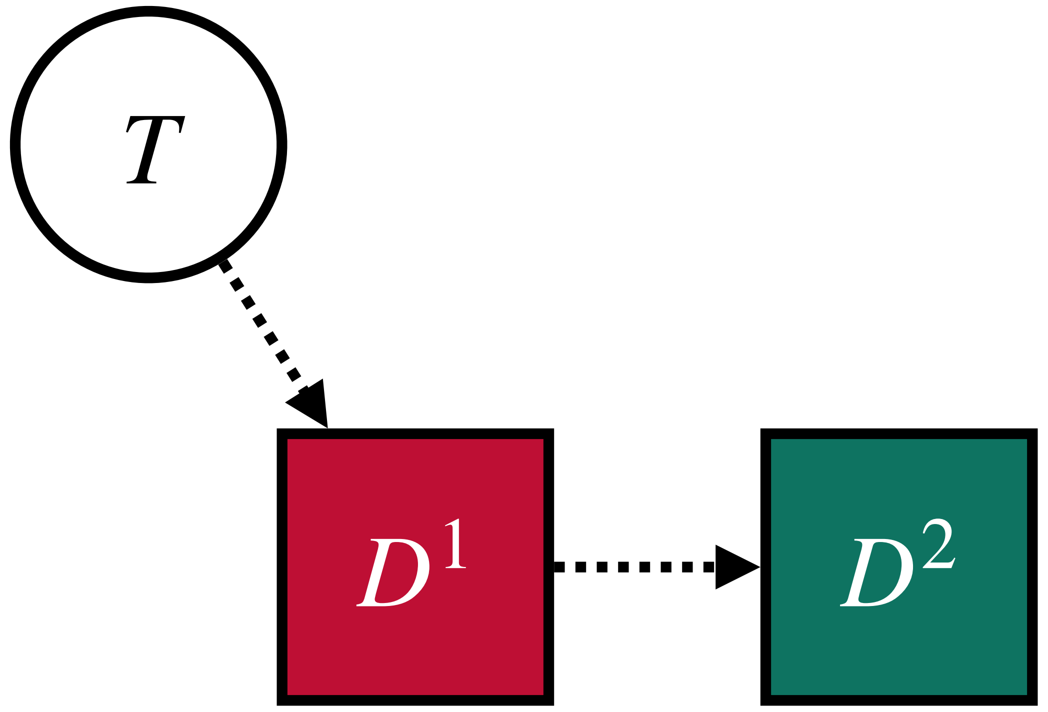
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

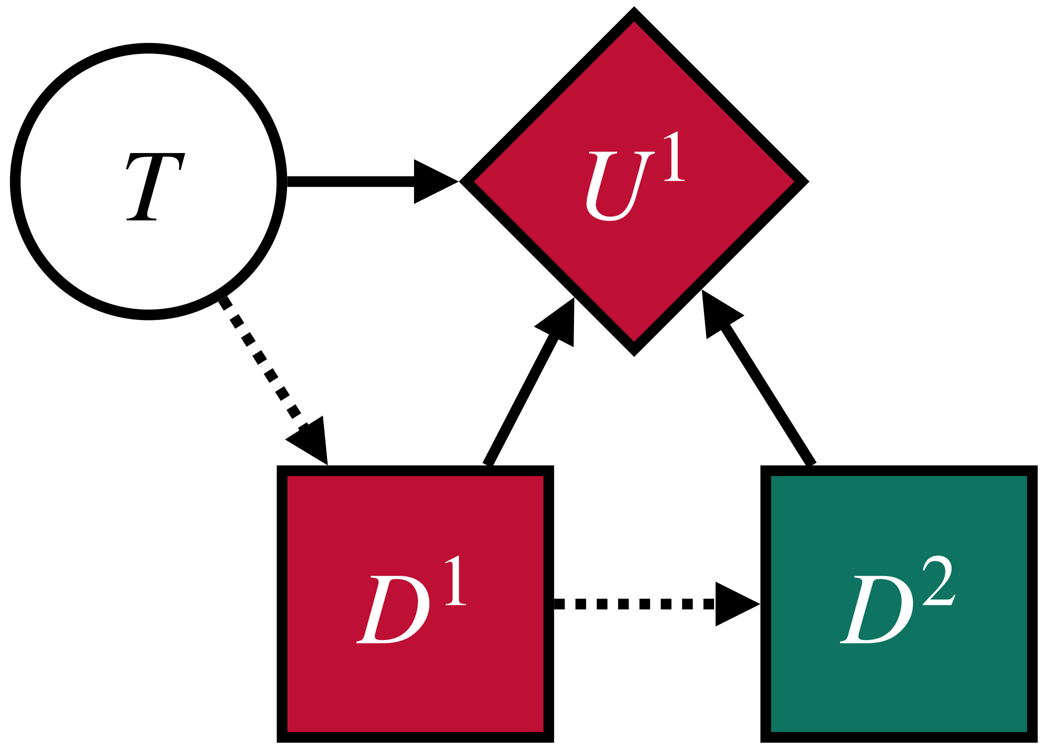
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

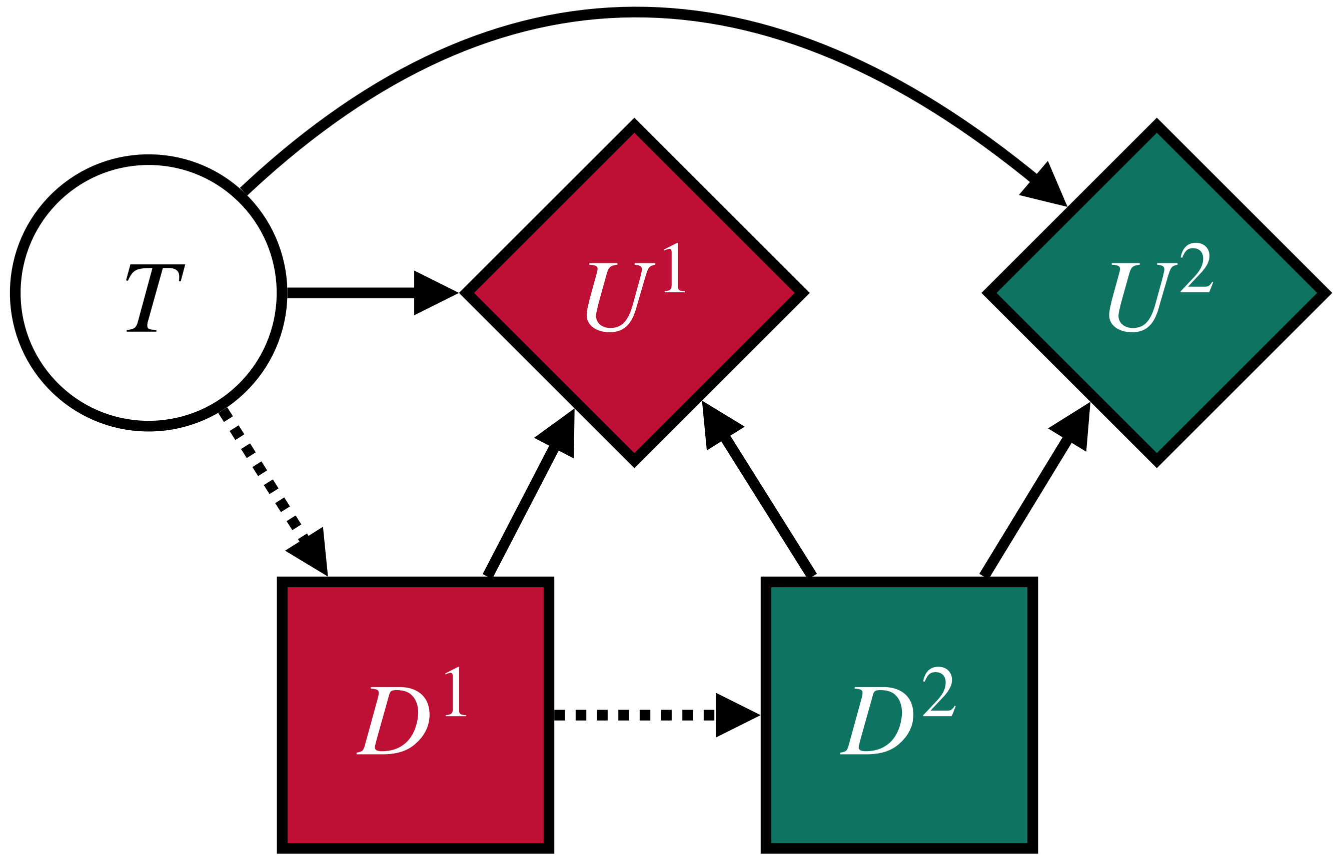
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]

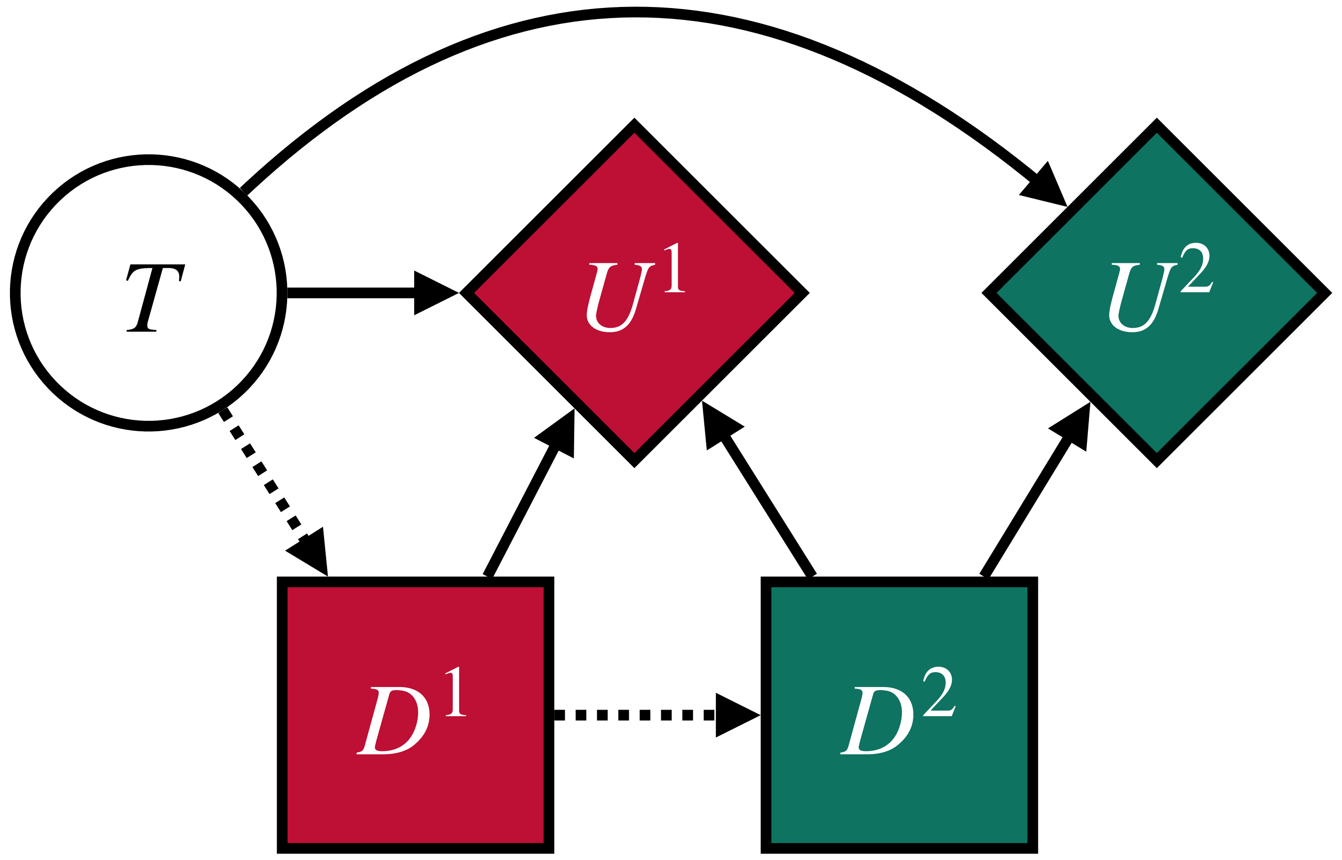
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathscr{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:

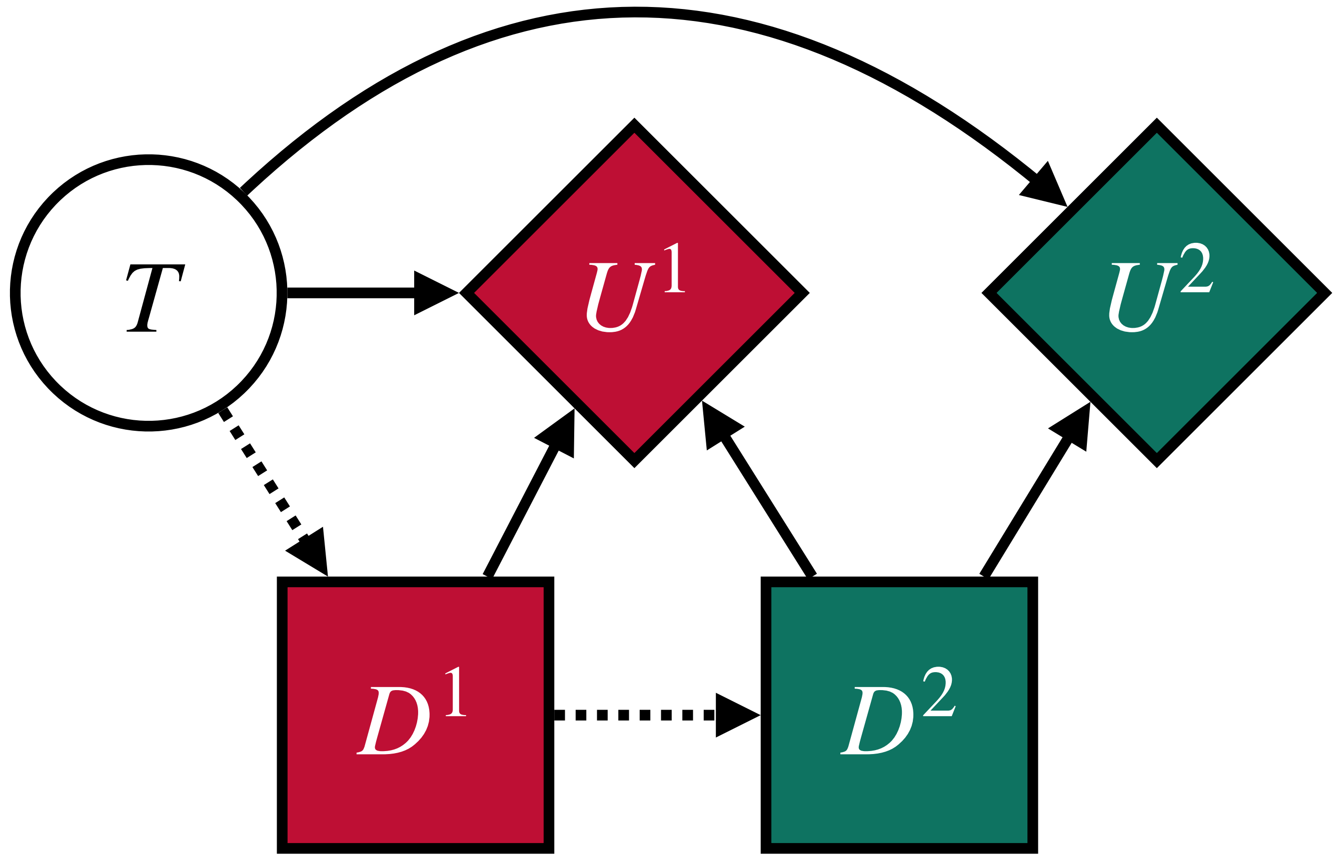
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1, \ldots, n\}$

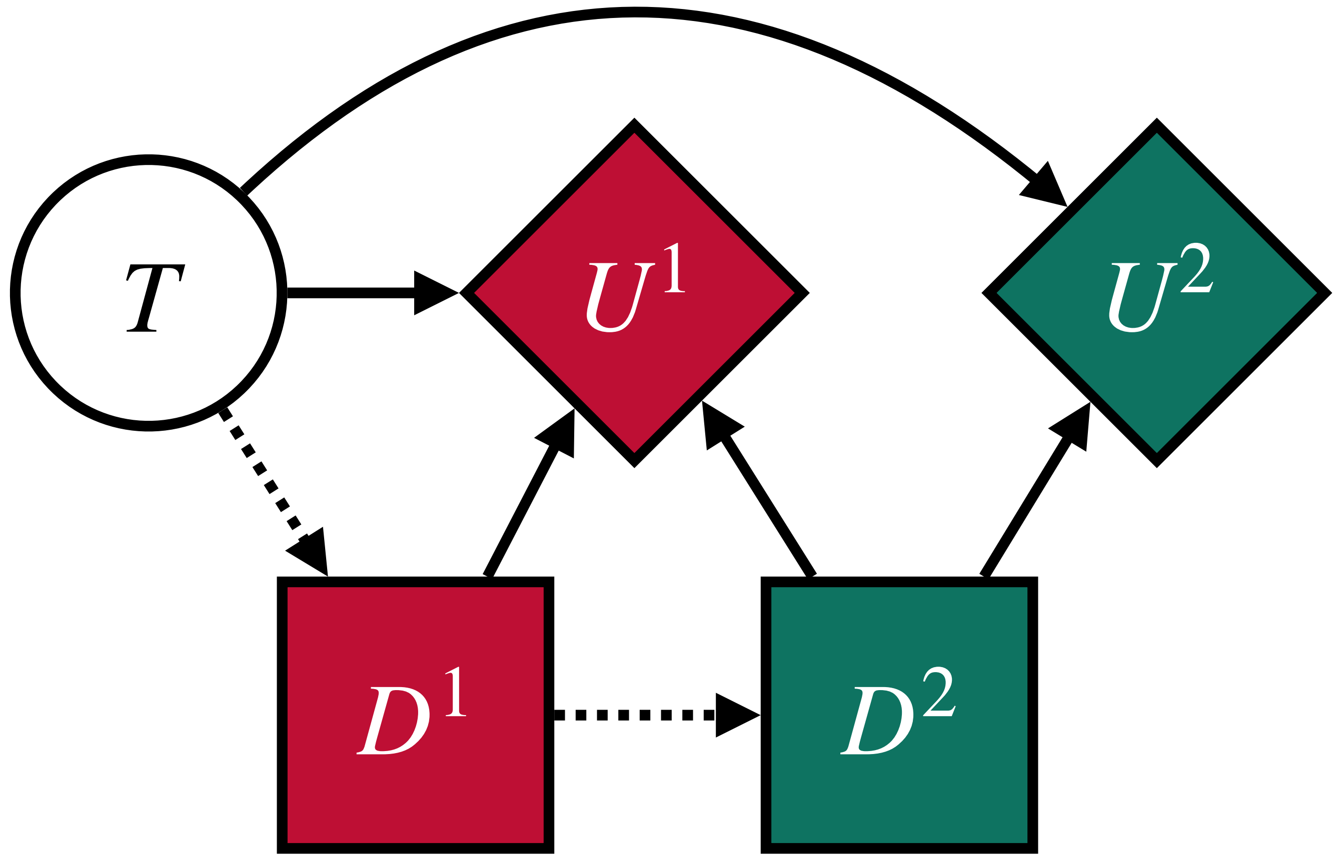
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

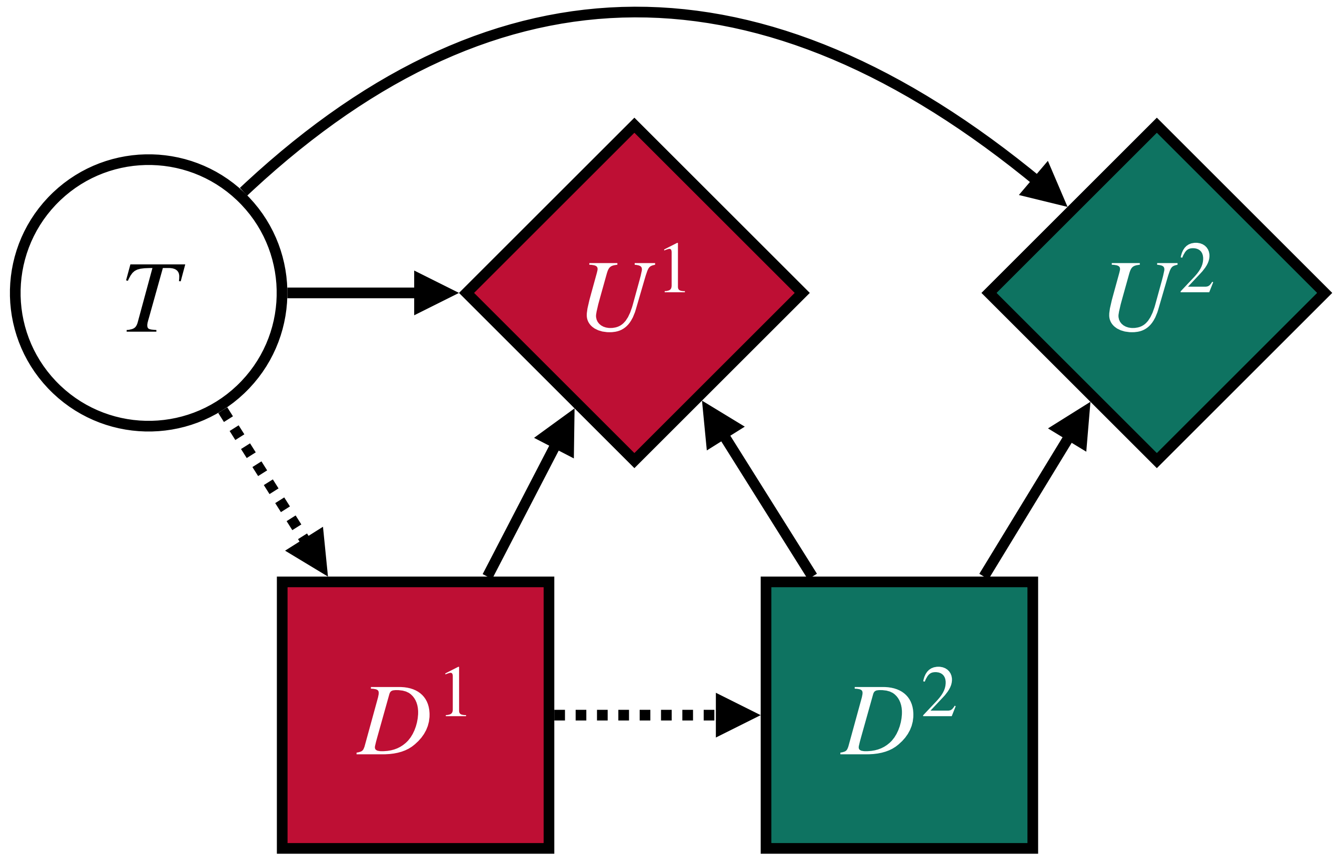
- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1, \ldots, n\}$
  - $\mathbf{V} = \mathbf{X} \cup \bigcup_{i \in N} \mathbf{D}^i \cup \bigcup_{i \in N} \mathbf{U}^i$

- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

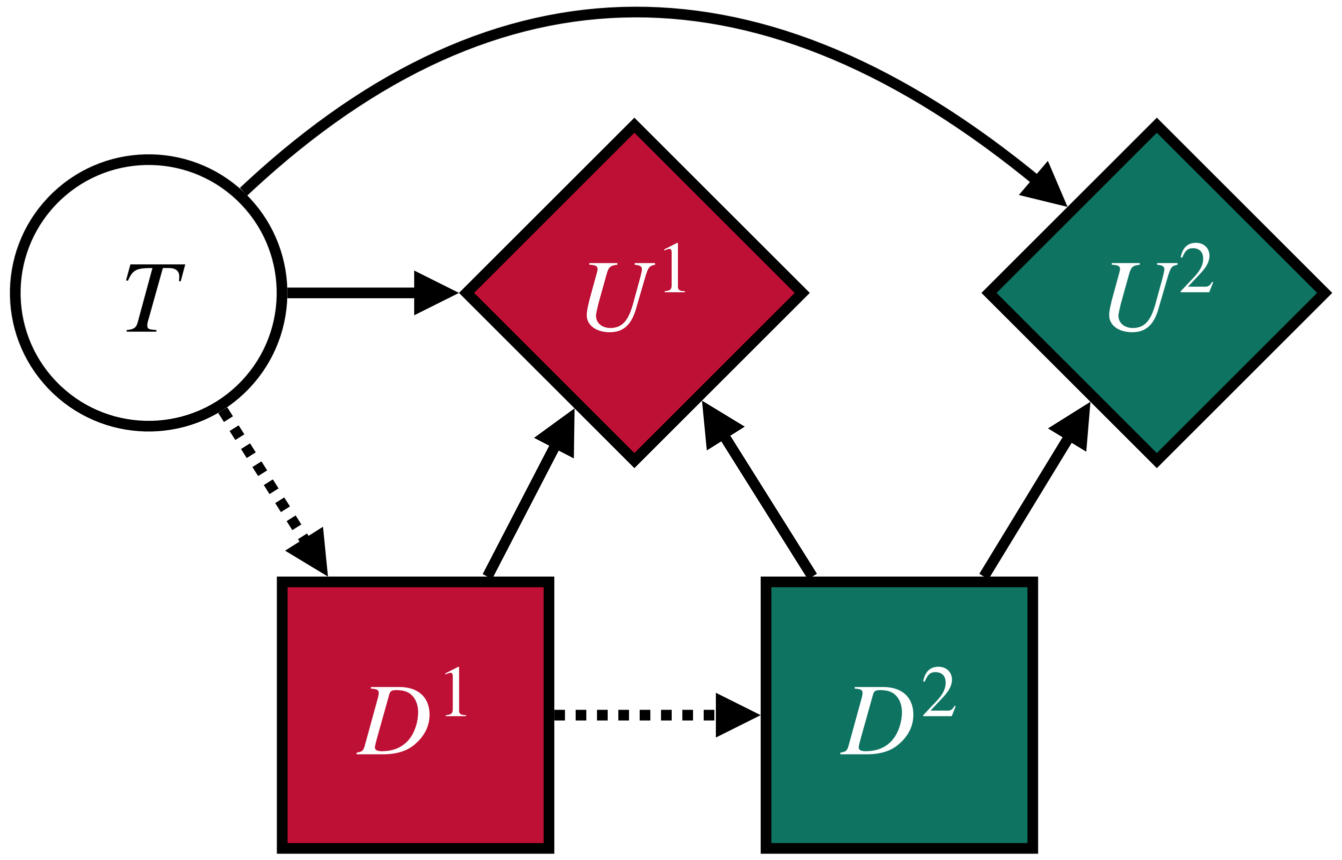
- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1, \ldots, n\}$
  - $\mathbf{V} = \mathbf{X} \cup \bigcup_{i \in N} \mathbf{D}^i \cup \bigcup_{i \in N} \mathbf{U}^i$
  - $\mathbb{E} \subset \mathbf{V} \times \mathbf{V}$
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1, \ldots, n\}$
  - $\mathbf{V} = \mathbf{X} \cup \bigcup_{i \in N} \mathbf{D}^i \cup \bigcup_{i \in N} \mathbf{U}^i$
  - $\mathbb{E} \subset \mathbf{V} \times \mathbf{V}$
- A MAIM $\mathcal{M} = (\mathcal{G}, \theta)$ consists of:

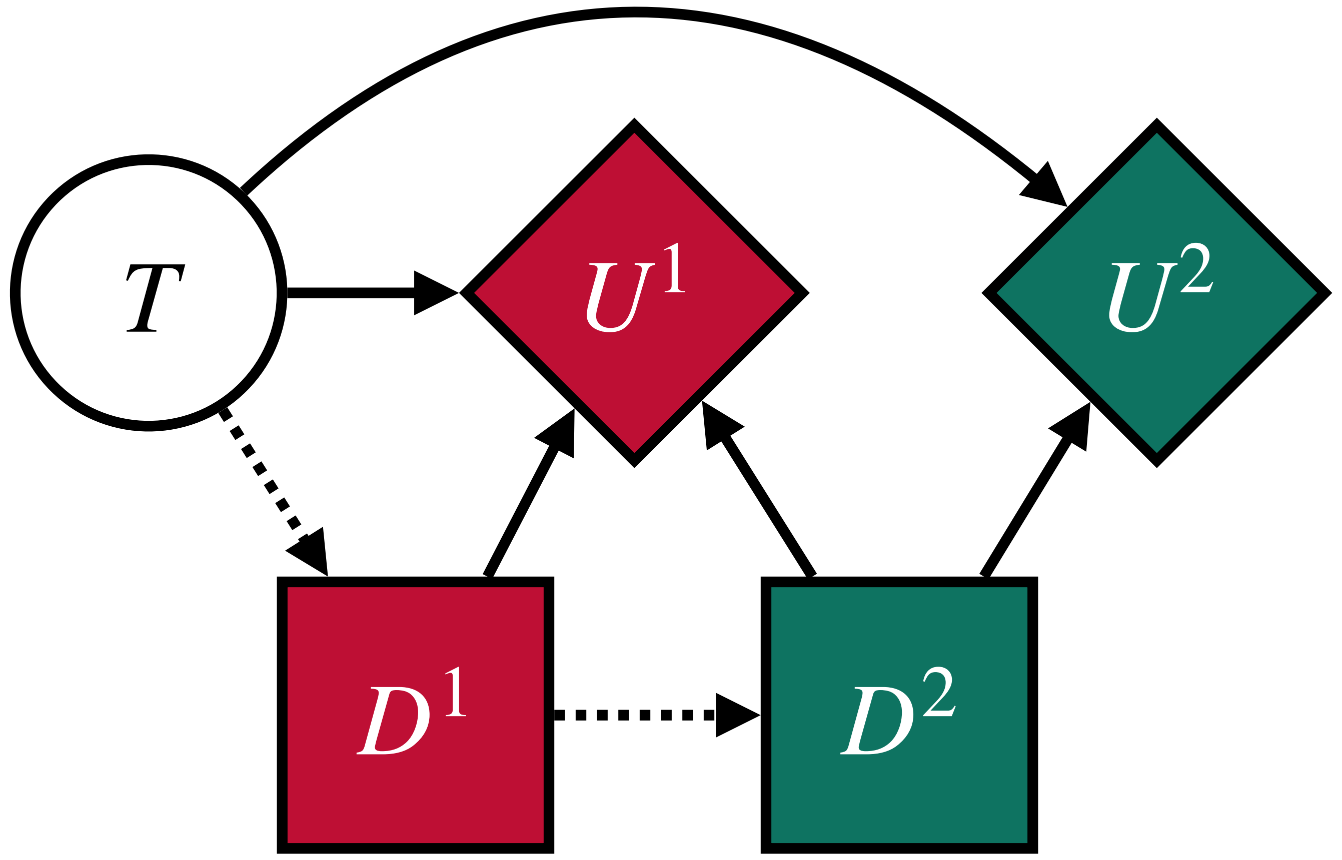
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1, \ldots, n\}$
  - $\mathbf{V} = \mathbf{X} \cup \bigcup_{i \in N} \mathbf{D}^i \cup \bigcup_{i \in N} \mathbf{U}^i$
  - $\mathbb{E} \subset \mathbf{V} \times \mathbf{V}$
- A MAIM $\mathcal{M} = (\mathcal{G}, \theta)$ consists of:
  - A MAID $\mathcal{G}$

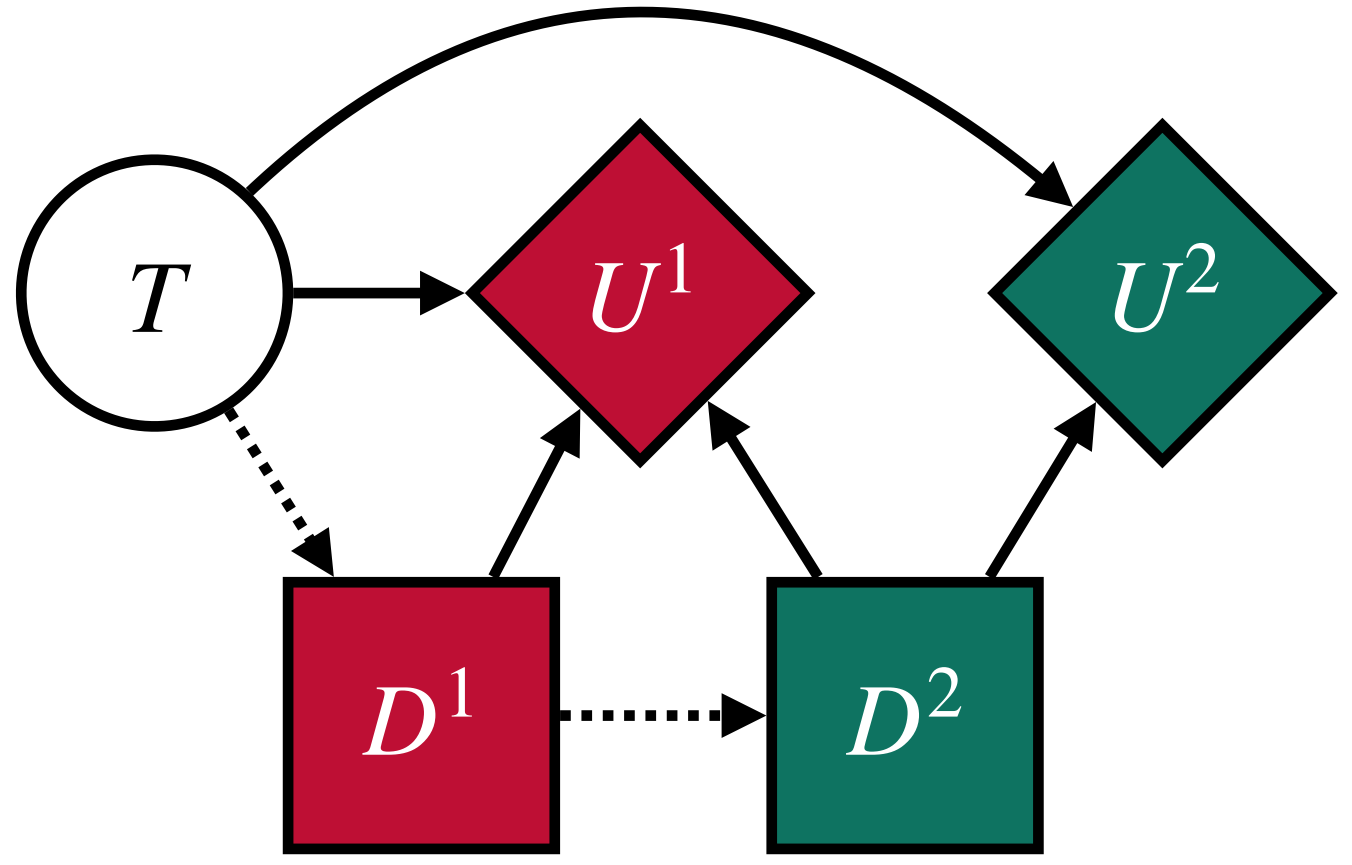
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

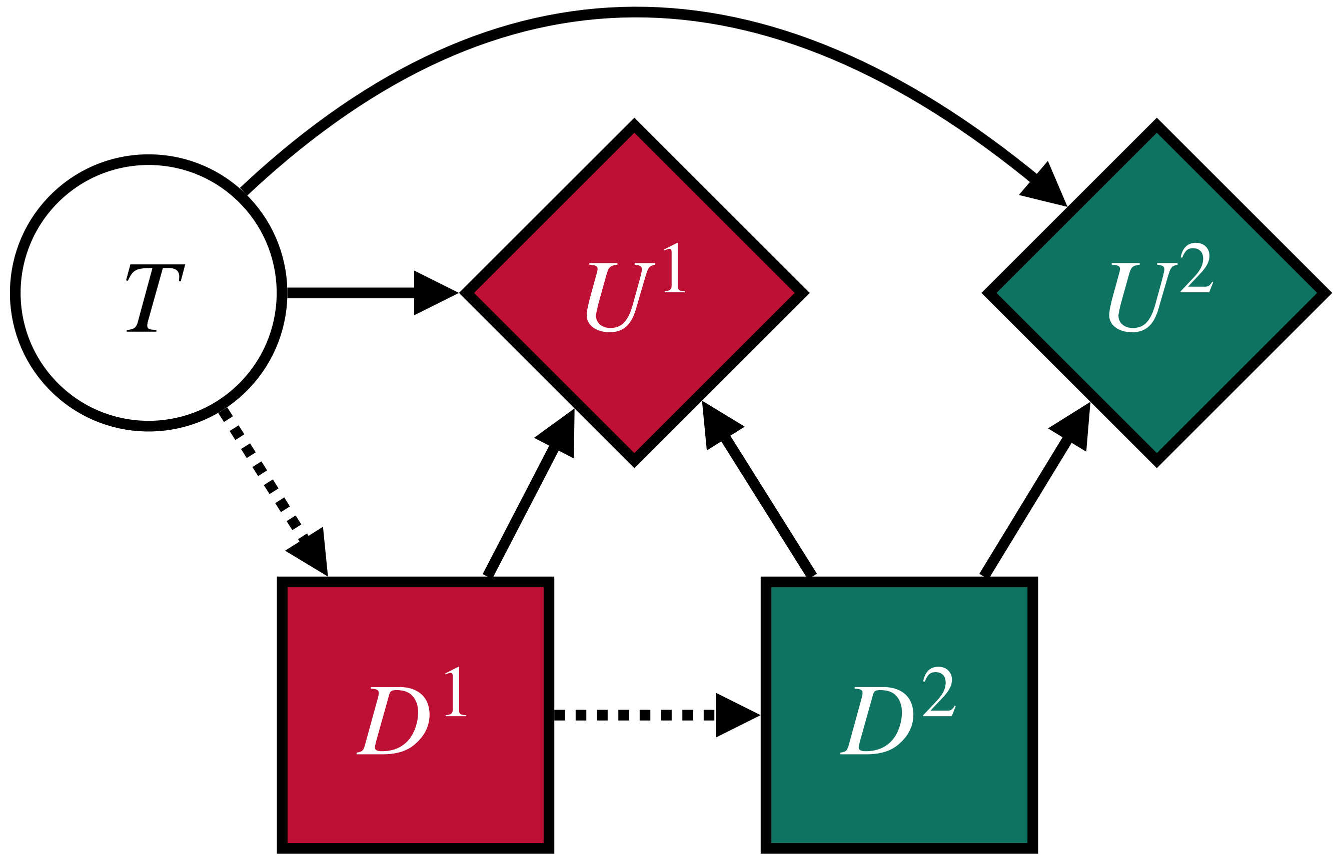
- Assuming basic knowledge of Pearl's hierarchy (BNs, CBNs, SCMs) [11]
- A MAID [8] $\mathscr{G} = (N, \mathbf{V}, \mathbb{E})$ consists of:
  - $N = \{1,\ldots,n\}$
  - $\mathbf{V} = \mathbf{X} \cup \bigcup_{i\in N} \mathbf{D}^i \cup \bigcup_{i\in N} \mathbf{U}^i$
  - $\mathbb{E} \subset \mathbf{V} \times \mathbf{V}$
- A MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ consists of:
  - A MAID $\mathscr{G}$
  - $\Pr(\mathbf{x}, \mathbf{u} : \mathbf{d}) := \prod_{V\in\mathbf{X}\cup\mathbf{U}} \Pr(v \mid \mathbf{pa}_V; \theta_V)$

- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

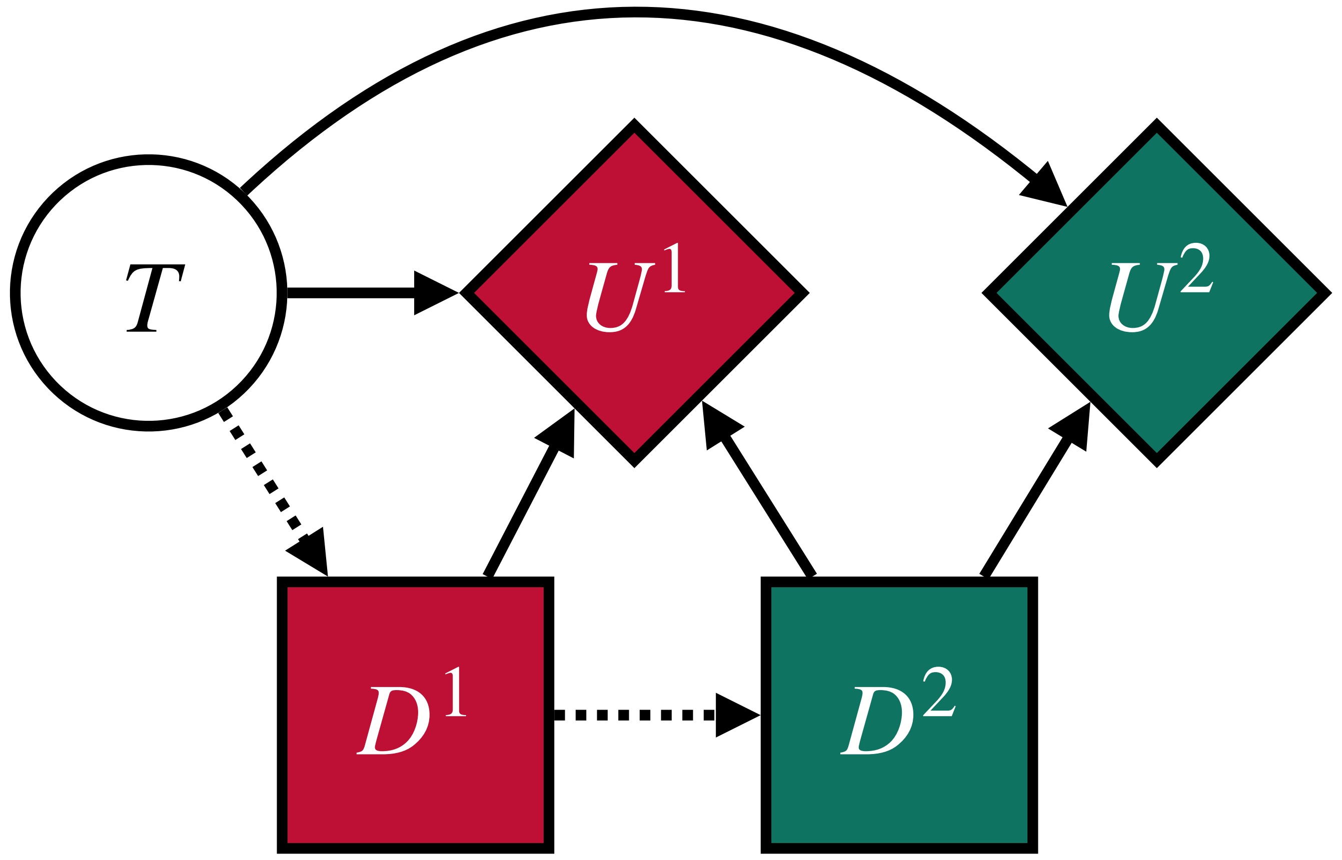
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Each agent i plays by selecting a policy π^i, made up of decision rules π_D

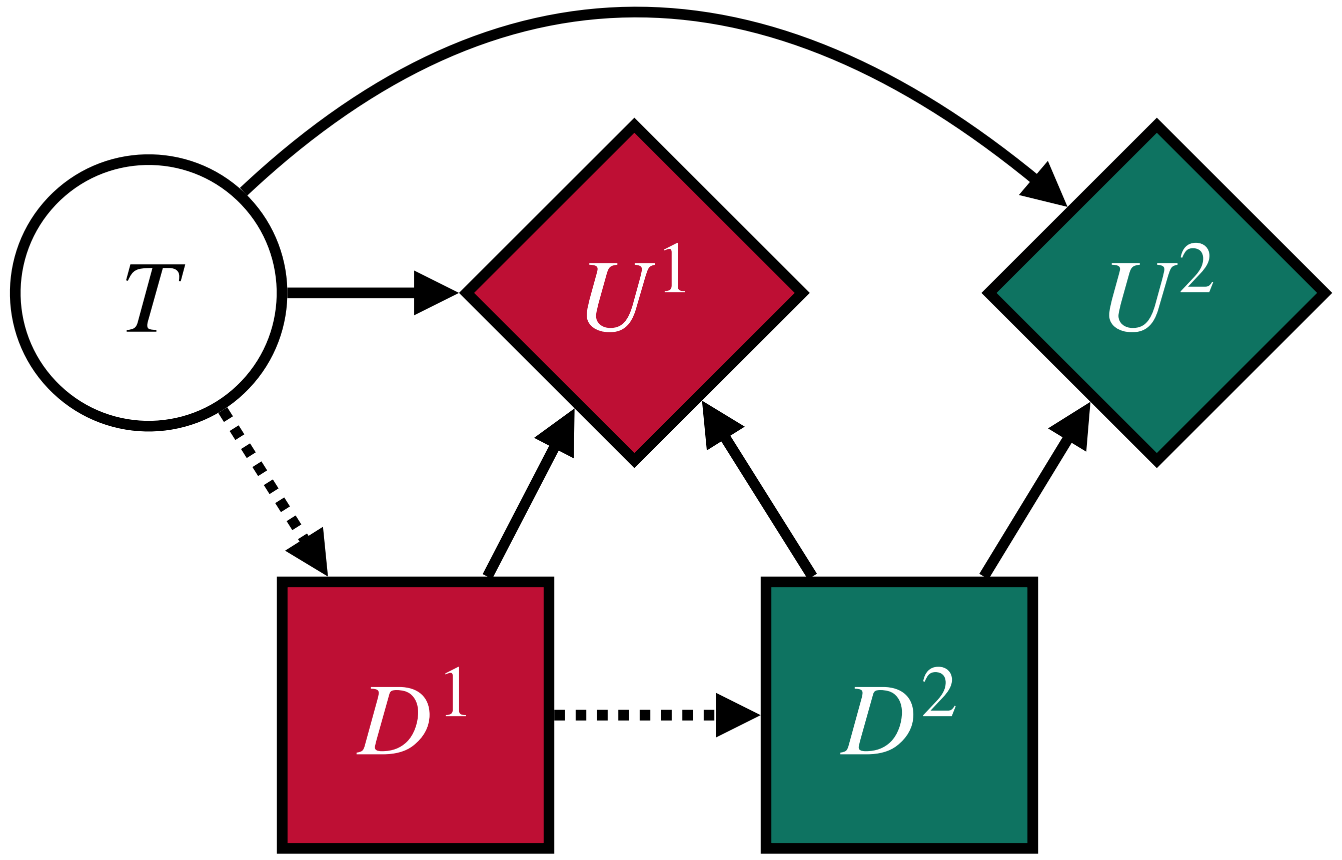
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Each agent i plays by selecting a policy π^i, made up of decision rules π_D
  - $\pi^i(\mathbf{d}^i \mid \mathbf{pa}_{\mathbf{D}^i}) := \prod_{D \in \mathbf{D}^i} \pi_D(d \mid \mathbf{pa}_D)$

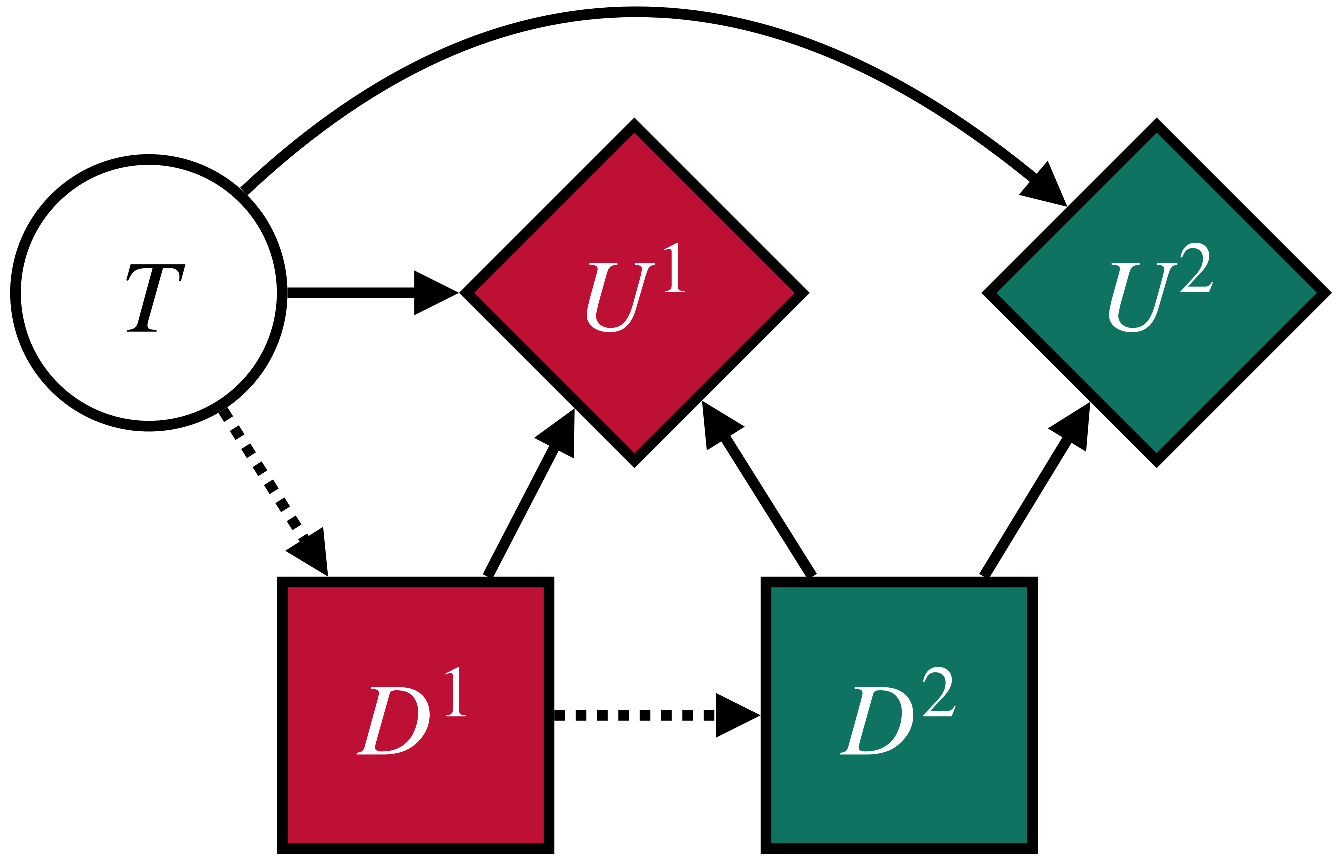
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Each agent i plays by selecting a policy π^i, made up of decision rules π_D
  - $\pi^i(\mathbf{d}^i \mid \mathbf{pa}_{\mathbf{D}^i}) := \prod_{D \in \mathbf{D}^i} \pi_D(d \mid \mathbf{pa}_D)$
  - This gives rise to a joint policy $\pi = (\pi^1, \ldots, \pi^n)$ and thus $\pi(\mathbf{D} : \mathbf{X}, \mathbf{U})$

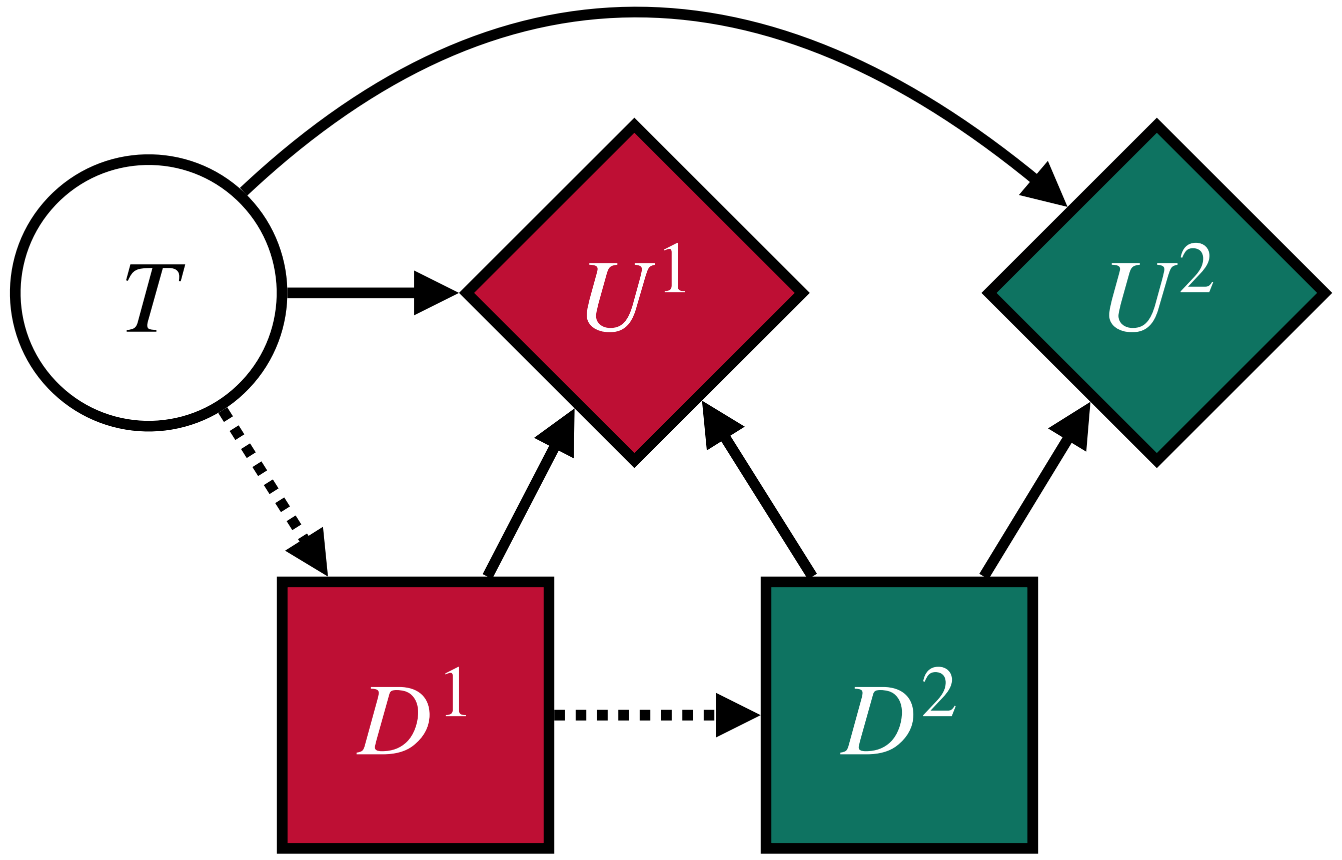
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Each agent i plays by selecting a policy π^i, made up of decision rules π_D
  - $\pi^i(\mathbf{d}^i \mid \mathbf{pa}_{\mathbf{D}^i}) := \prod_{D \in \mathbf{D}^i} \pi_D(d \mid \mathbf{pa}_D)$
  - This gives rise to a joint policy $\pi = (\pi^1, \ldots, \pi^n)$ and thus $\pi(\mathbf{D} : \mathbf{X}, \mathrm{U})$
  - Given π we have a joint distribution $\Pr^\pi(\mathbf{x}, \mathbf{d}, \mathbf{u}) := \Pr(\mathbf{x}, \mathbf{u} : \mathbf{d})\pi(\mathbf{d} : \mathbf{x}, \mathbf{u})$

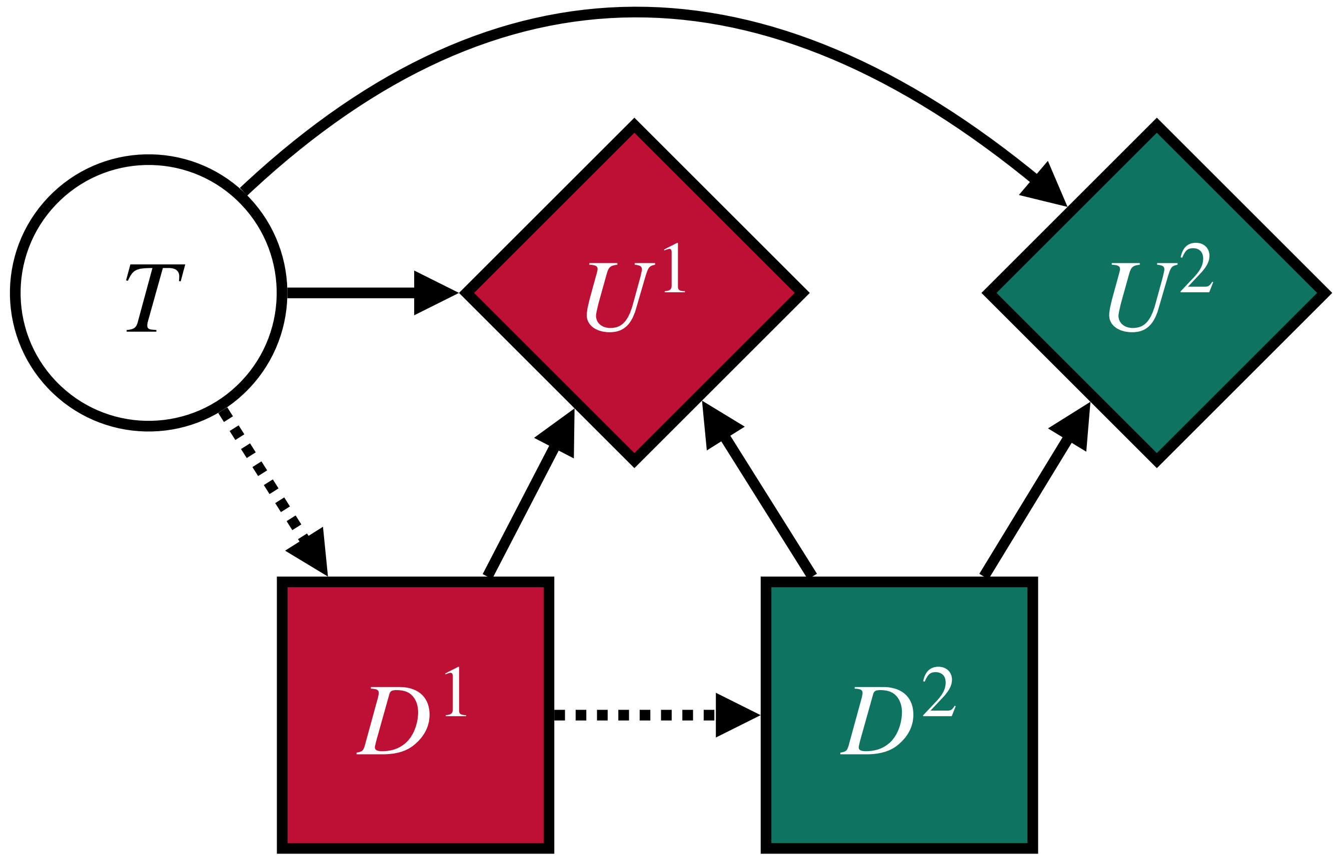
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Background

- Each agent i plays by selecting a policy π^i, made up of decision rules π_D
  - $\pi^i(\mathbf{d}^i \mid \mathbf{pa}_{\mathbf{D}^i}) := \prod_{D \in \mathbf{D}^i} \pi_D(d \mid \mathbf{pa}_D)$
  - This gives rise to a joint policy $\pi = (\pi^1, \ldots, \pi^n)$ and thus $\pi(\mathbf{D} : \mathbf{X}, \mathbf{U})$
  - Given π we have a joint distribution $\Pr^\pi(\mathbf{x}, \mathbf{d}, \mathbf{u}) := \Pr(\mathbf{x}, \mathbf{u} : \mathbf{d})\pi(\mathbf{d} : \mathbf{x}, \mathbf{u})$
- The expected utility for agent i under π is given by $\mathbb{E}_\pi\left[\sum_{U \in \mathbf{U}^i} u\right]$

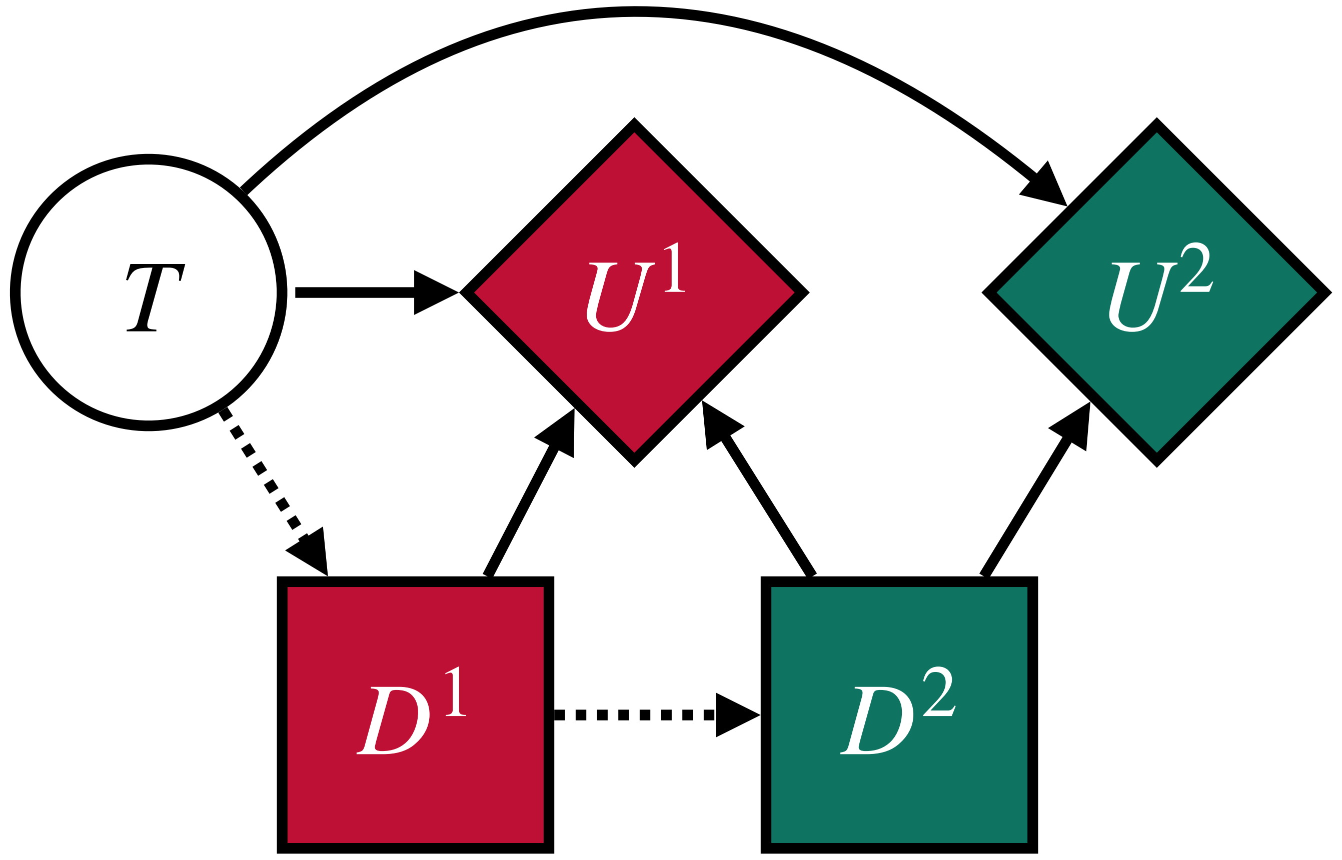
- Example: Job market signalling [16]
- The worker is either hard-working or lazy (T), and chooses to go to university or not (D^1). The firm chooses to hire the worker or not (D^2)



Representing Strategic Interactions

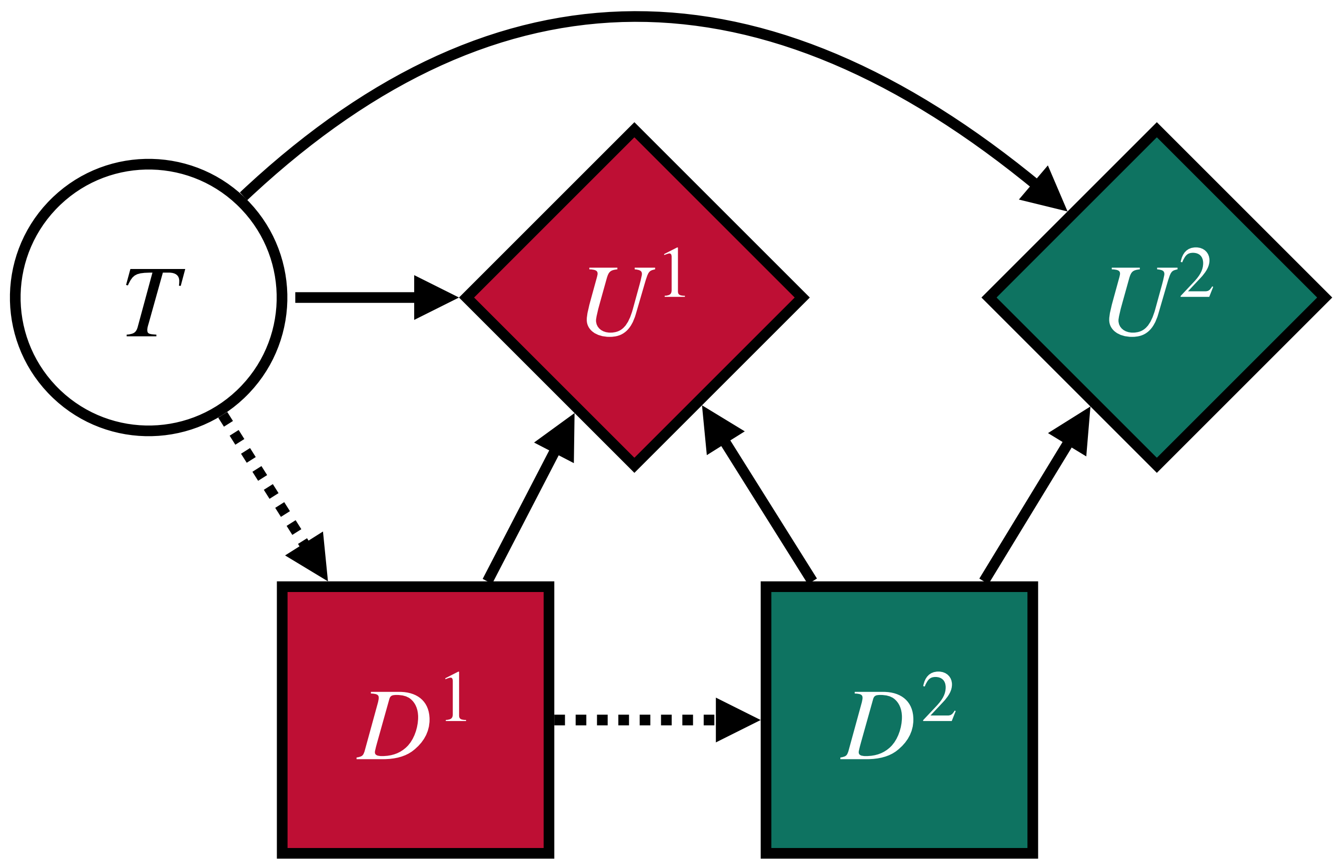
Extended Models

Extended Models



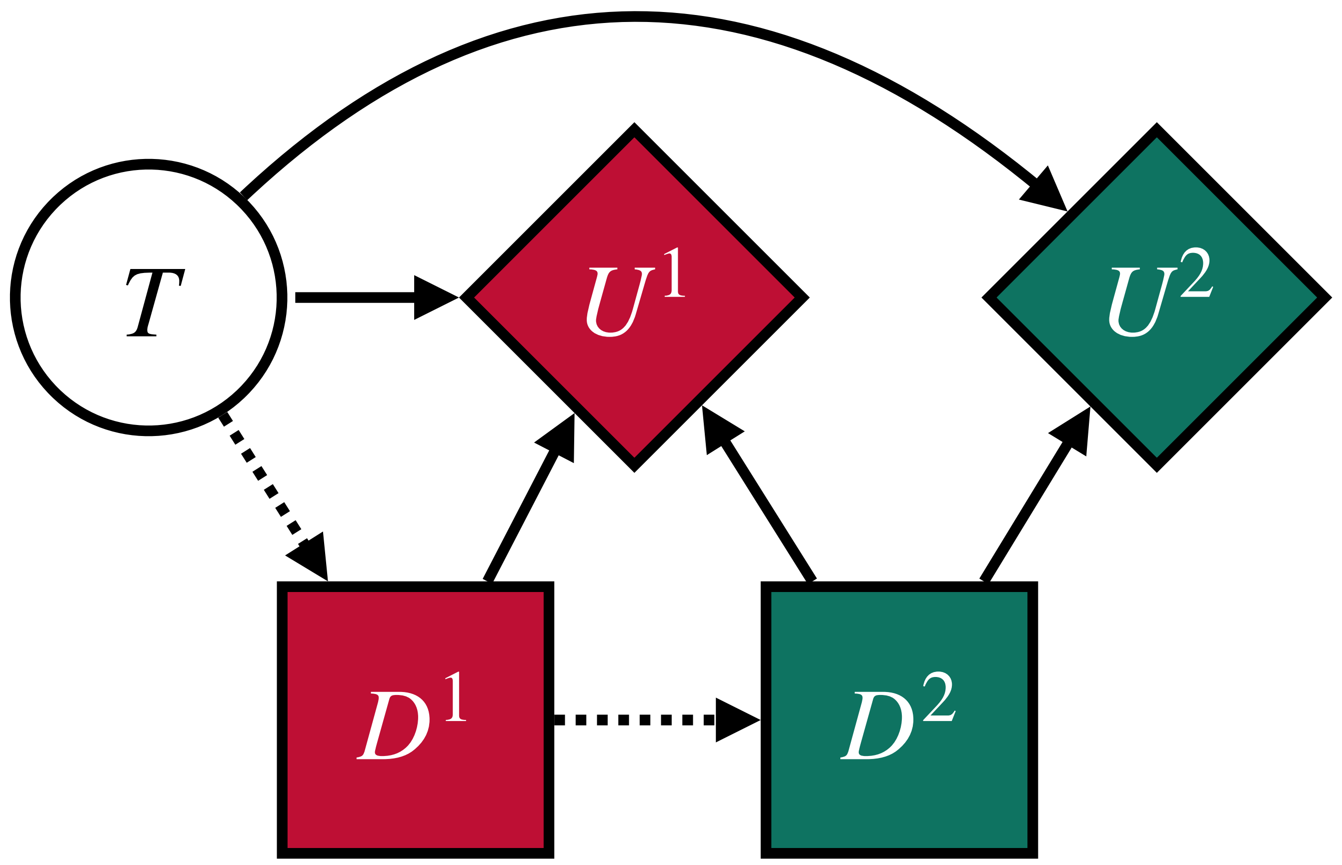
Extended Models

- This graph doesn't tell the whole story



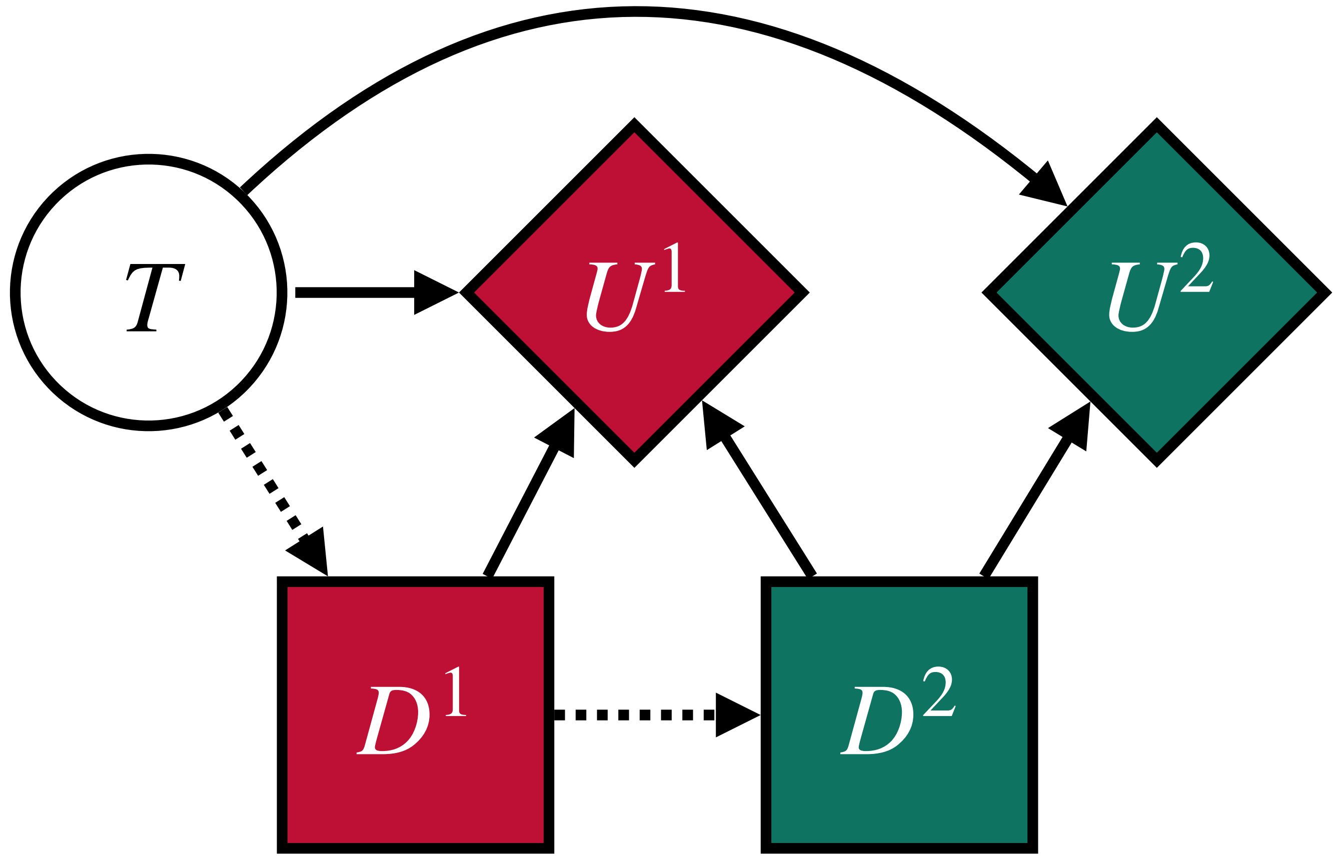
Extended Models

- This graph doesn't tell the whole story
- In any non-trivial equilibrium of the game the choice of each decision rule π_D will depend on:



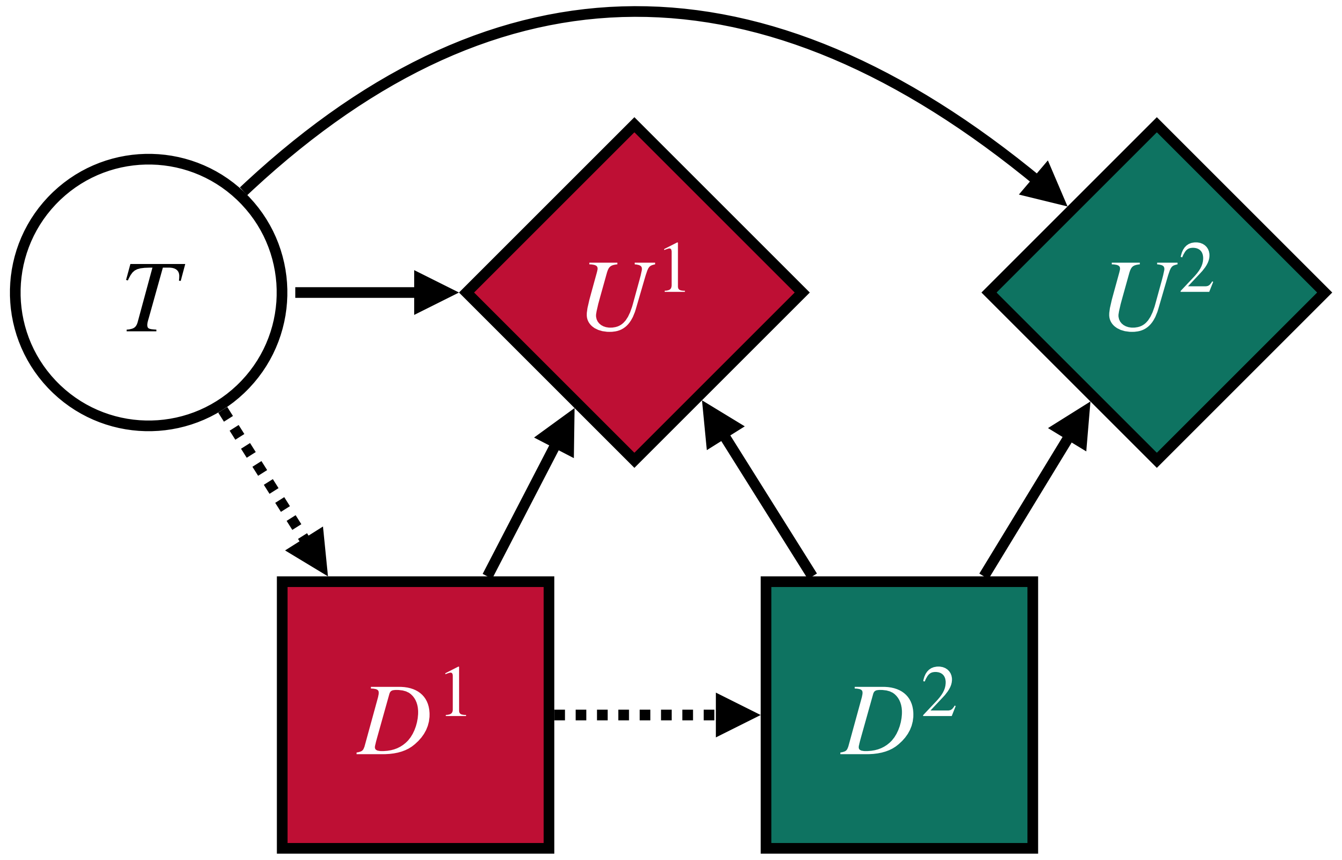
Extended Models

- This graph doesn't tell the whole story
- In any non-trivial equilibrium of the game the choice of each decision rule π_D will depend on:
  - The other decision rules π_{-D}



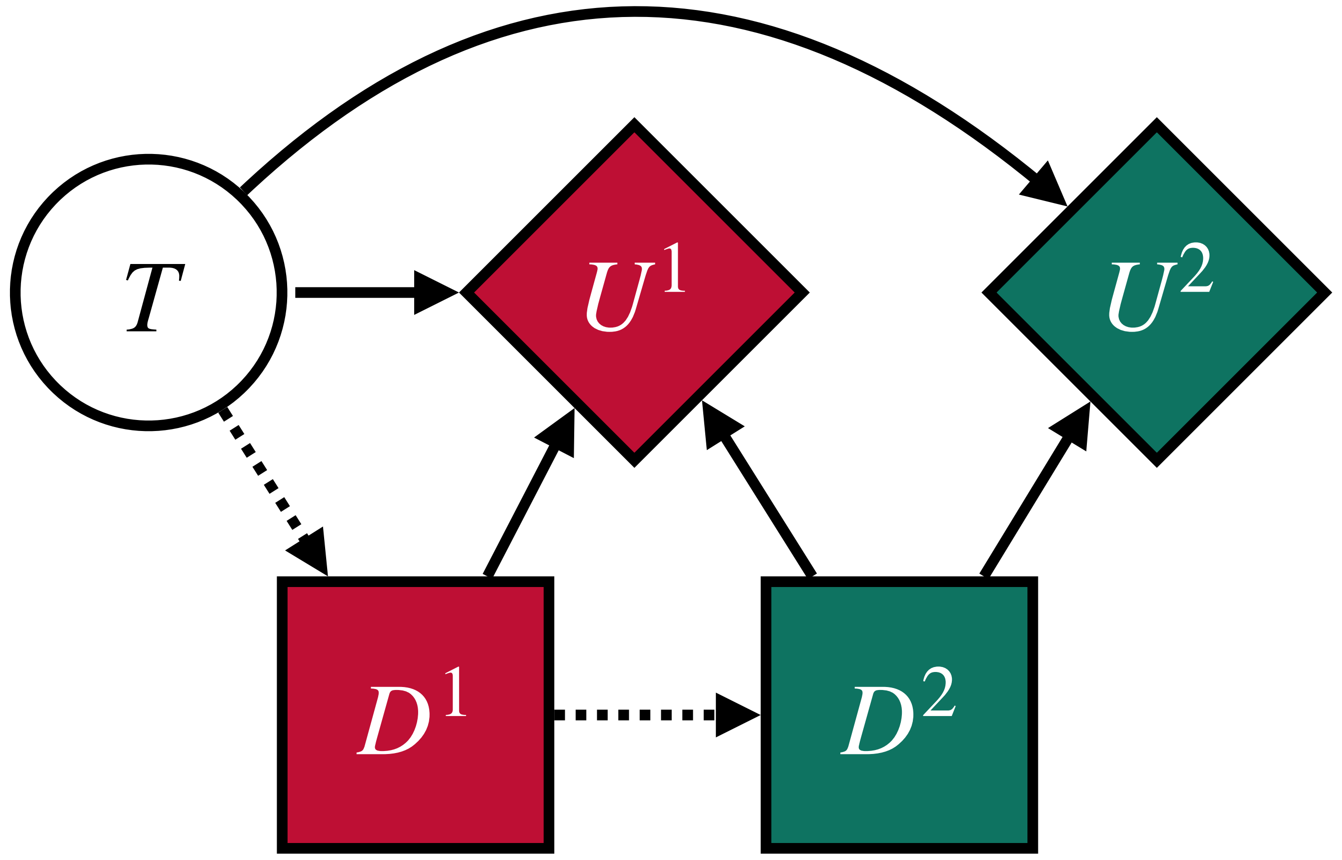
Extended Models

- This graph doesn't tell the whole story
- In any non-trivial equilibrium of the game the choice of each decision rule π_D will depend on:
  - The other decision rules π_{-D}
  - The parameterisation of the game θ



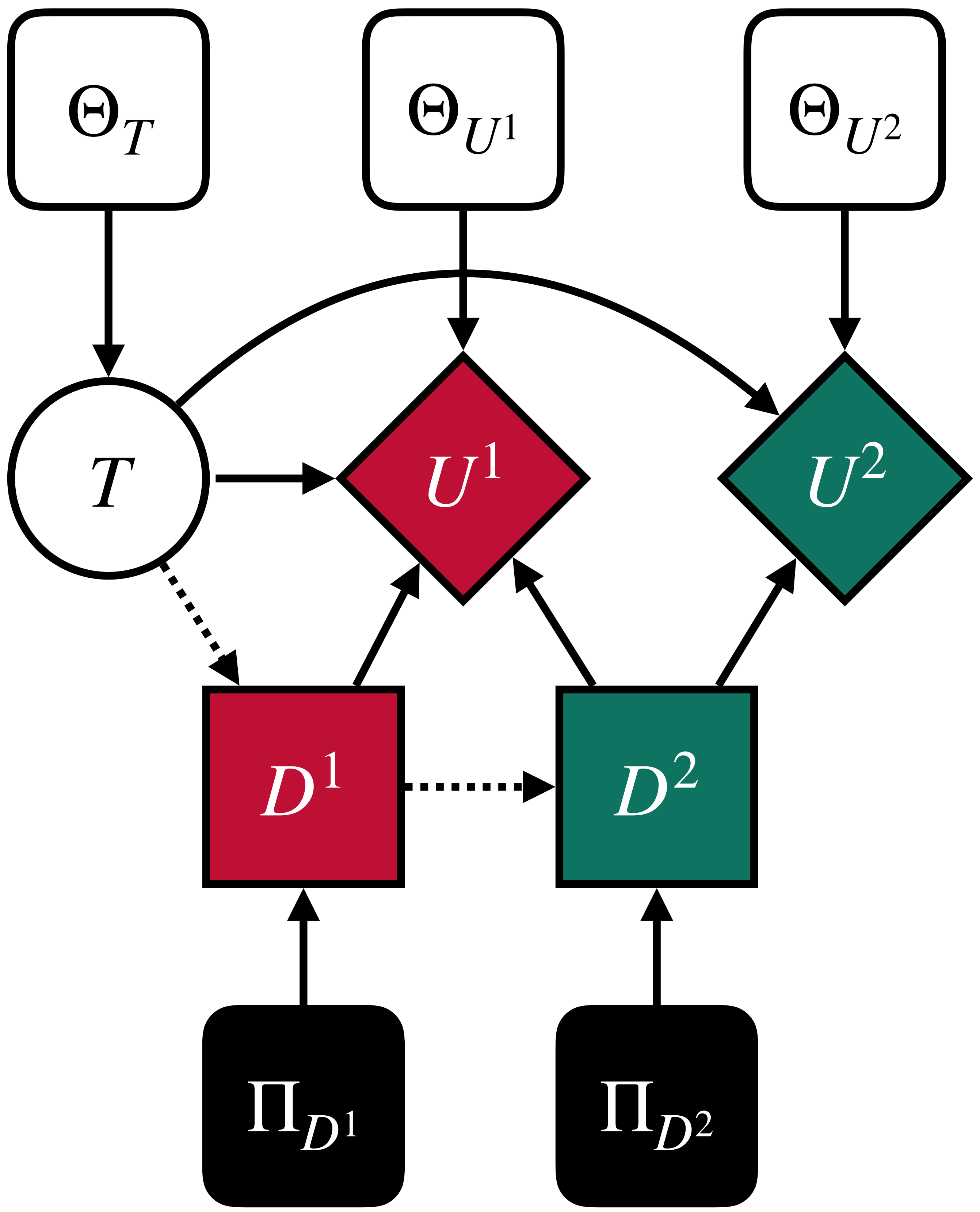
Extended Models

- This graph doesn't tell the whole story
- In any non-trivial equilibrium of the game the choice of each decision rule π_D will depend on:
  - The other decision rules π_{-D}
  - The parameterisation of the game θ
- We represent these dependencies using <u>mechanism variables</u> $\mathsf{M}_{\mathbf{V}} = \{\mathsf{M}_V\}_{V \in \mathbf{V}}$, denoting M_V as Π_V if $V \in \mathbf{D}$ and as Θ_V otherwise

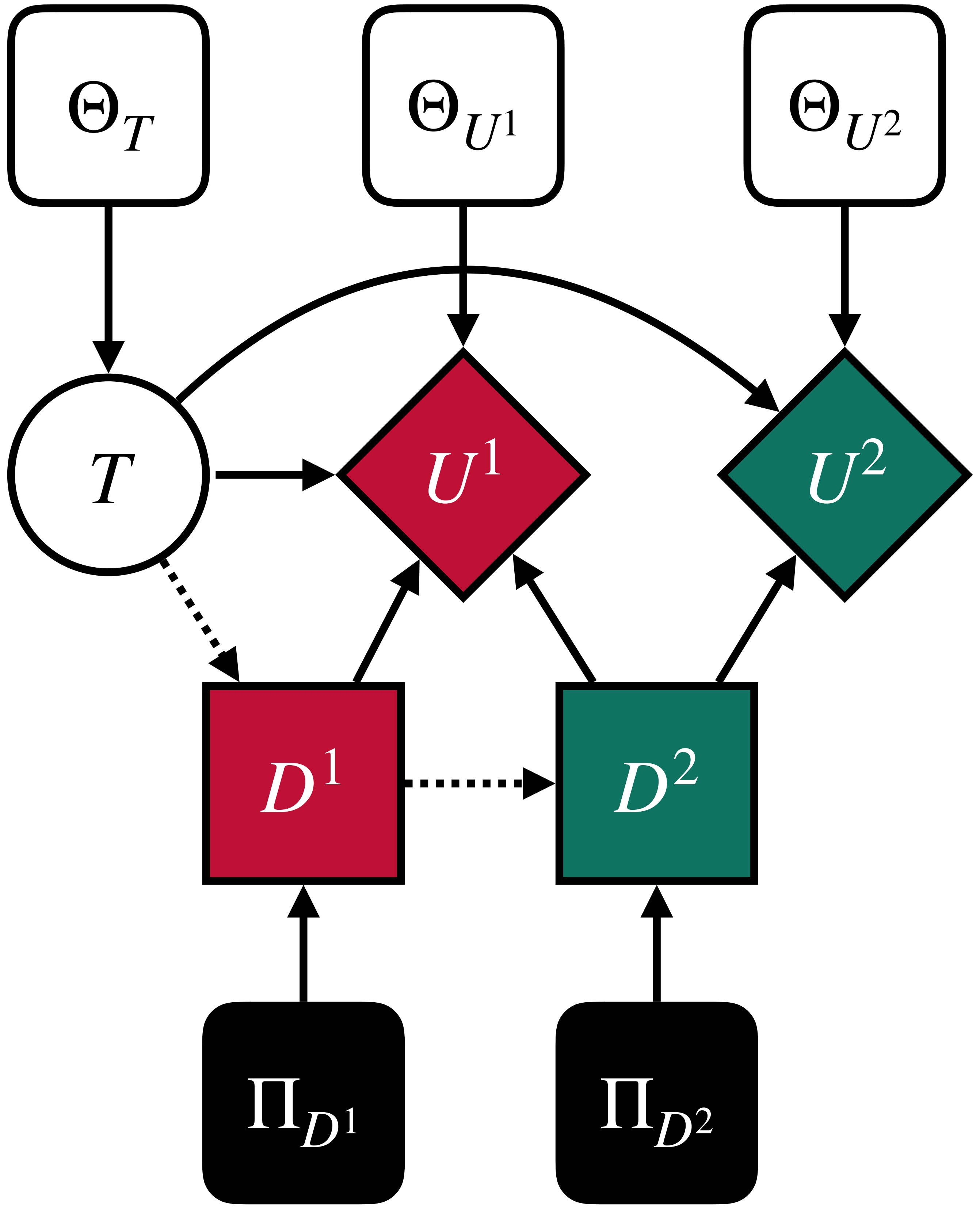


Extended Models

- This graph doesn't tell the whole story
- In any non-trivial equilibrium of the game the choice of each decision rule π_D will depend on:
  - The other decision rules π_{-D}
  - The parameterisation of the game θ
- We represent these dependencies using <u>mechanism variables</u> $\mathsf{M}_{\mathbf{V}} = \{\mathsf{M}_V\}_{V \in \mathbf{V}}$, denoting M_V as Π_V if $V \in \mathbf{D}$ and as Θ_V otherwise

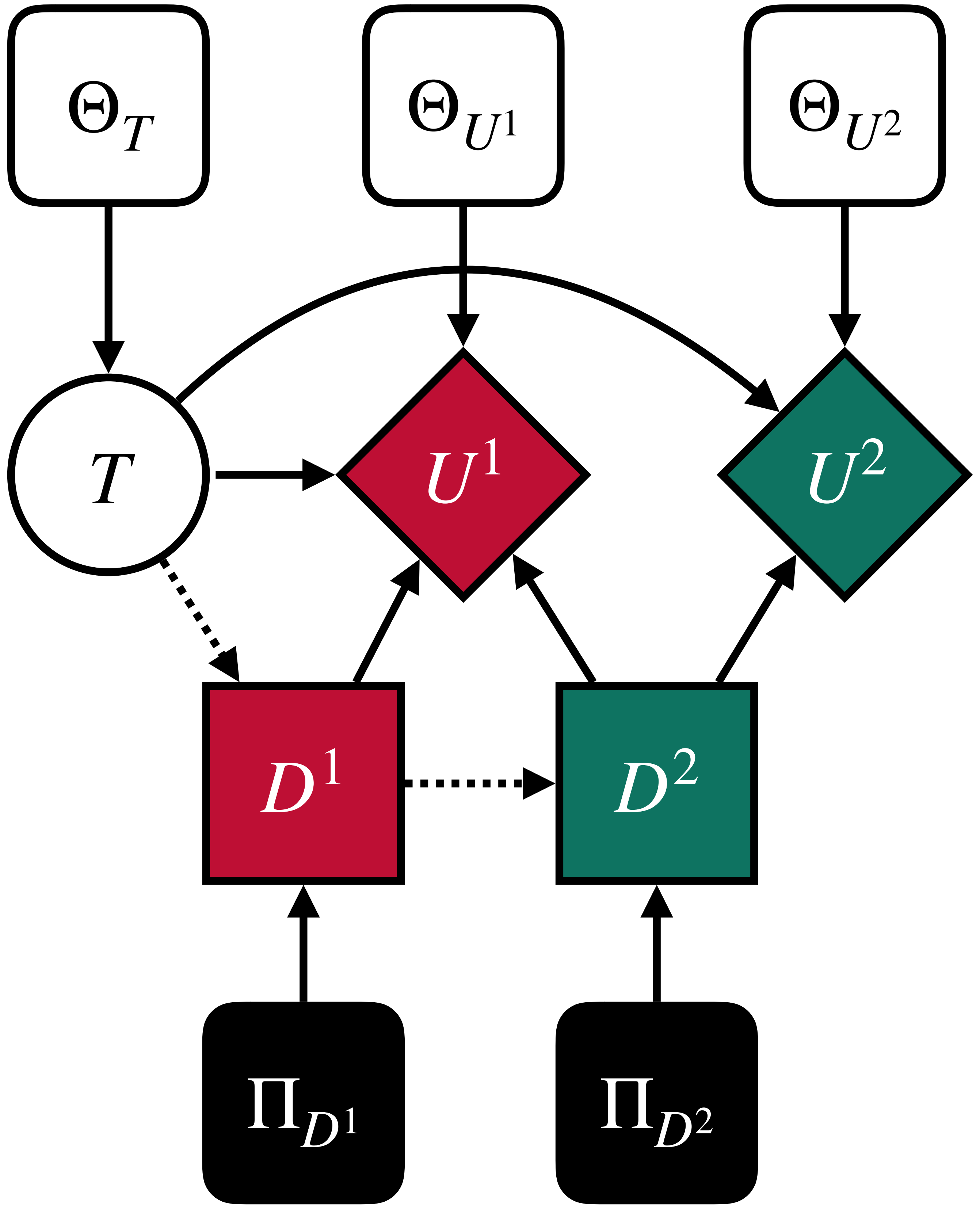


Extended Models



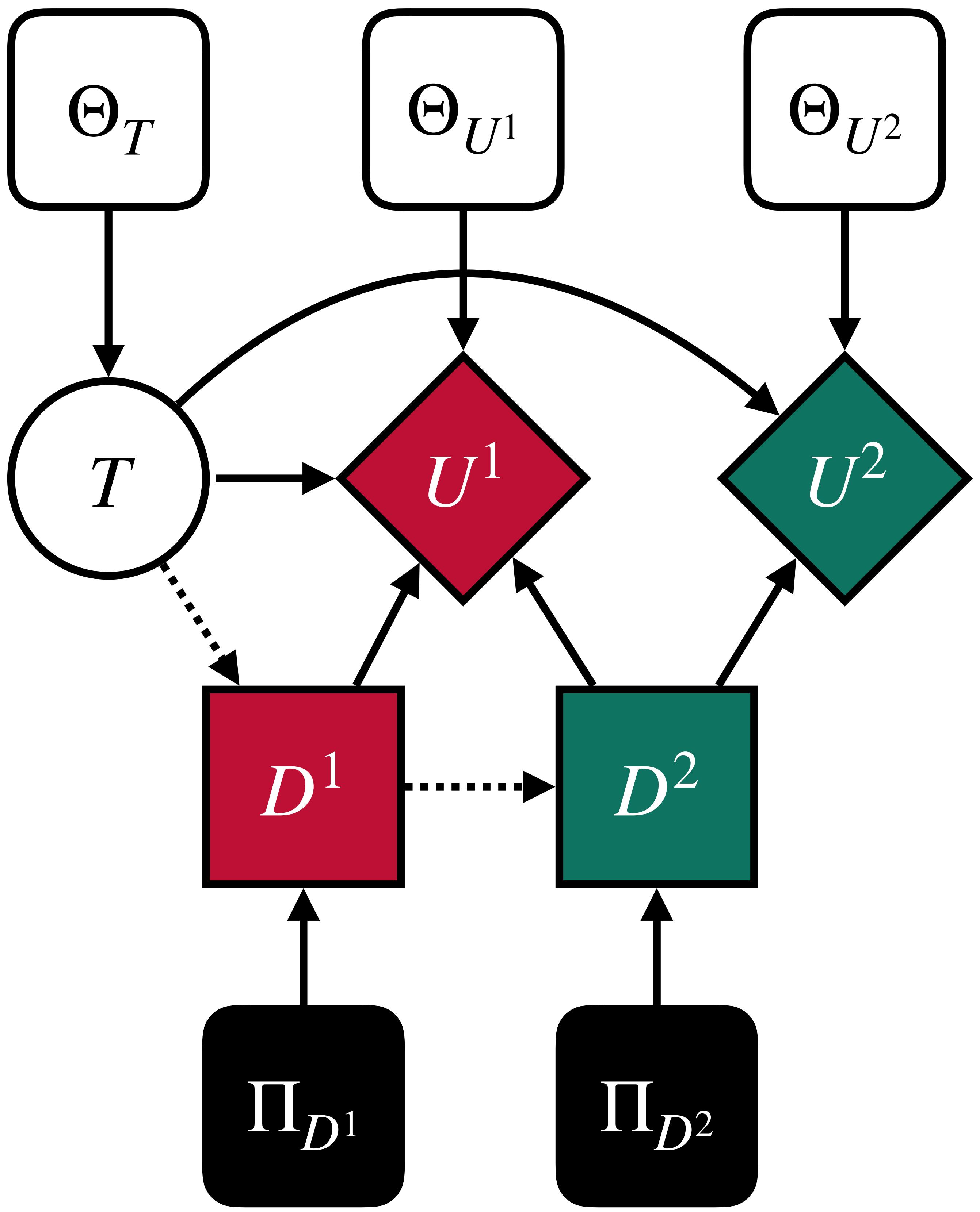
Extended Models

- $\Pr^{\pi}(\mathbf{v}; \theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V \in \mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$



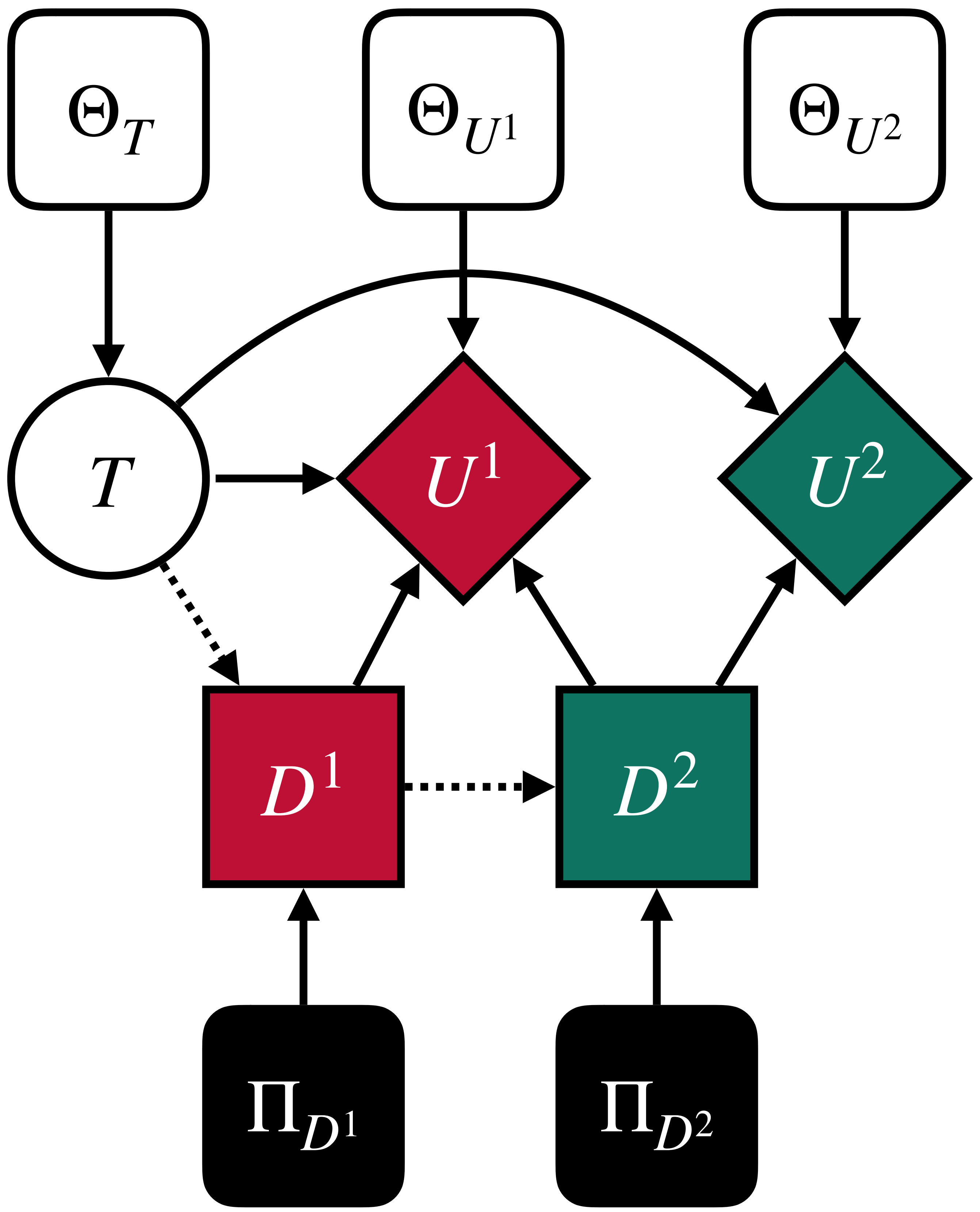
Extended Models

- $\Pr^{\pi}(\mathbf{v};\theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V \in \mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$
- Each Θ_V is governed by a point distribution $\delta(\Theta_V = \theta_V)$



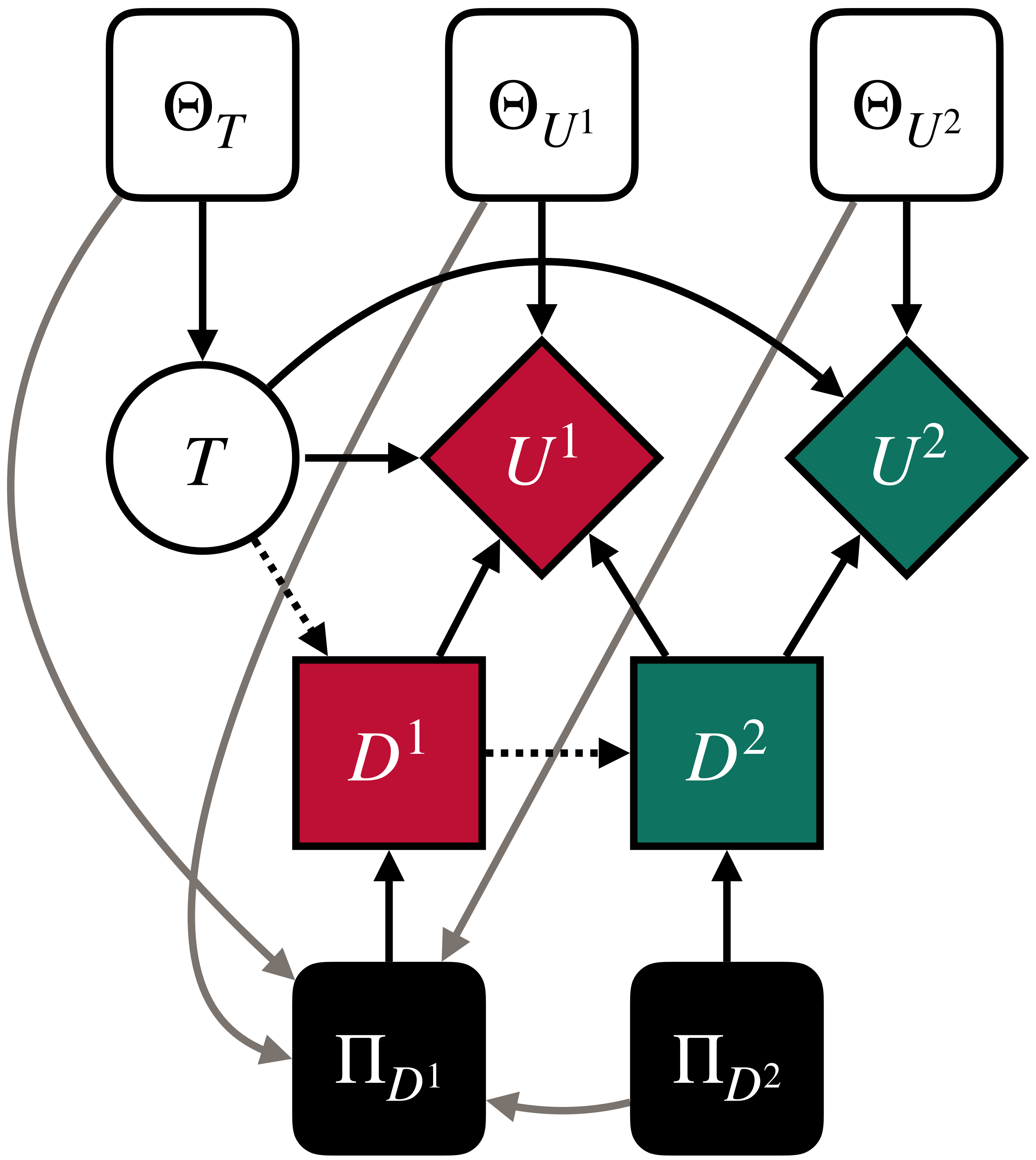
Extended Models

- $\Pr^{\pi}(\mathbf{v};\theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V\in\mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$
- Each Θ_V is governed by a point distribution $\delta(\Theta_V = \theta_V)$
- Each Π_D is governed by a <u>rationality relation</u> $r_D \subseteq \mathrm{dom}(\mathbf{Pa}_{\Pi_D}) \times \mathrm{dom}(\Pi_D)$ that is serial (i.e., a many-valued function)



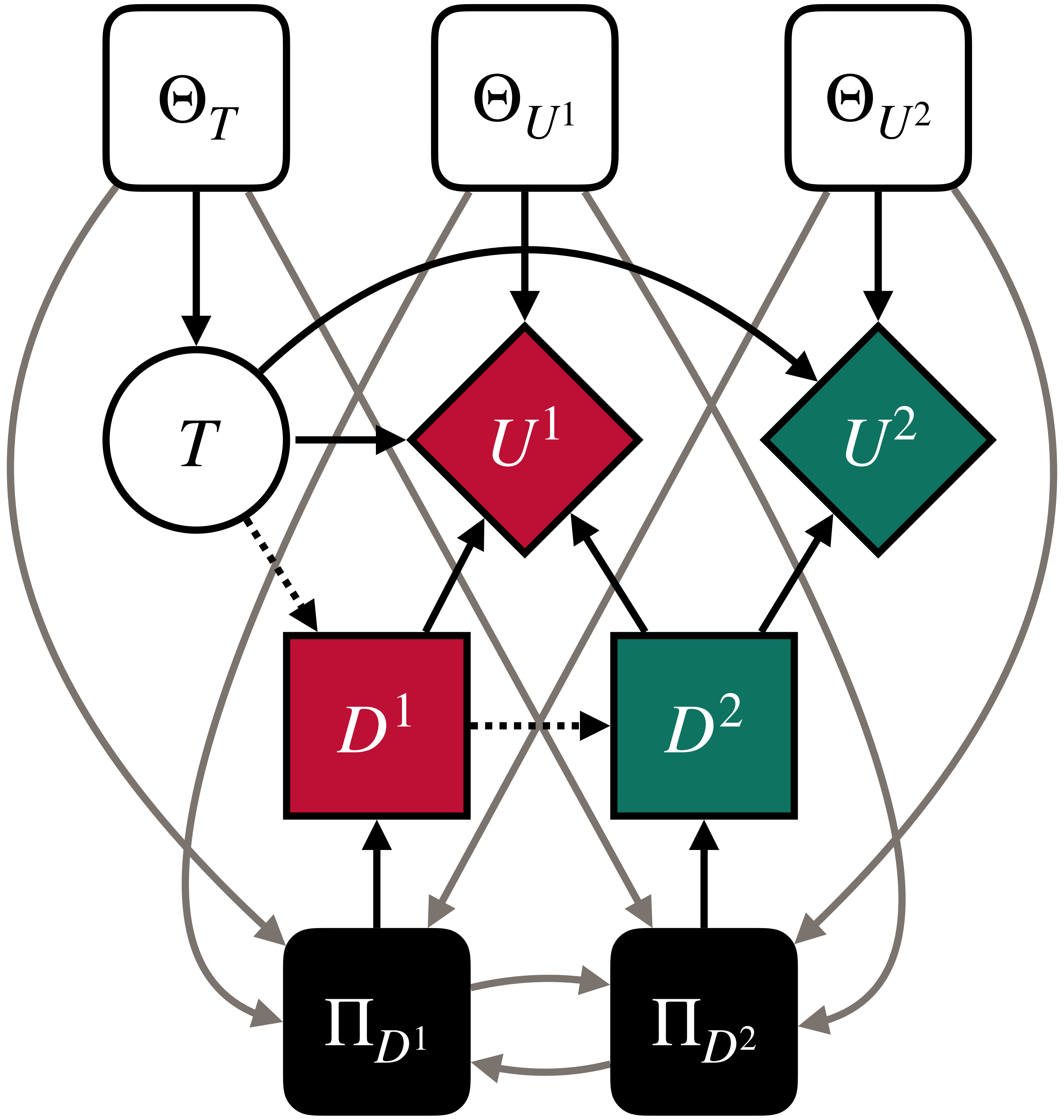
Extended Models

- $\Pr^{\pi}(\mathbf{v};\theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V\in\mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$
- Each Θ_V is governed by a point distribution $\delta(\Theta_V = \theta_V)$
- Each Π_D is governed by a <u>rationality relation</u> $r_D \subseteq \mathrm{dom}(\mathbf{Pa}_{\Pi_D}) \times \mathrm{dom}(\Pi_D)$ that is serial (i.e., a many-valued function)



Extended Models

- $\Pr^{\pi}(\mathbf{v};\theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V\in\mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$
- Each Θ_V is governed by a point distribution $\delta(\Theta_V = \theta_V)$
- Each Π_D is governed by a rationality relation $r_D \subseteq \mathrm{dom}(\mathbf{Pa}_{\Pi_D}) \times \mathrm{dom}(\Pi_D)$ that is serial (i.e., a many-valued function)

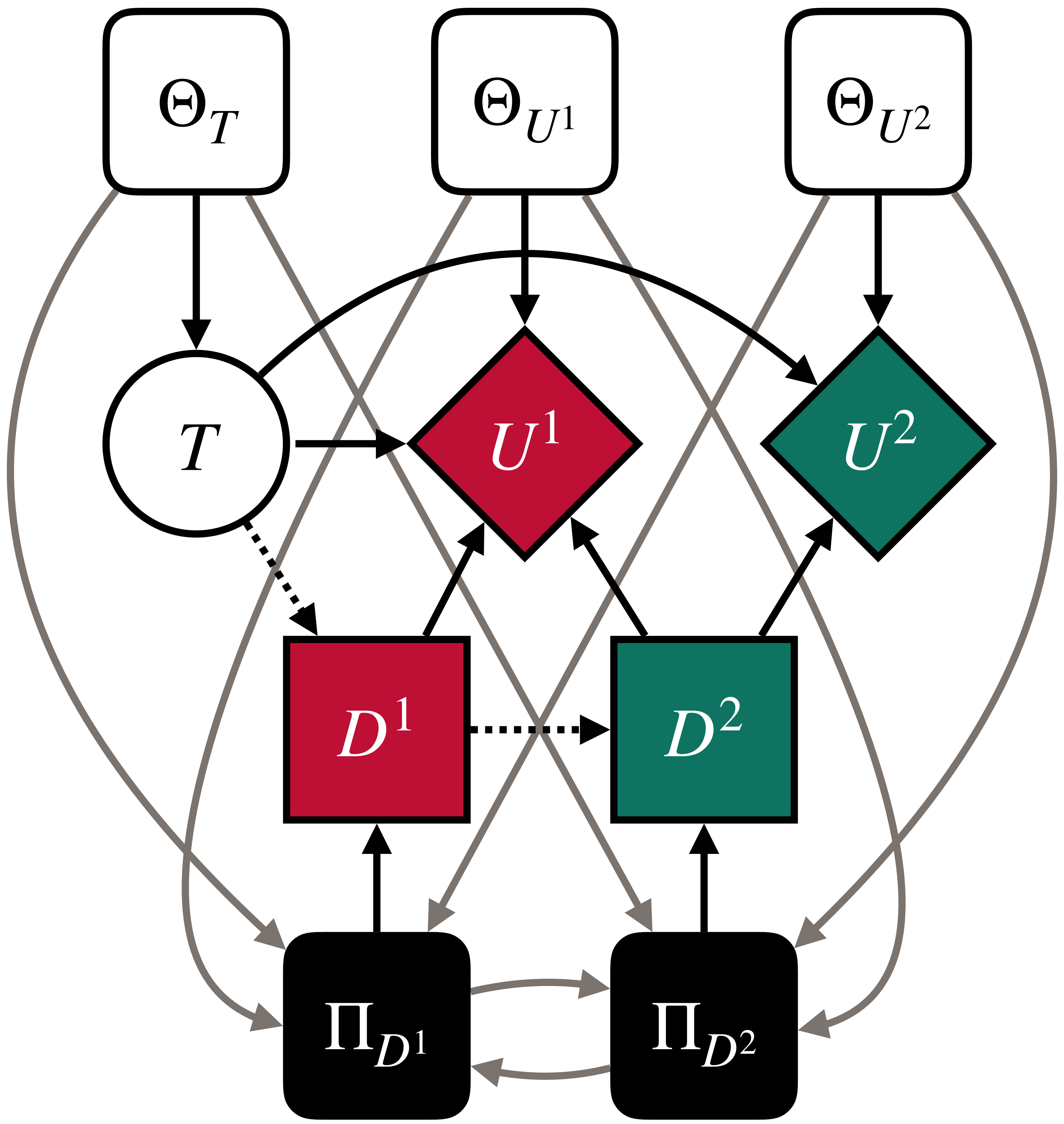


Extended Models

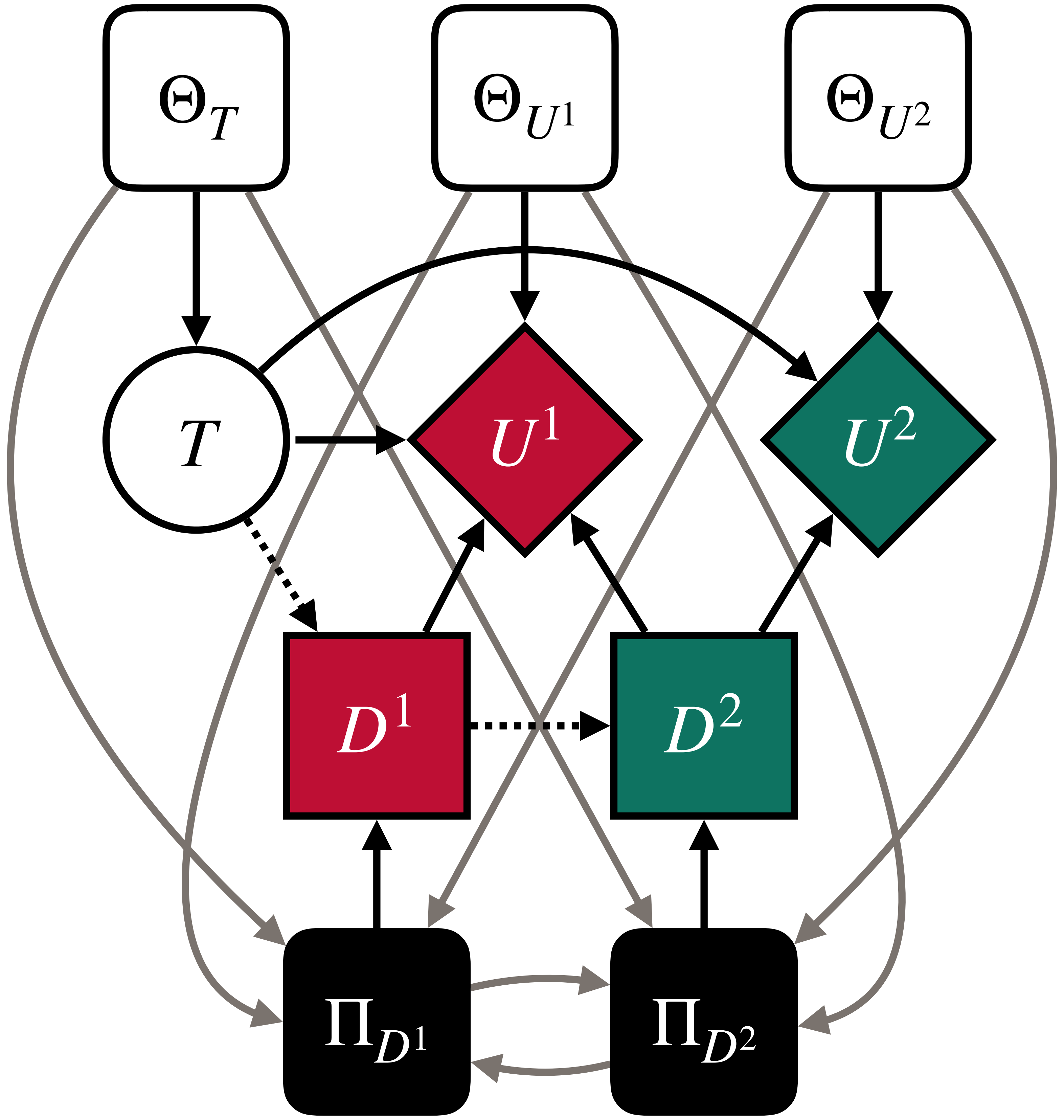
- $\Pr^{\pi}(\mathbf{v};\theta) = \Pr(\mathbf{v} \mid \mathsf{m}) := \prod_{V \in \mathbf{V}} \Pr(v \mid \mathbf{pa}_V, \mathsf{m}_V)$
- Each Θ_V is governed by a point distribution $\delta(\Theta_V = \theta_V)$
- Each Π_D is governed by a rationality relation $r_D \subseteq \text{dom}(\mathbf{Pa}_{\Pi_D}) \times \text{dom}(\Pi_D)$ that is serial (i.e., a many-valued function)

$$(\mathbf{pa}_{\Pi_D}, \pi_D) \in r_D^{NE} \Leftrightarrow \pi_D \in r_D^{NE}(\mathbf{pa}_{\Pi_D})$$

$$\Leftrightarrow \pi^i \in \text{argmax}_{\hat{\pi}^i \in \text{dom}(\Pi^i)} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}\Big[\sum_{U \in \mathbf{U}^i} u\Big]$$

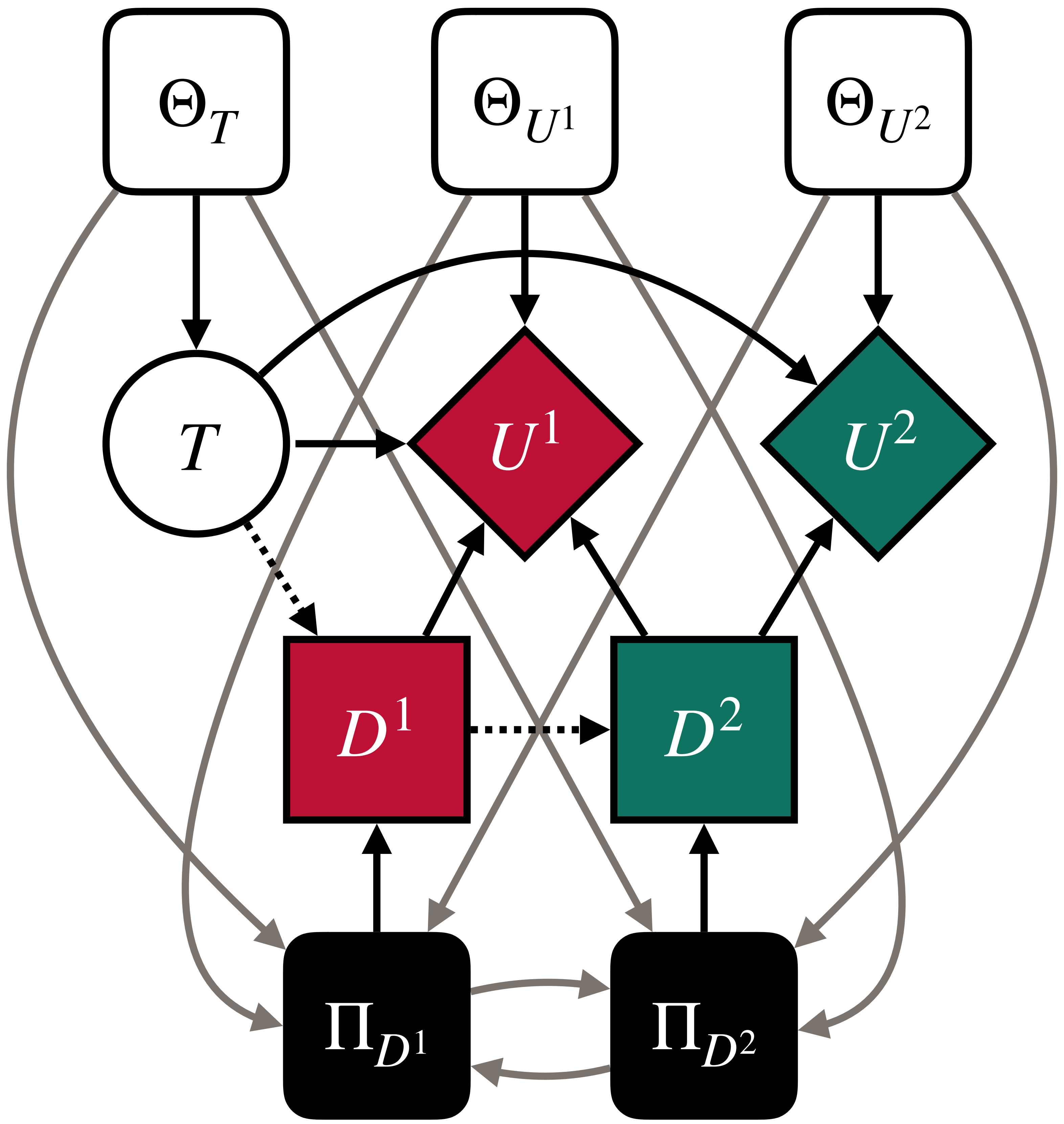


Extended Models



Extended Models

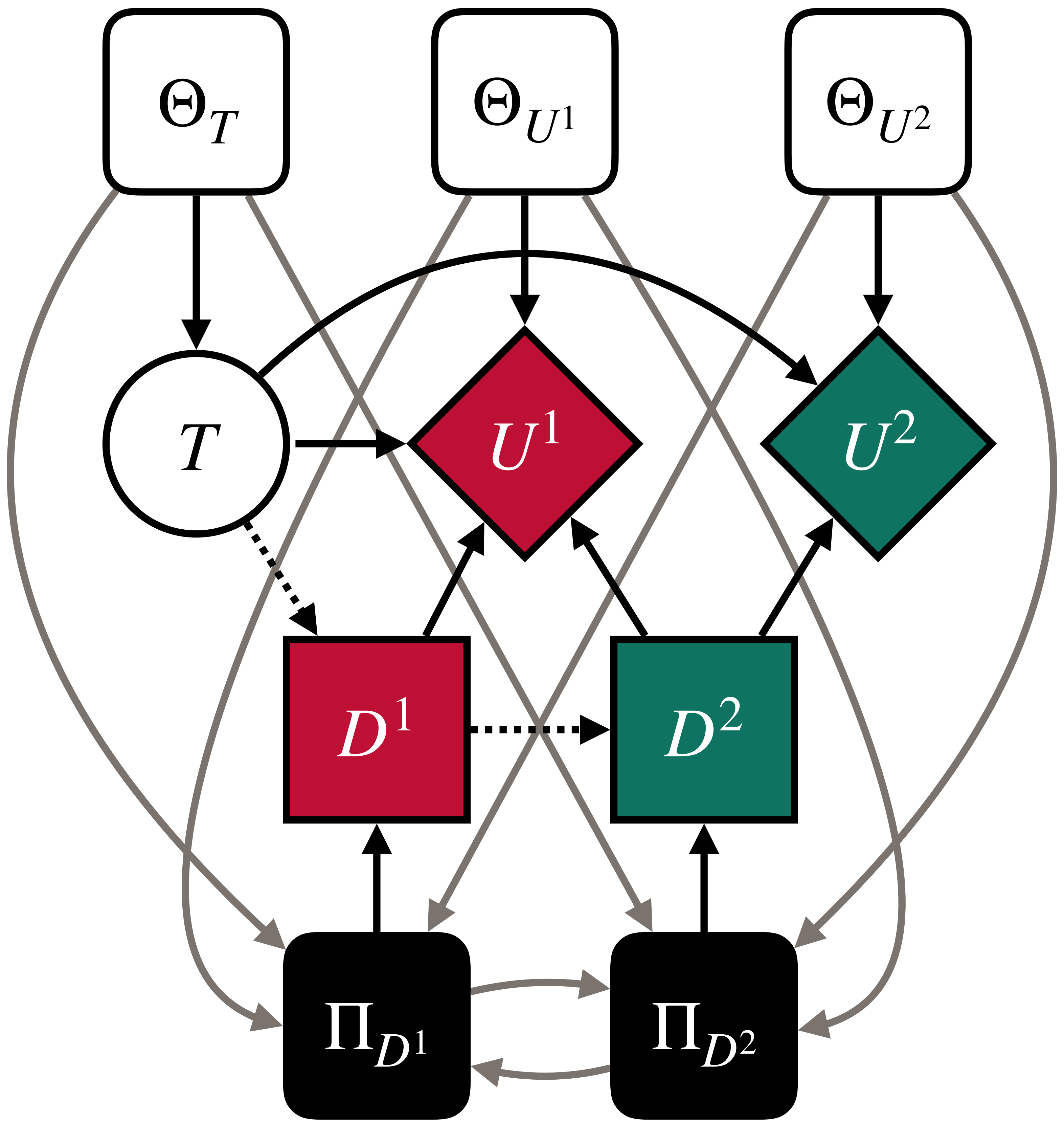
- Given a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ over a MAID $\mathscr{G} = (N, \mathbf{V}, \mathbb{E})$ and a set of rationality relations $\mathscr{R} = \{r_D\}_{D \in \mathbf{D}}$ we call the result of this construction an <u>extended MAIM</u> x$\mathscr{M} = ($x$\mathscr{G}, \theta, \mathscr{R})$ over an <u>extended MAID</u> x$\mathscr{G} = (N, \mathbf{V} \cup$ M, x$\mathbb{E})$



Extended Models

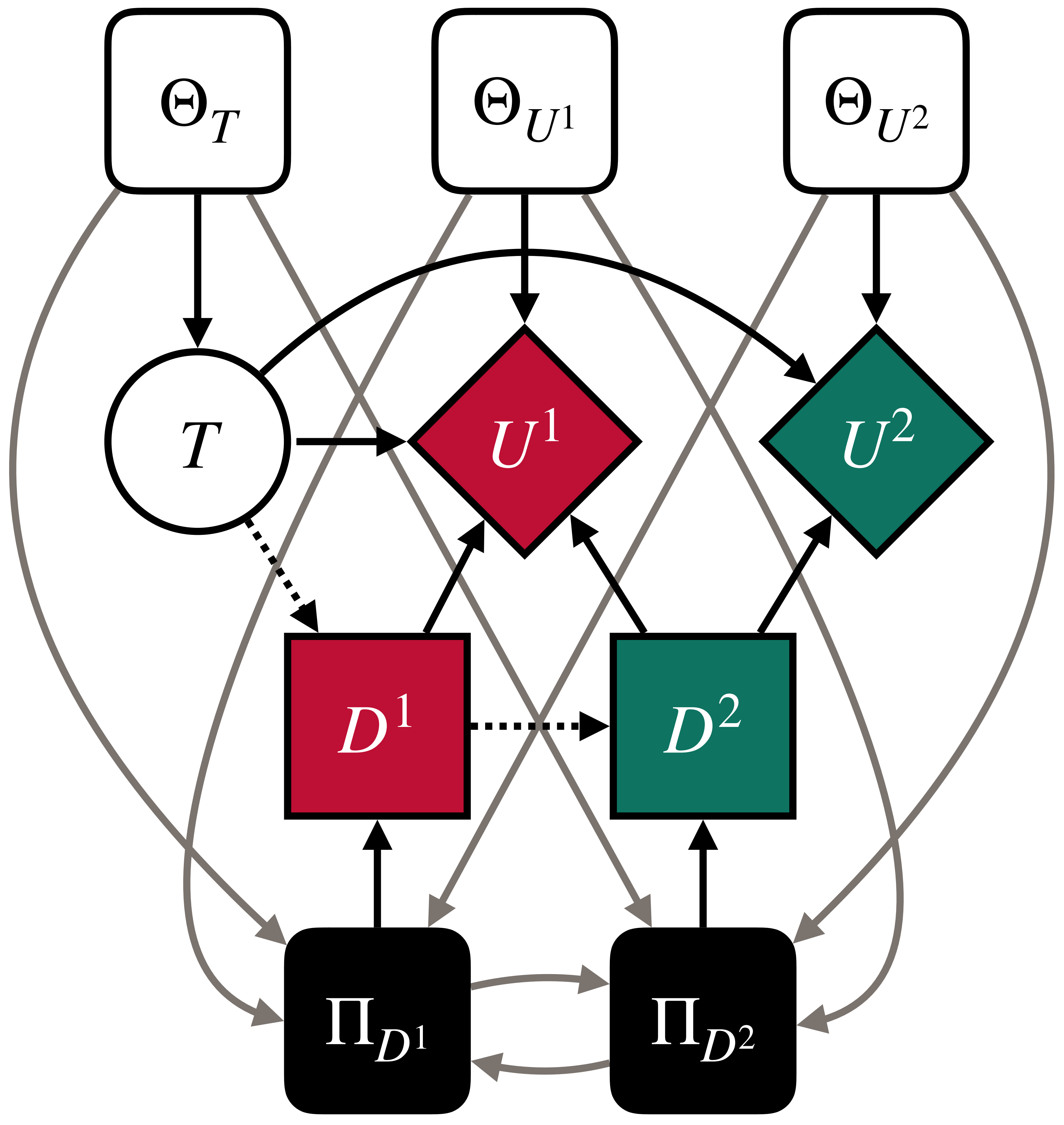
- Given a MAIM $\mathcal{M} = (\mathcal{G}, \theta)$ over a MAID $\mathcal{G} = (N, \mathbf{V}, \mathbb{E})$ and a set of rationality relations $\mathcal{R} = \{r_D\}_{D \in \mathbf{D}}$ we call the result of this construction an <u>extended MAIM</u> x$\mathcal{M}$ $= ($x$\mathcal{G}, \theta, \mathcal{R})$ over an <u>extended MAID</u> x$\mathcal{G}$ $= (N, \mathbf{V} \cup$ M, x$\mathbb{E})$

- We denote by $\mathcal{R}($x$\mathcal{M})$ the <u>rational outcomes</u> of the game, where $\pi \in \mathcal{R}($x$\mathcal{M})$ if $\pi_D \in r_D(\mathbf{pa}_{\Pi_D})$ for every $D \in \mathbf{D}$



Extended Models

- Given a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ over a MAID $\mathscr{G} = (N, \mathbf{V}, \mathbb{E})$ and a set of rationality relations $\mathscr{R} = \{r_D\}_{D \in \mathbf{D}}$ we call the result of this construction an <u>extended MAIM</u> $\mathrm{x}\mathscr{M} = (\mathrm{x}\mathscr{G}, \theta, \mathscr{R})$ over an <u>extended MAID</u> $\mathrm{x}\mathscr{G} = (N, \mathbf{V} \cup \mathsf{M}, \mathrm{x}\mathbb{E})$
- We denote by $\mathscr{R}(\mathrm{x}\mathscr{M})$ the <u>rational outcomes</u> of the game, where $\pi \in \mathscr{R}(\mathrm{x}\mathscr{M})$ if $\pi_D \in r_D(\mathbf{pa}_{\Pi_D})$ for every $D \in \mathbf{D}$
  - For example, $\mathscr{R}^{NE}(\mathrm{x}\mathscr{M})$ are the NEs of $\mathscr{M}$



Answering Queries

A Causal Hierarchy for Games

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

1. Predictions

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

1. Predictions

   a) Given that the worker went to university, what is their wellbeing?

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

1. Predictions

   a) Given that the worker went to university, what is their wellbeing?

a)

1. $\Pr^{\pi}(u^1 \mid g)$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

1. Predictions

   a) Given that the worker went to university, what is their wellbeing?

   b) Given that the worker always decides to go to university, what are the firm's profits?

a)

1. $\Pr^{\pi}(u^1 \mid g)$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

1. Predictions

   a) Given that the worker went to university, what is their wellbeing?

   b) Given that the worker always decides to go to university, what are the firm's profits?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

2. Interventions

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

2. Interventions

   a) Given that the worker is forced to go to university, what is their wellbeing?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

2. Interventions

   a) Given that the worker is forced to go to university, what is their wellbeing?

	a)	b)
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

2. Interventions

   a) Given that the worker is forced to go to university, what is their wellbeing?

   b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

2. Interventions

   a) Given that the worker is forced to go to university, what is their wellbeing?

   b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

   a) Given that the worker didn't go to university, what would be their wellbeing if they had?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

   a) Given that the worker didn't go to university, what would be their wellbeing if they had?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$
3.	$\Pr^{\pi}(u^1_g \mid \neg g)$	

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

   a) Given that the worker didn't go to university, what would be their wellbeing if they had?

   b) Given that the worker never decides to go to university, what would be the firm's profits if they always decided to go to university?

	a)	**b)**
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$
3.	$\Pr^{\pi}(u^1_g \mid \neg g)$	

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

   a) Given that the worker didn't go to university, what would be their wellbeing if they had?

   b) Given that the worker never decides to go to university, what would be the firm's profits if they always decided to go to university?

	a)	b)
1.	$\Pr^{\pi}(u^1 \mid g)$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\Pr(u^2_{\hat{\pi}_{D^1}})$
3.	$\Pr^{\pi}(u^1_g \mid \neg g)$	$\Pr(u^2_{\bar{\pi}_{D^1}} \mid \tilde{\pi}_{D^1})$

A Causal Hierarchy for Games

- There are three main kinds of questions we might want to ask in games, with two variants of each

3. Counterfactuals

   a) Given that the worker didn't go to university, what would be their wellbeing if they had?

   b) Given that the worker never decides to go to university, what would be the firm's profits if they always decided to go to university?

	a)	b)
1.	$\boxed{\Pr^{\pi}(u^1 \mid g)}$	$\Pr(u^2 \mid \bar{\pi}_{D^1})$
2.	$\Pr^{\pi}(u^1_g)$	$\boxed{\Pr(u^2_{\hat{\pi}_{D^1}})}$
3.	$\boxed{\Pr^{\pi}(u^1_g \mid \neg g)}$	$\Pr(u^2_{\bar{\pi}_{D^1}} \mid \tilde{\pi}_{D^1})$

A Causal Hierarchy for Games

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy

SCM

CBN

BN

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy

SCM

CBN

BN

Model

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy

SCM	
CBN	**Model**
BN	
DAG	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]

SCM	
CBN	**Model**
BN	
DAG	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]

SCM	SCIM	
CBN	CIM	**Model**
BN	IM	
DAG	ID	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]
- By using MAIDs we generalise this hierarchy to a set of three models for representing games

SCM	SCIM	
CBN	CIM	**Model**
BN	IM	
DAG	ID	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]
- By using MAIDs we generalise this hierarchy to a set of three models for representing games

SCM	SCIM	MASCIM	
CBN	CIM	MACIM	**Model**
BN	IM	MAIM	
DAG	ID	MAID	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]
- By using MAIDs we generalise this hierarchy to a set of three models for representing games

Structure ↑

SCM	SCIM	MASCIM	
CBN	CIM	MACIM	**Model**
BN	IM	MAIM	
DAG	ID	MAID	**Graph**

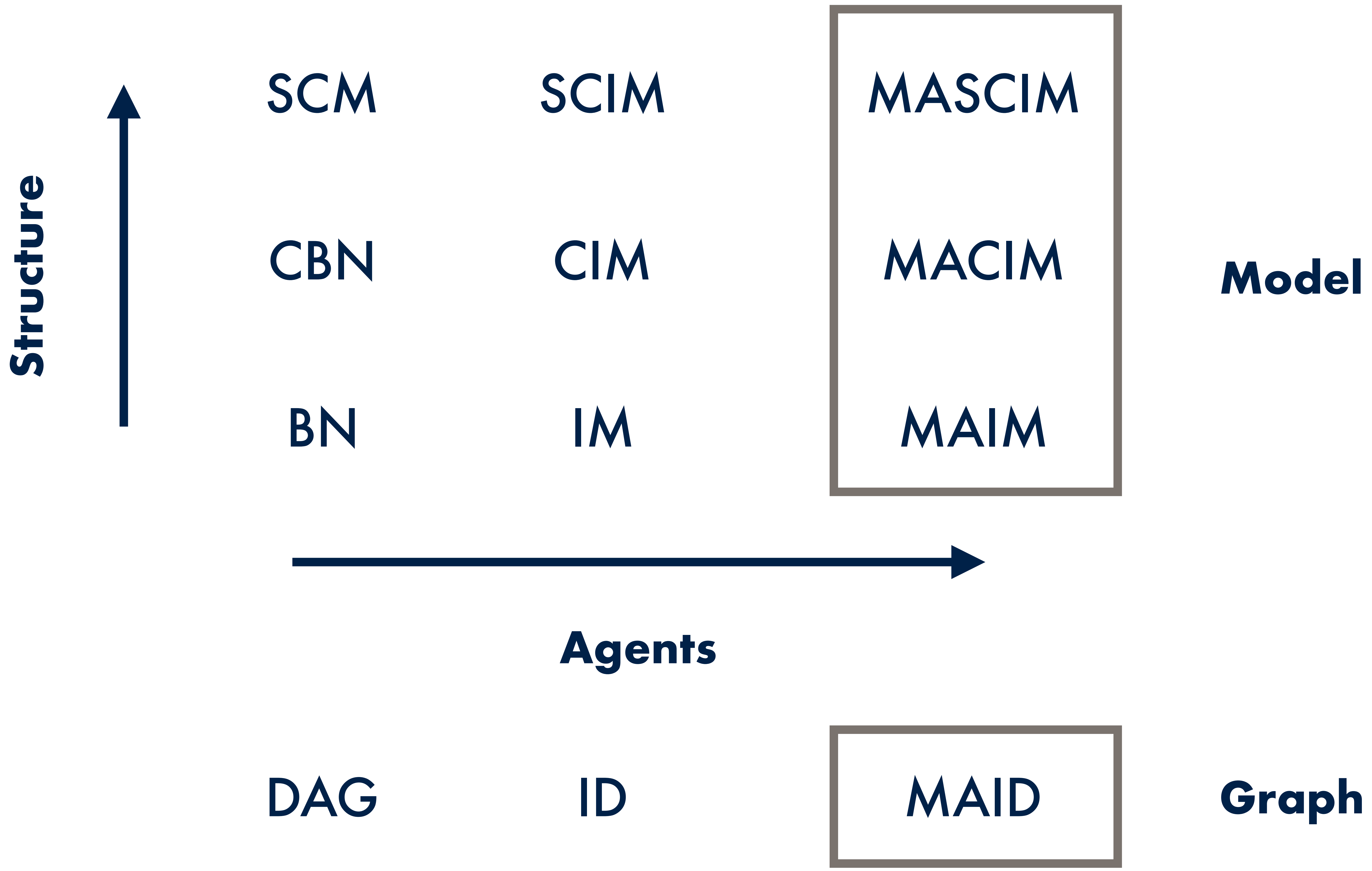
A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]
- By using MAIDs we generalise this hierarchy to a set of three models for representing games

Structure ↑	SCM	SCIM	MASCIM	
	CBN	CIM	MACIM	**Model**
	BN	IM	MAIM	
	Agents →			
	DAG	ID	MAID	**Graph**

A Causal Hierarchy for Games

- We start with Pearl's causal hierarchy
- Considering a (single) decision-maker leads to Influence Diagrams and resulting models [2,4]
- By using MAIDs we generalise this hierarchy to a set of three models for representing games



Predictions

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathcal{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:
  - There is no game-theoretic basis for only choosing a *single* policy

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:
  - There is no game-theoretic basis for only choosing a *single* policy
  - Given $\mathbf{z}$, we *learn about* π

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:
  - There is no game-theoretic basis for only choosing a *single* policy
  - Given $\mathbf{z}$, we *learn about* π
- Given an extended MAIM $\mathrm{x}\mathscr{M}$ with rationality relations $\mathscr{R}$, the answer to a conditional query of $\mathbf{x}$ given observation $\mathbf{z}$ is given by the set $\Pr^{\mathscr{R}}(\mathbf{x} \mid \mathbf{z}) := \left\{ \Pr^{\pi}(\mathbf{x} \mid \mathbf{z}) \right\}_{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z})}$

Predictions

- <u>Query:</u> Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^{\pi}(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:
  - There is no game-theoretic basis for only choosing a *single* policy
  - Given $\mathbf{z}$, we *learn about* π

- Given an extended MAIM $\mathrm{x}\mathscr{M}$ with rationality relations $\mathscr{R}$, the <u>answer to a conditional query</u> of $\mathbf{x}$ given observation $\mathbf{z}$ is given by the set $\Pr^{\mathscr{R}}(\mathbf{x} \mid \mathbf{z}) := \left\{ \Pr^{\pi}(\mathbf{x} \mid \mathbf{z}) \right\}_{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z})}$
- $\mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z}) := \{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M}) : \Pr^{\pi}(\mathbf{z}) > 0\}$ are the <u>conditional rational outcomes</u>

Predictions

- Query: Given an observation $\mathbf{z}$, what is the probability of $\mathbf{x}$?
- In a MAIM $\mathscr{M}$ this is simple if we have a particular policy π, as then we simply have a BN, giving $\Pr^\pi(\mathbf{x} \mid \mathbf{z})$, but what policy should we choose?
- Our main insights:
  - There is no game-theoretic basis for only choosing a *single* policy
  - Given $\mathbf{z}$, we *learn about* π
- Given an extended MAIM $\mathrm{x}\mathscr{M}$ with rationality relations $\mathscr{R}$, the answer to a conditional query of $\mathbf{x}$ given observation $\mathbf{z}$ is given by the set $\Pr^{\mathscr{R}}(\mathbf{x} \mid \mathbf{z}) := \left\{ \Pr^\pi(\mathbf{x} \mid \mathbf{z}) \right\}_{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z})}$
- $\mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z}) := \{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M}) : \Pr^\pi(\mathbf{z}) > 0\}$ are the conditional rational outcomes
- Generally, $\mathbf{Z} \subseteq \mathbf{V} \cup \mathbf{M}$ so we compute $\Pr^\pi(\mathbf{x} \mid \mathbf{z})$ in $\mathscr{M}$ as $\Pr(\mathbf{X} \mid \mathbf{z}, \mathbf{m}')$ in $\mathrm{x}\mathscr{M}$, where $\mathbf{M}' = \mathbf{M} \setminus \mathbf{Z}$ and $\mathbf{M}_\mathbf{D} = \pi$

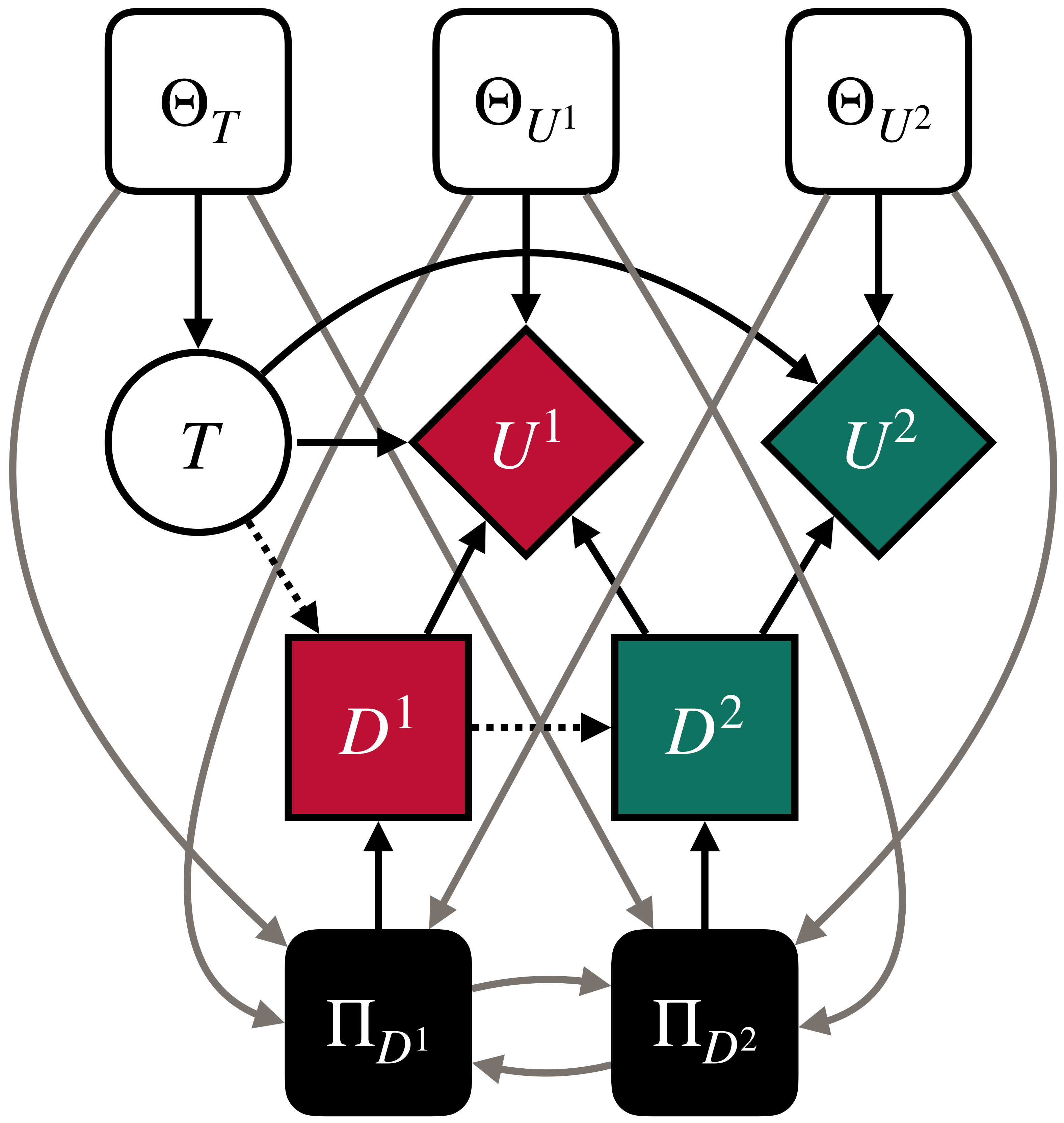
Predictions

Predictions

1. a) Given that the worker went to university, what is their wellbeing?

Predictions

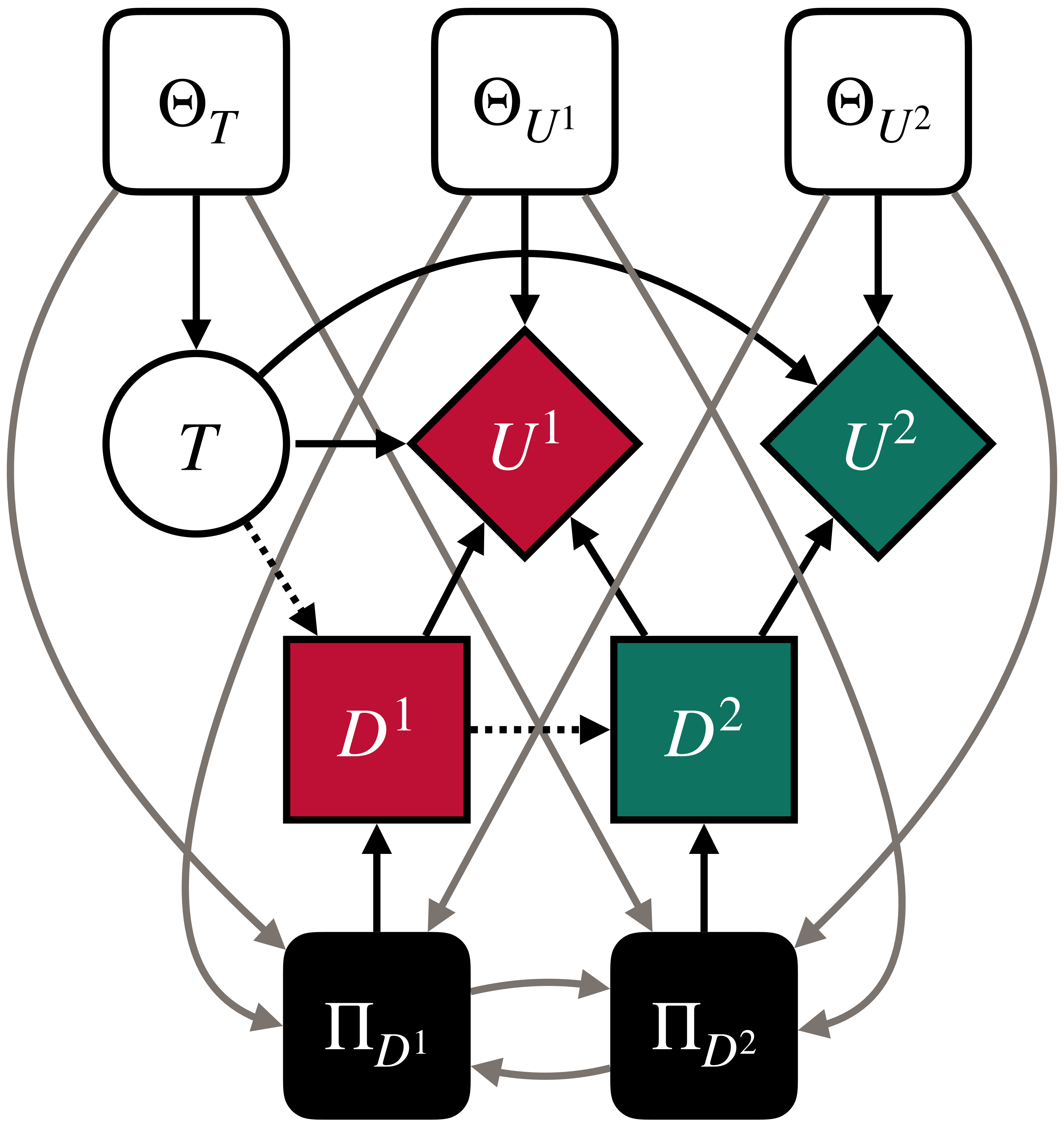
1. a) Given that the worker went to university, what is their wellbeing?



Predictions

1. a) Given that the worker went to university, what is their wellbeing?

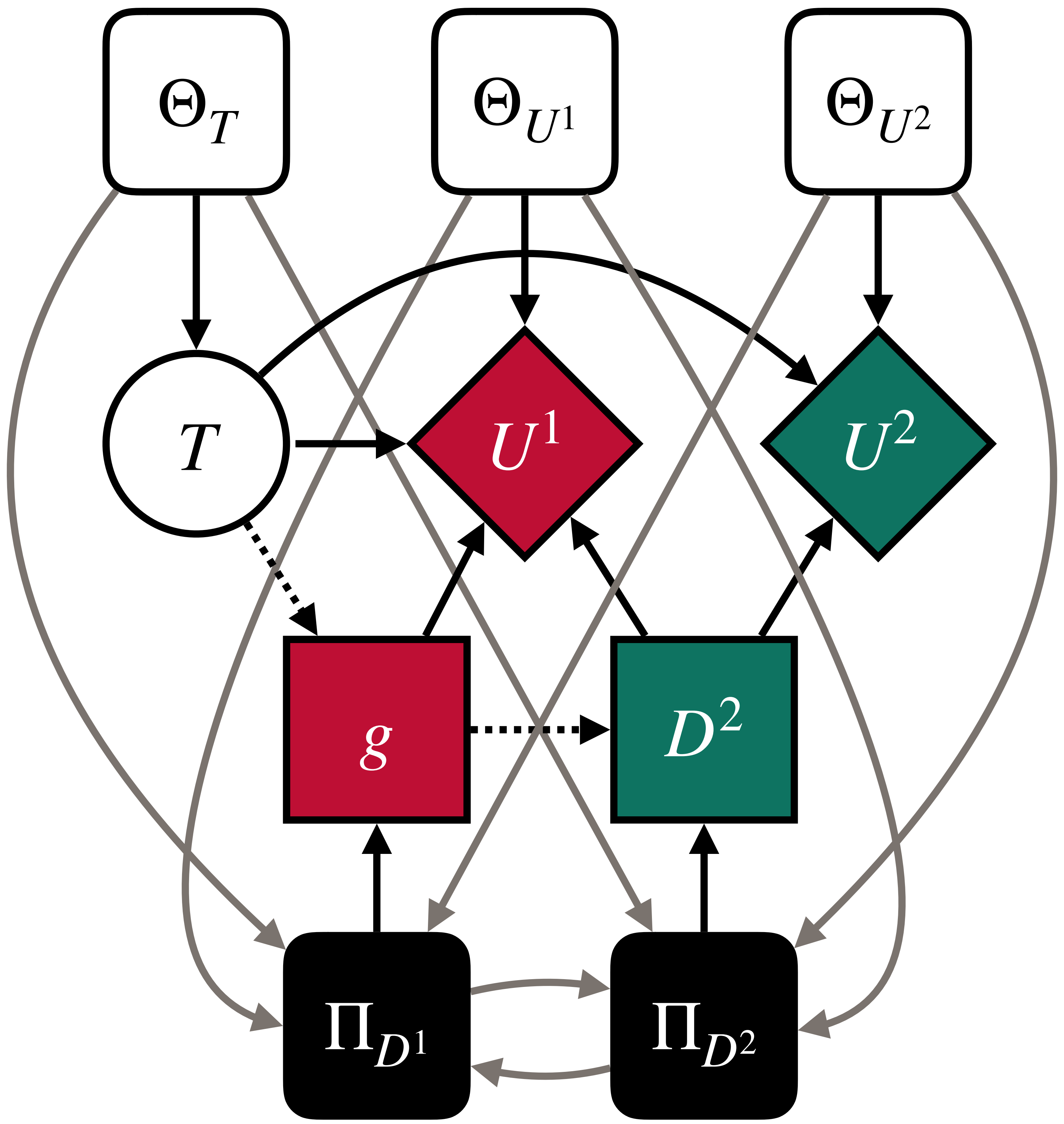
- Observe g and predict u^1



Predictions

1. a) Given that the worker went to university, what is their wellbeing?

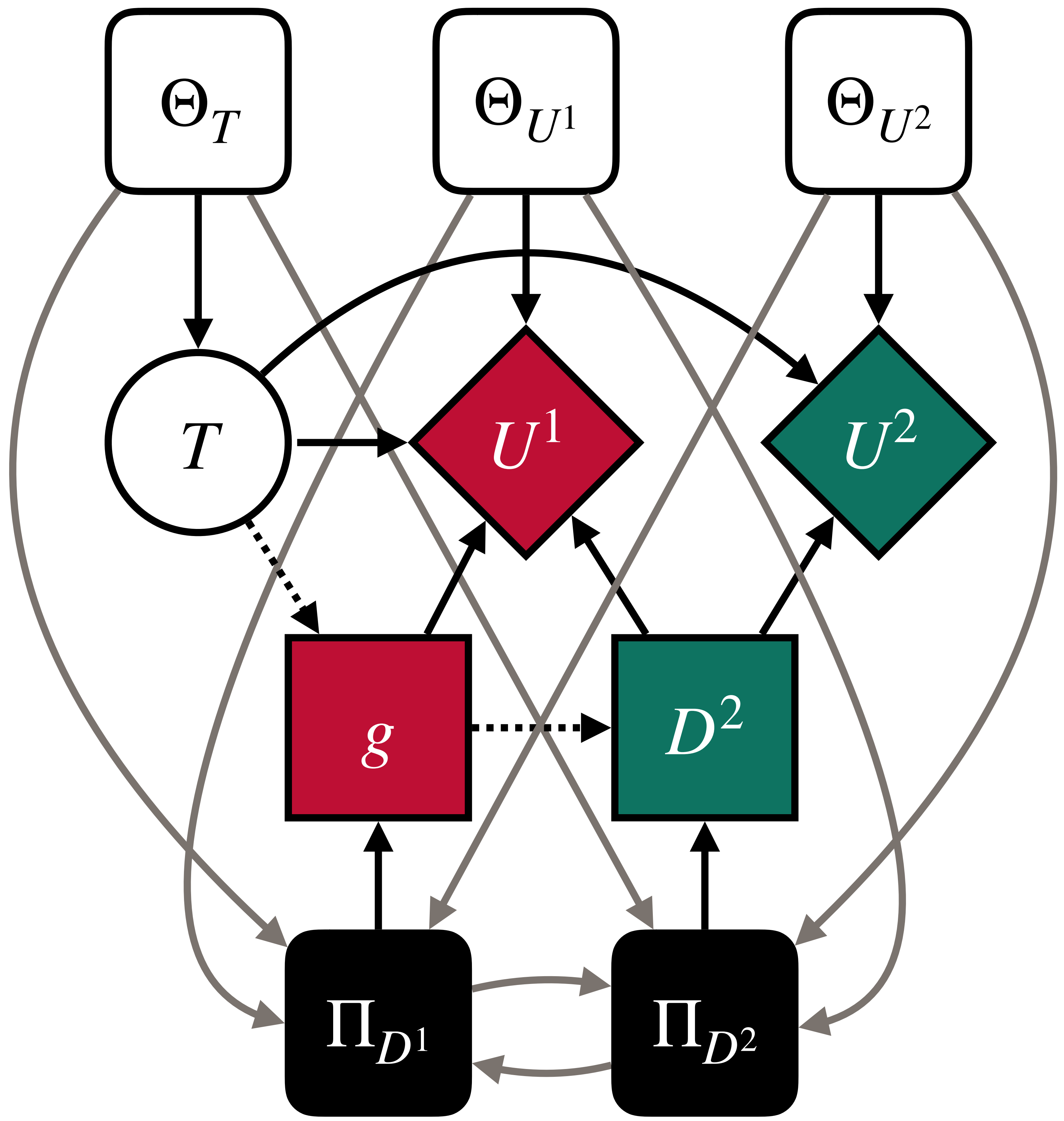
- Observe g and predict u^1



Predictions

1. a) Given that the worker went to university, what is their wellbeing?

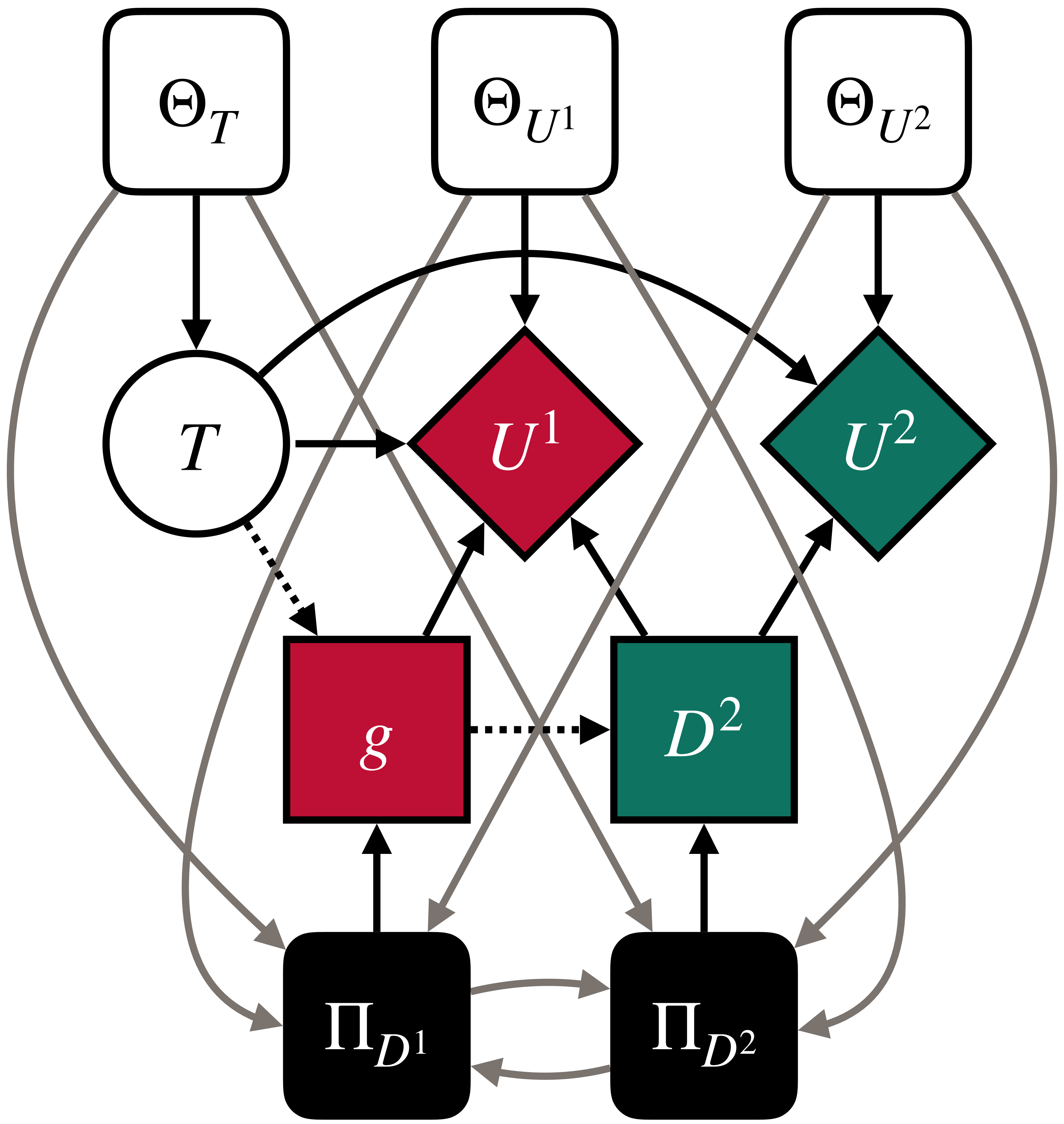
- Observe g and predict u^1
- First find $\mathscr{R}(\mathrm{x}\mathscr{M} \mid g)$, the set of all rational outcomes such that $D^1 = g$ with non-zero probability



Predictions

1. a) Given that the worker went to university, what is their wellbeing?

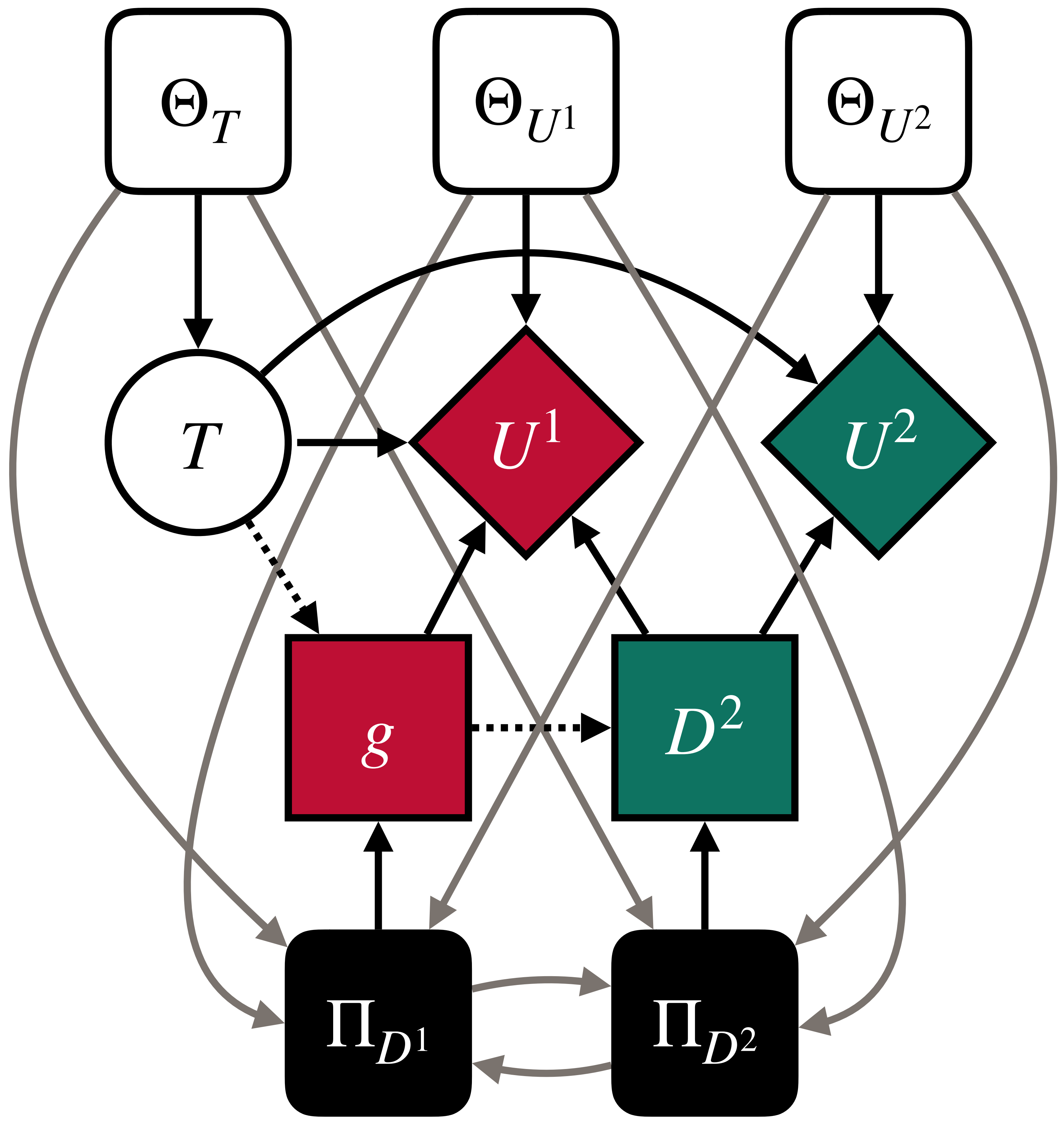
- Observe g and predict u^1
- First find $\mathscr{R}(\mathrm{x}\mathscr{M} \mid g)$, the set of all rational outcomes such that $D^1 = g$ with non-zero probability
  - E.g., NEs π where $\Pr^{\pi}(g) > 0$



Predictions

1. a) Given that the worker went to university, what is their wellbeing?

- Observe g and predict u^1
- First find $\mathscr{R}(\text{x}\mathscr{M} \mid g)$, the set of all rational outcomes such that $D^1 = g$ with non-zero probability
  - E.g., NEs π where $\Pr^\pi(g) > 0$
- Then for each $\pi \in \mathscr{R}(\text{x}\mathscr{M} \mid g)$, compute $\Pr^\pi(u^1 \mid g)$



Interventions

Interventions

- A MACIM is a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathscr{G}$ where:

Interventions

- A MACIM is a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathscr{G}$ where:
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = 1$ when $V \in \mathbf{Y}$ and v is consistent with $\mathbf{y}$

Interventions

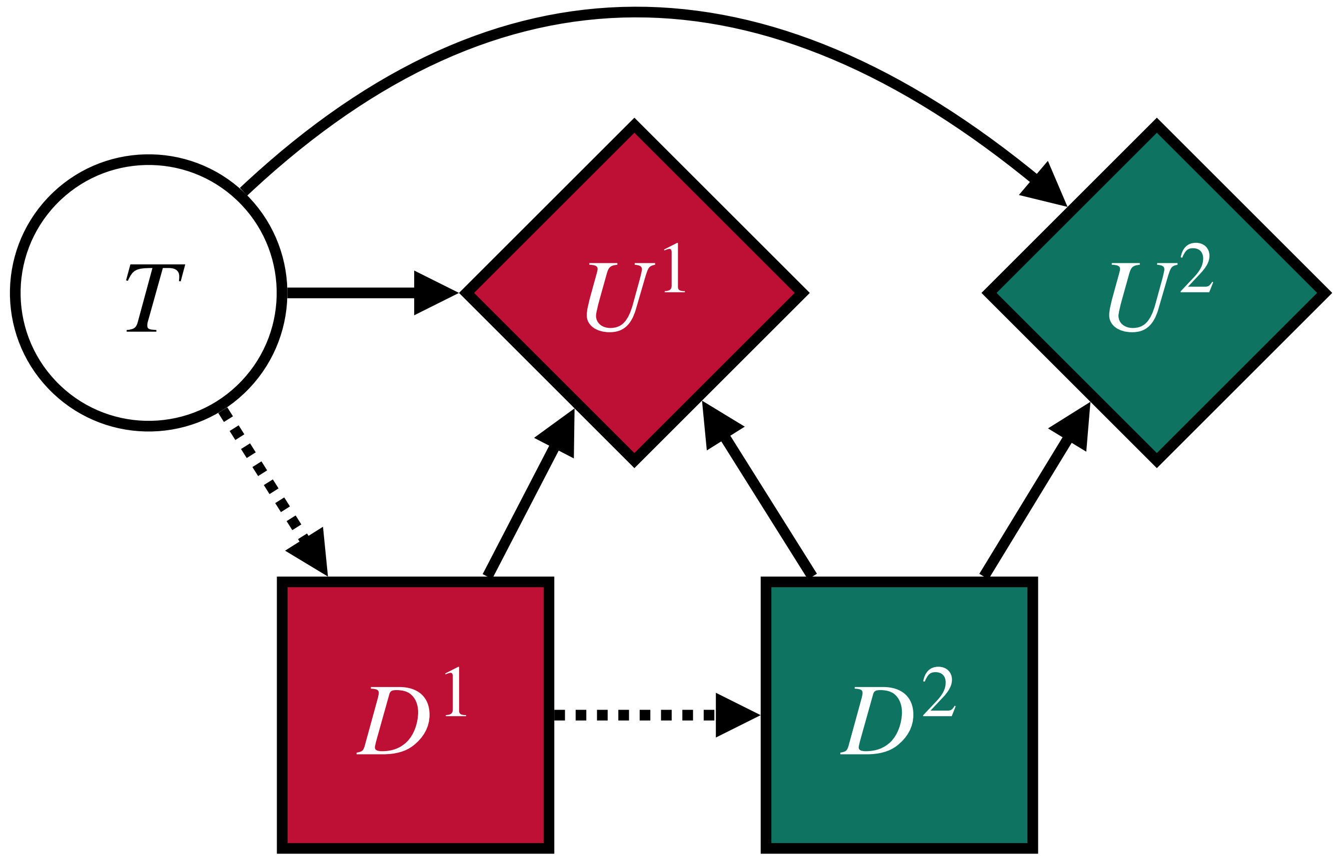
- A <u>MACIM</u> is a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathscr{G}$ where:
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = 1$ when $V \in \mathbf{Y}$ and v is consistent with $\mathbf{y}$
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = \Pr^{\pi}(v \mid \mathbf{pa}_V)$ when $V \in \mathbf{Y}$ and $\mathbf{pa}_V$ is consistent with $\mathbf{y}$

Interventions

- A MACIM is a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathscr{G}$ where:
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = 1$ when $V \in \mathbf{Y}$ and v is consistent with $\mathbf{y}$
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = \Pr^{\pi}(v \mid \mathbf{pa}_V)$ when $V \in \mathbf{Y}$ and $\mathbf{pa}_V$ is consistent with $\mathbf{y}$
- $\mathscr{M}$ is a CBN without parameters $\theta_{\mathbf{D}}$

Interventions

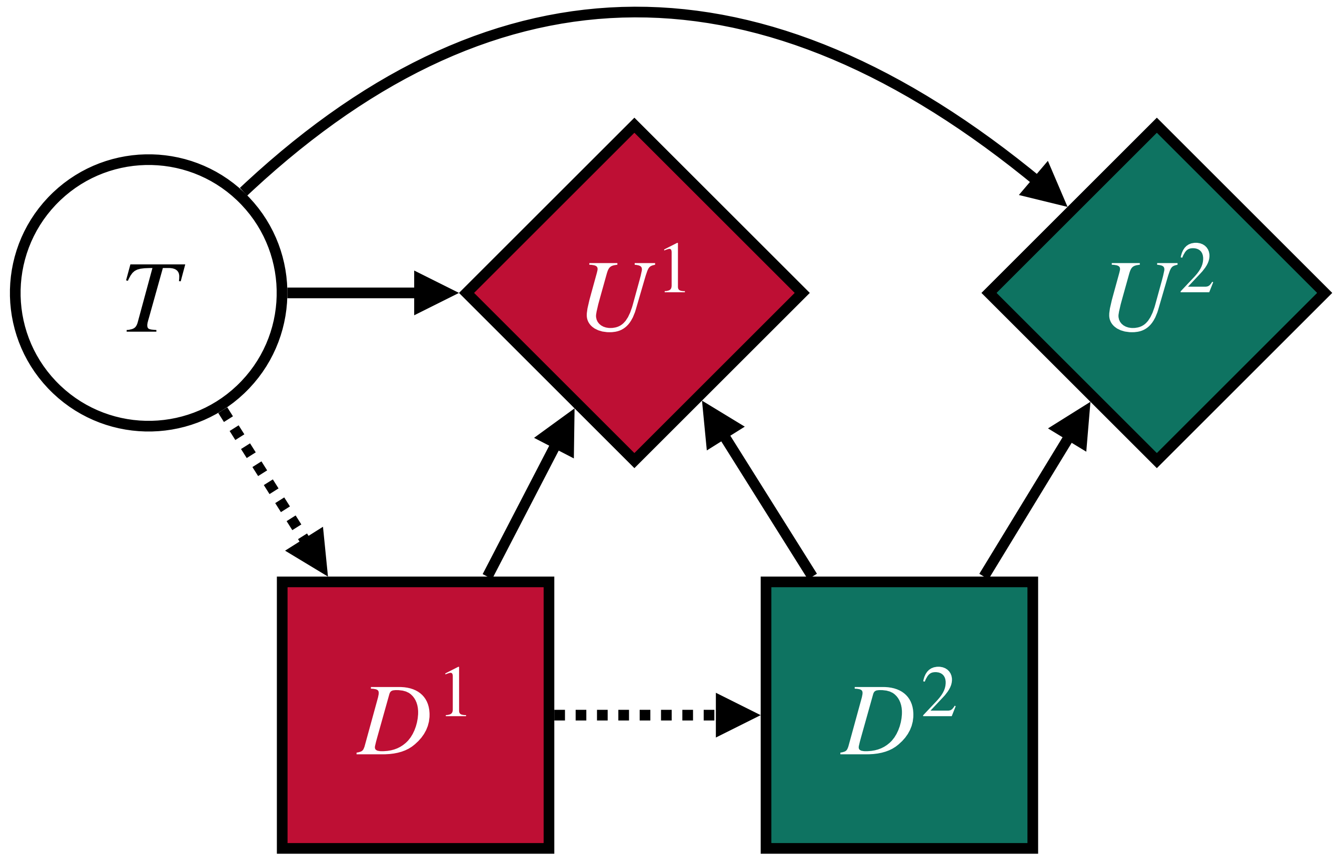
- A MACIM is a MAIM $\mathscr{M} = (\mathscr{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathscr{G}$ where:
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = 1$ when $V \in \mathbf{Y}$ and v is consistent with $\mathbf{y}$
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = \Pr^{\pi}(v \mid \mathbf{pa}_V)$ when $V \in \mathbf{Y}$ and $\mathbf{pa}_V$ is consistent with $\mathbf{y}$
- $\mathscr{M}$ is a CBN without parameters $\theta_{\mathbf{D}}$



Interventions

- A <u>MACIM</u> is a MAIM $\mathcal{M} = (\mathcal{G}, \theta)$ such that each interventional distribution $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}$ arising from an atomic intervention $\mathbf{y}$ and policy π is Markov-compatible with the MAID $\mathcal{G}$ where:
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = 1$ when $V \in \mathbf{Y}$ and v is consistent with $\mathbf{y}$
  - $\Pr^{\pi}_{\mathbf{Y} \leftarrow \mathbf{y}}(v \mid \mathbf{pa}_V) = \Pr^{\pi}(v \mid \mathbf{pa}_V)$ when $V \in \mathbf{Y}$ and $\mathbf{pa}_V$ is consistent with $\mathbf{y}$
- $\mathcal{M}$ is a CBN without parameters $\theta_{\mathbf{D}}$

- Moreover, the additional mechanism variables and their outgoing edges in an *extended MACIM* also represent causal (though potentially non-deterministic) processes



Interventions

Interventions

- <u>Query:</u> Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?

Interventions

- Query: Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?

Interventions

- Query: Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?
- Our main insight:

Interventions

- Query: Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?
- Our main insight:
  - Interventions on $\mathbf{V}$ correspond to *post-policy interventions*, and those on M correspond to *pre-policy interventions*

Interventions

- <u>Query:</u> Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?
- Our main insight:
  - Interventions on $\mathbf{V}$ correspond to *post-policy interventions*, and those on M correspond to *pre-policy interventions*
- After an intervention $\mathbf{Y} \leftarrow \mathbf{y}$ on $\mathbf{V}$ or M (i.e., $\mathbf{Y} \subseteq \mathbf{V} \cup$ M) we denote the resulting extended MACIM by $\mathrm{x}\mathscr{M}_{\mathbf{y}}$

Interventions

- <u>Query:</u> Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?
- Our main insight:
  - Interventions on $\mathbf{V}$ correspond to *post-policy interventions*, and those on M correspond to *pre-policy interventions*
- After an intervention $\mathbf{Y} \leftarrow \mathbf{y}$ on $\mathbf{V}$ or M (i.e., $\mathbf{Y} \subseteq \mathbf{V} \cup \mathrm{M}$) we denote the resulting extended MACIM by $\mathrm{x}\mathscr{M}_{\mathbf{y}}$
- Given an extended MACIM $\mathrm{x}\mathscr{M}$ with rationality relations $\mathscr{R}$, the <u>answer to an interventional query</u> of $\mathbf{x}$ given intervention $\mathbf{y}$ is given by the set $\Pr^{\mathscr{R}}(\mathbf{x}_{\mathbf{y}}) := \left\{ \Pr^{\pi}(\mathbf{x}_{\mathbf{y}}) \right\}_{\pi \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}})}$

Interventions

- Query: Given an intervention $\mathbf{Y} \leftarrow \mathbf{y}$, what is the probability of $\mathbf{x}$?
- As for predictions, we can answer these as first-order queries where π is a free variable, but 'when' in the course of play is the intervention made?
- Our main insight:
    - Interventions on $\mathbf{V}$ correspond to *post-policy interventions*, and those on M correspond to *pre-policy interventions*
- After an intervention $\mathbf{Y} \leftarrow \mathbf{y}$ on $\mathbf{V}$ or M (i.e., $\mathbf{Y} \subseteq \mathbf{V} \cup \mathrm{M}$) we denote the resulting extended MACIM by $\mathrm{x}\mathcal{M}_{\mathbf{y}}$
- Given an extended MACIM $\mathrm{x}\mathcal{M}$ with rationality relations $\mathcal{R}$, the answer to an interventional query of $\mathbf{x}$ given intervention $\mathbf{y}$ is given by the set $\Pr^{\mathcal{R}}(\mathbf{x}_{\mathbf{y}}) := \left\{ \Pr^{\pi}(\mathbf{x}_{\mathbf{y}}) \right\}_{\pi \in \mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}})}$
- $\mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}})$ are the interventional rational outcomes

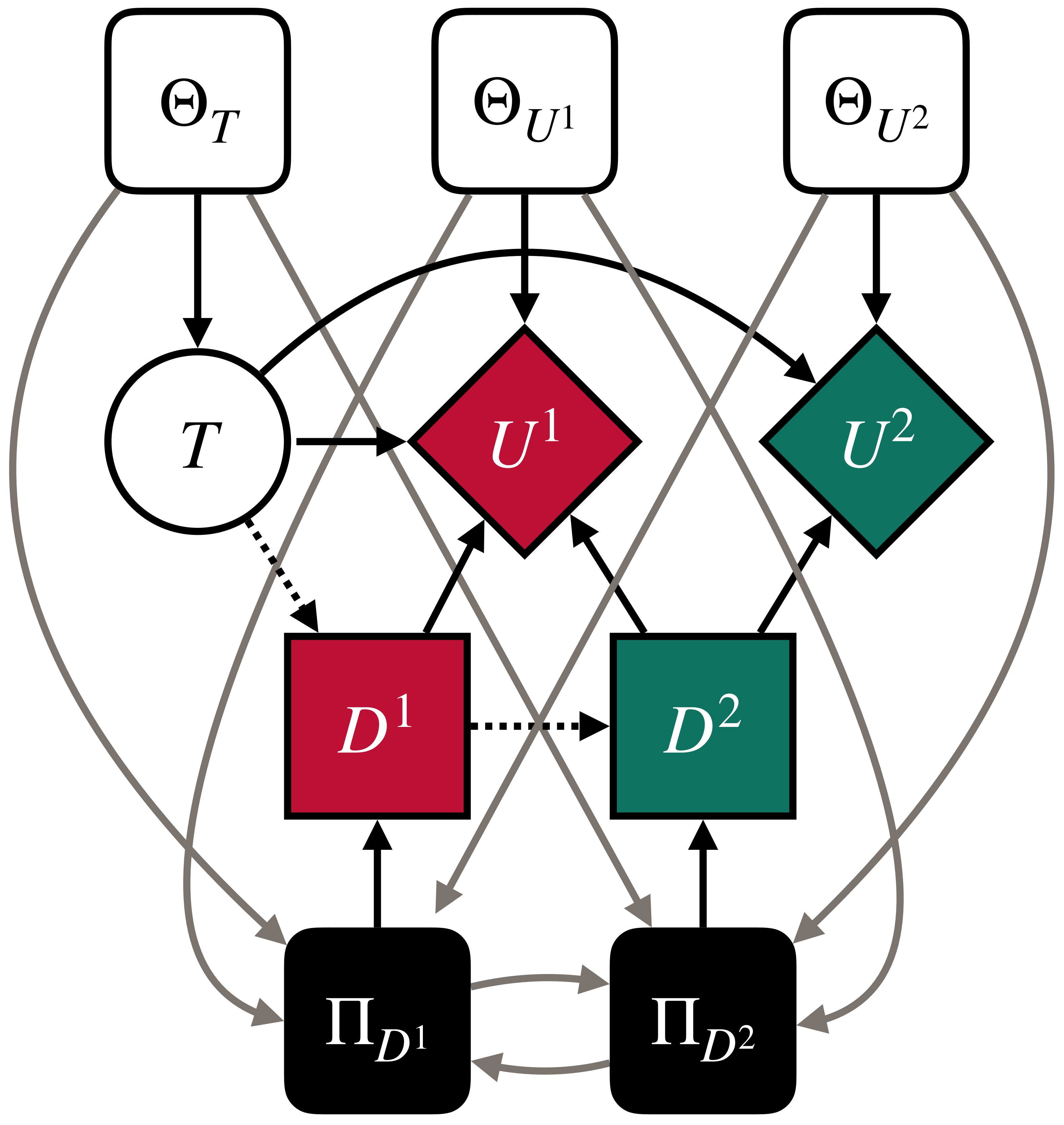
Interventions

Interventions

2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

Interventions

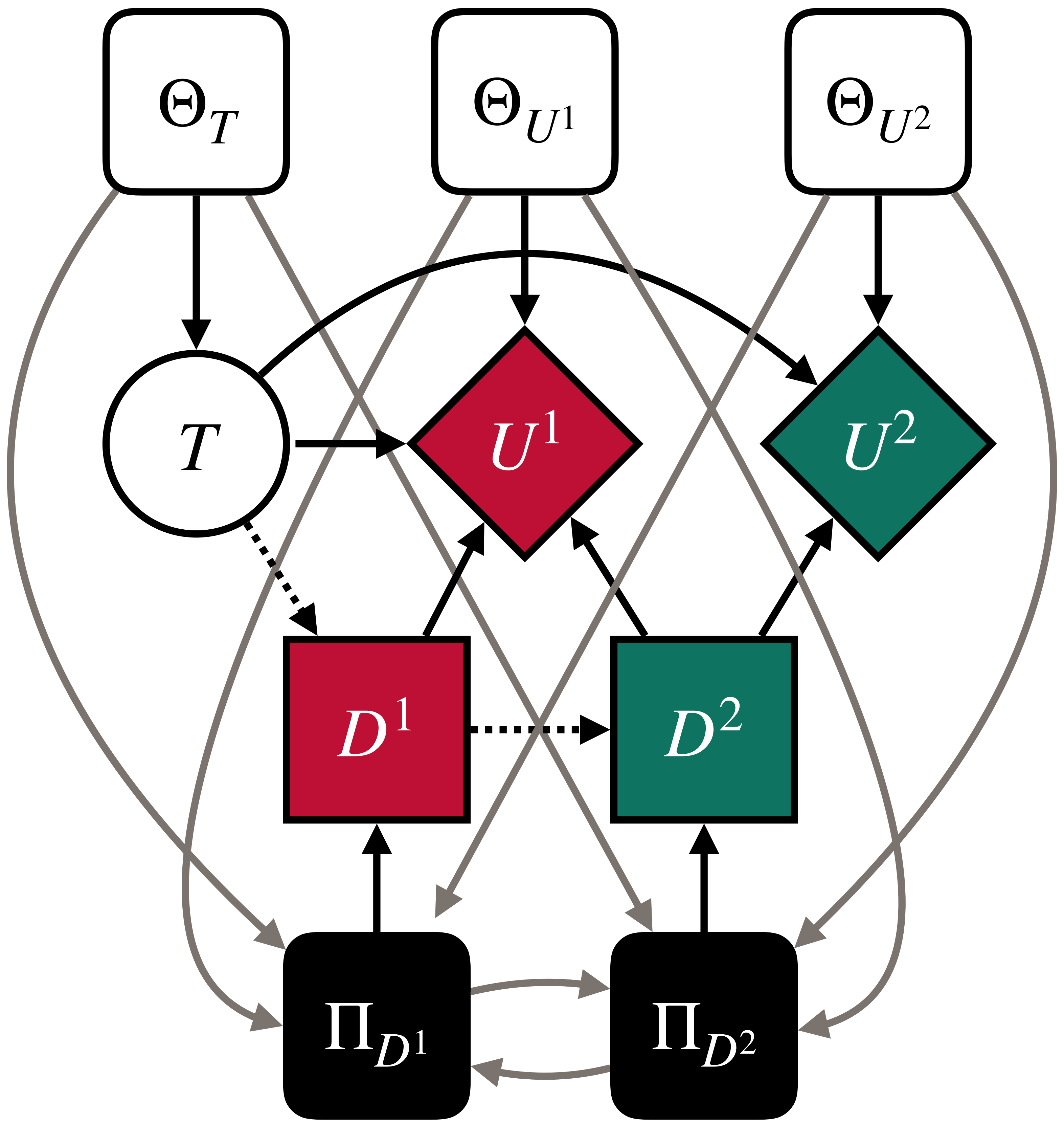
2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?



Interventions

2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

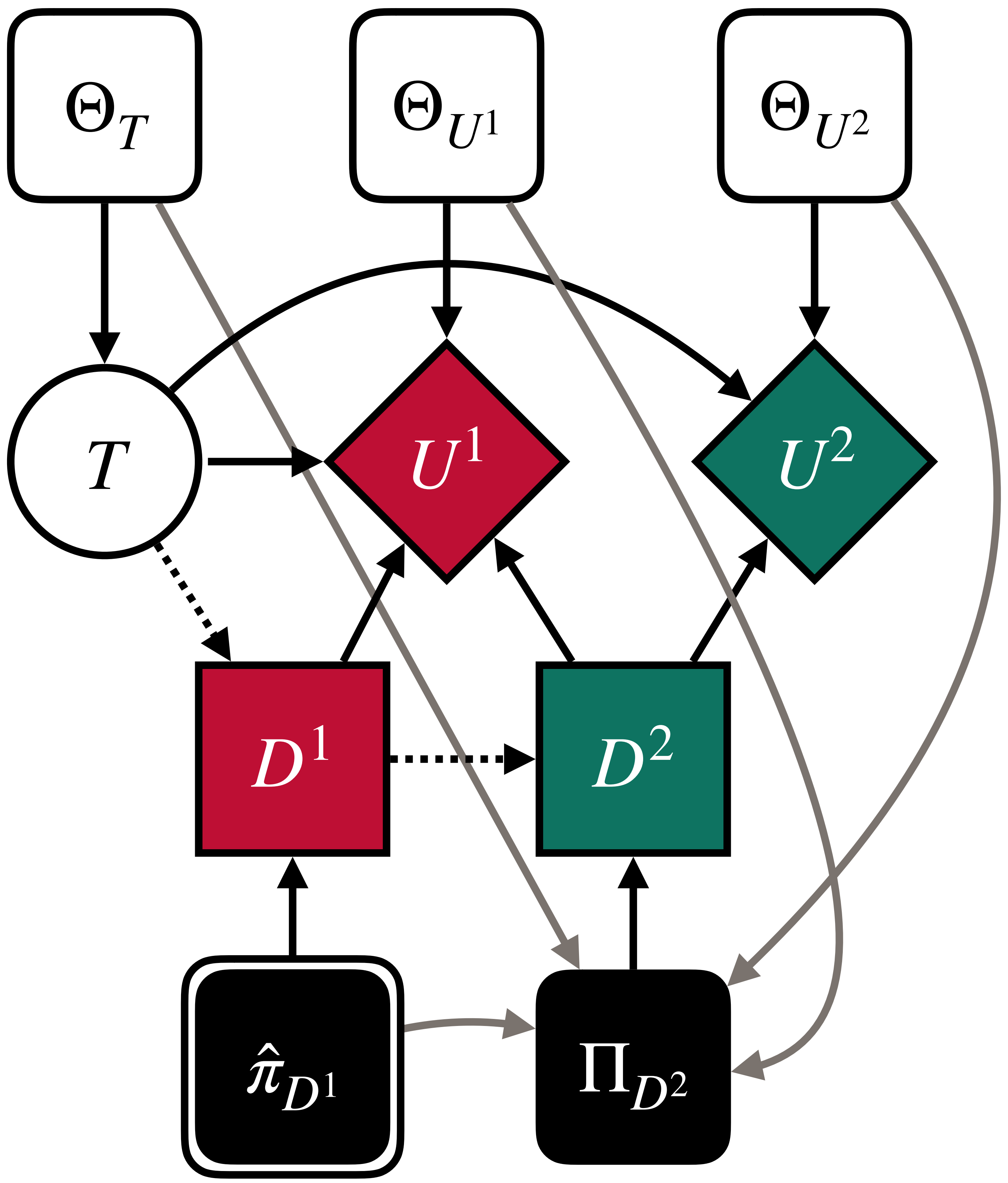
- Set $\Pi_{D^1} \leftarrow \hat{\pi}_{D^1}$ and predict u^1



Interventions

2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

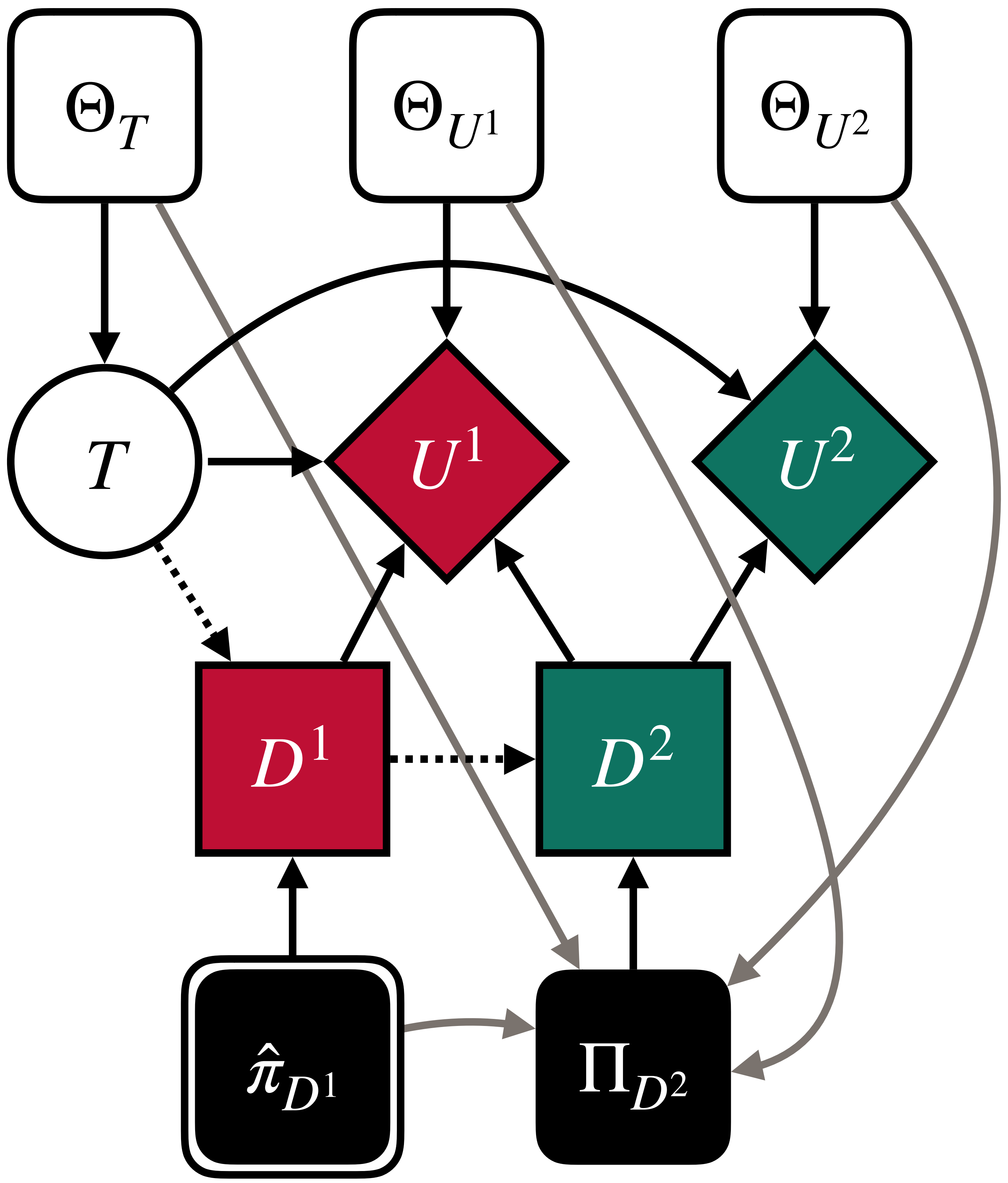
- Set $\Pi_{D^1} \leftarrow \hat{\pi}_{D^1}$ and predict u^1



Interventions

2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

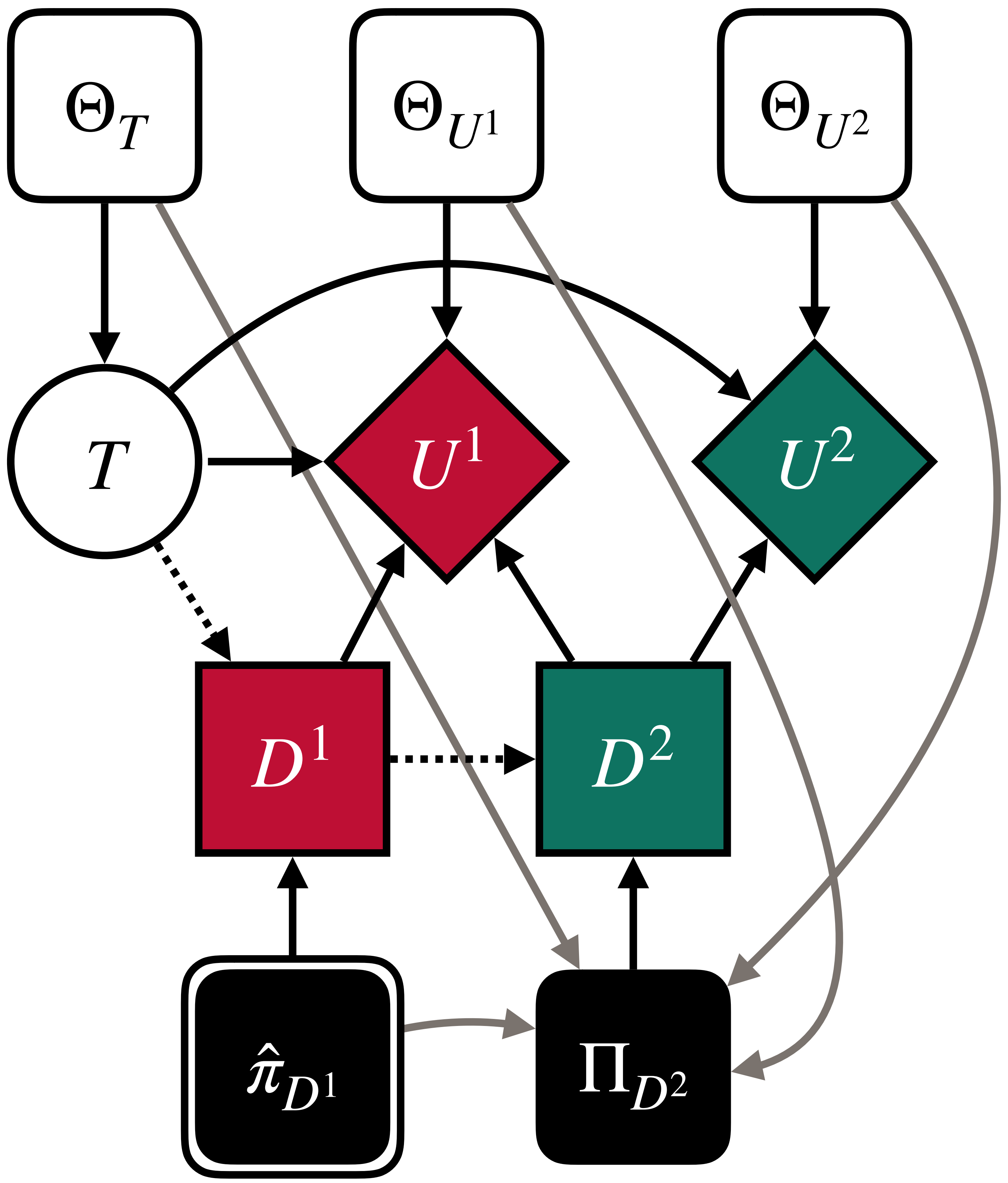
- Set $\Pi_{D^1} \leftarrow \hat{\pi}_{D^1}$ and predict u^1
- It is easy to see that we have $\mathscr{R}(\mathsf{x}\mathscr{M}_{\hat{\pi}_{D^1}}) = \left\{(\hat{\pi}_{D^1}, \pi_{D^2}) : \pi_{D^2} \in r_{D^2}(\hat{\pi}_{D^1})\right\}$



Interventions

2. b) Given that the worker goes to university if and only if they are selected via a lottery system, what are the firm's profits?

- Set $\Pi_{D^1} \leftarrow \hat{\pi}_{D^1}$ and predict u^1
- It is easy to see that we have $\mathscr{R}(\mathsf{x}\mathscr{M}_{\hat{\pi}_{D^1}}) = \left\{(\hat{\pi}_{D^1}, \pi_{D^2}) : \pi_{D^2} \in r_{D^2}(\hat{\pi}_{D^1})\right\}$
- Then for each $\pi \in \mathscr{R}(\mathsf{x}\mathscr{M}_{\hat{\pi}_{D^1}})$, compute $\Pr(u^2_{\hat{\pi}_{D^1}}) = \Pr^{(\hat{\pi}_{D^1}, \pi_{D^2})}(u^2)$



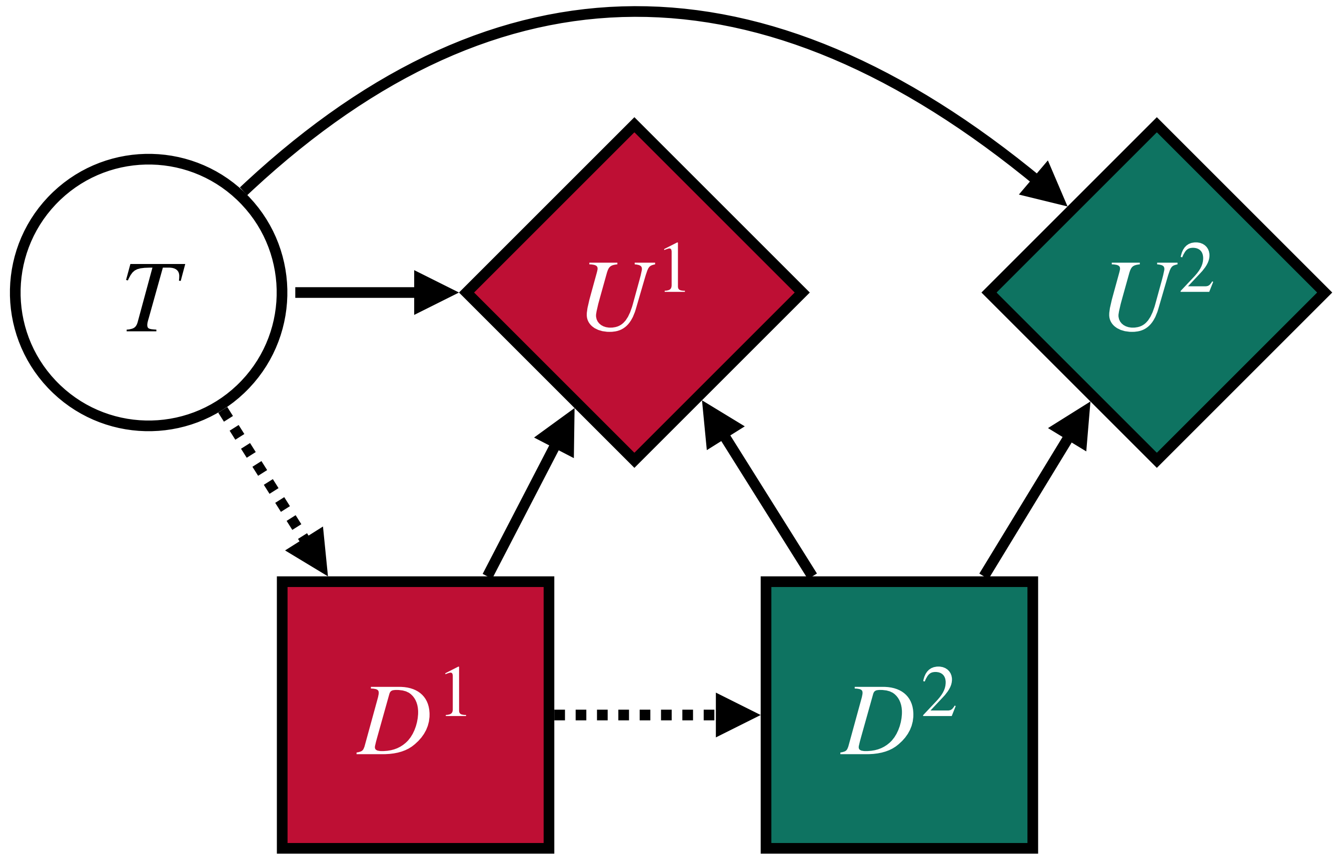
Counterfactuals

Counterfactuals

- A (Markovian) MASCIM is a MACIM $\mathcal{M} = (\mathcal{G}, \theta)$ where the MAID $\mathcal{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:

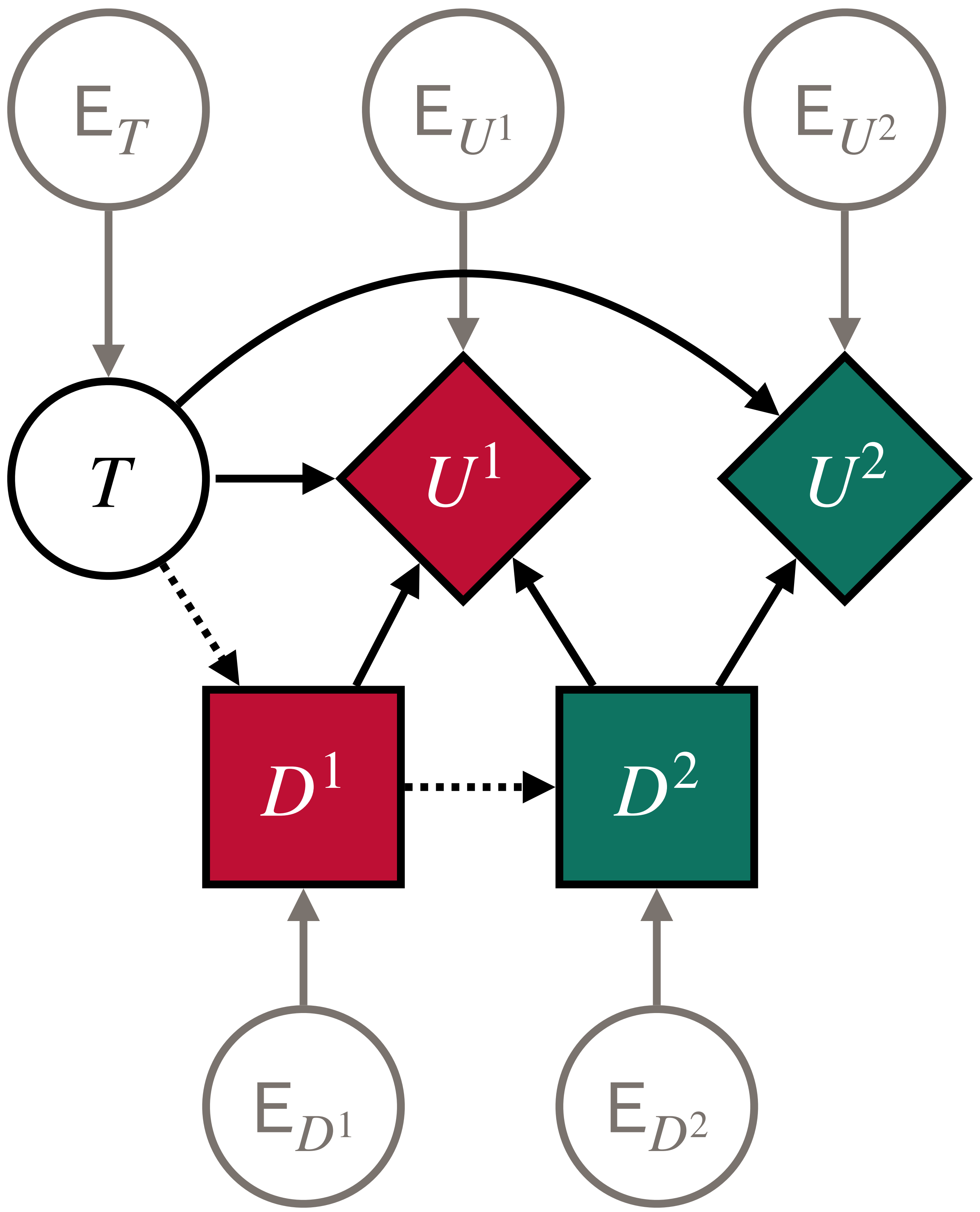
Counterfactuals

- A (Markovian) MASCIM is a MACIM $\mathscr{M} = (\mathscr{G}, \theta)$ where the MAID $\mathscr{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:



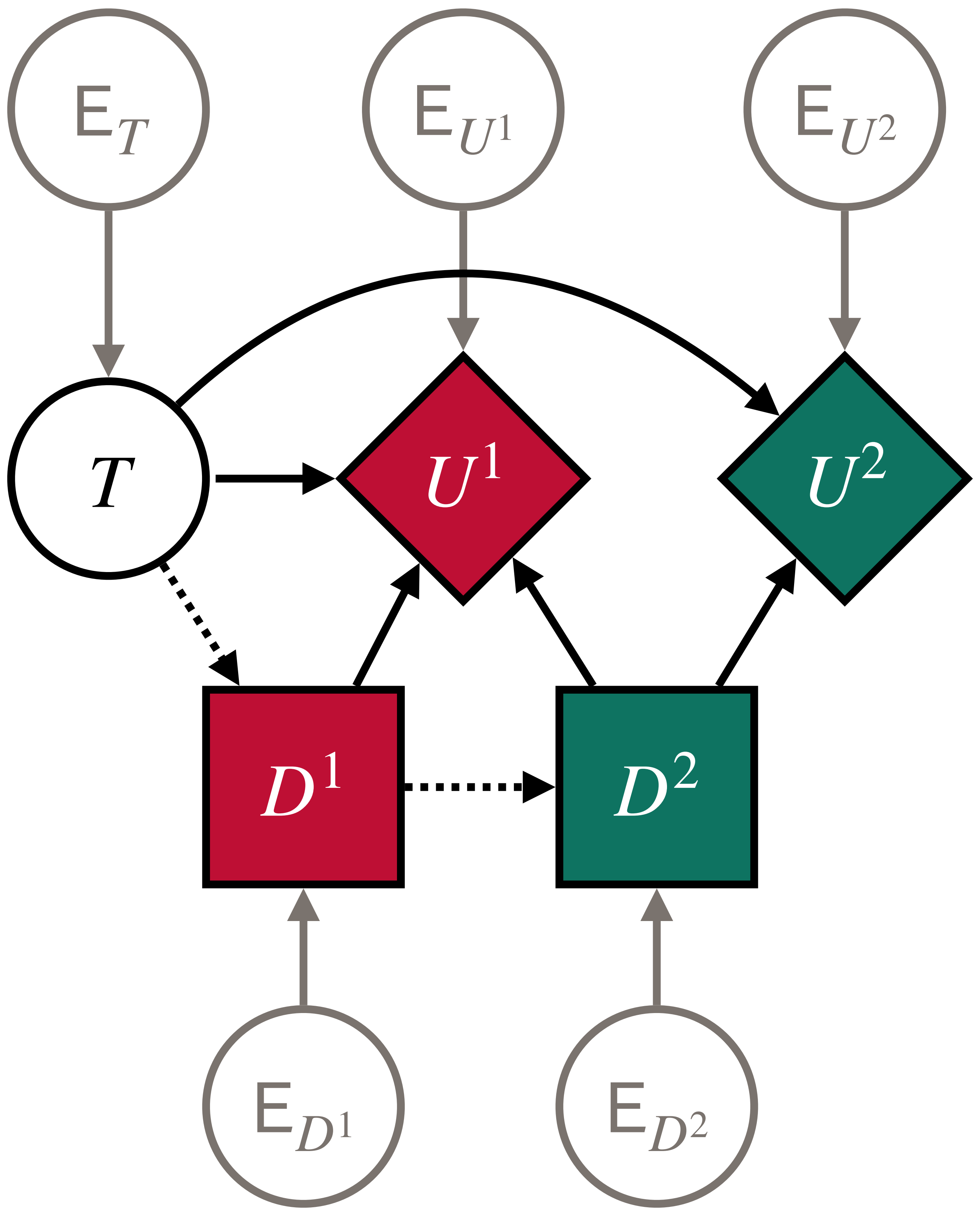
Counterfactuals

- A (Markovian) MASCIM is a MACIM $\mathscr{M} = (\mathscr{G}, \theta)$ where the MAID $\mathscr{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:



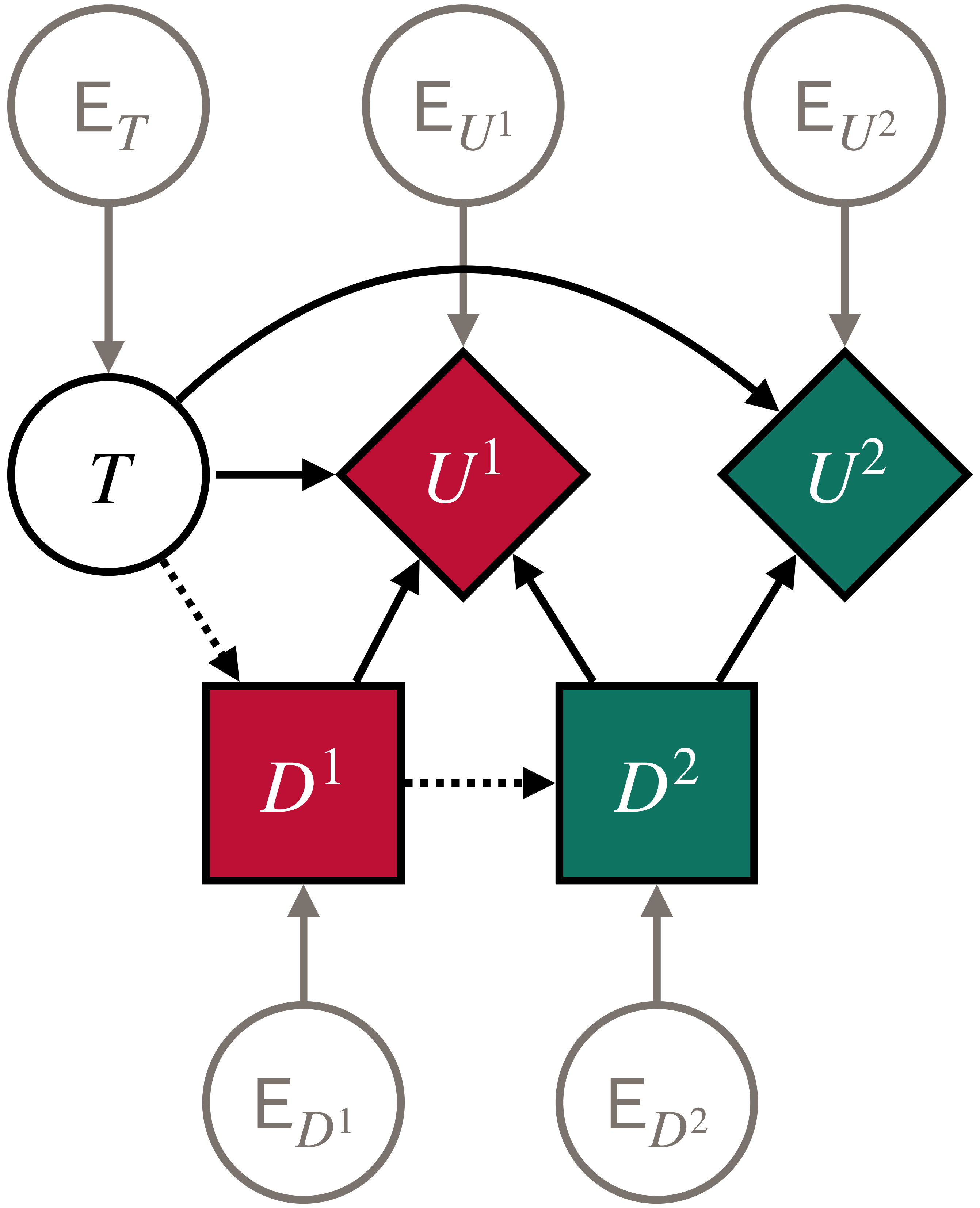
Counterfactuals

- A (Markovian) MASCIM is a MACIM $\mathcal{M} = (\mathcal{G}, \theta)$ where the MAID $\mathcal{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:
  - $\Pr^\pi(\mathbf{e}) = \prod_E \Pr^\pi(\mathsf{e}_V)$



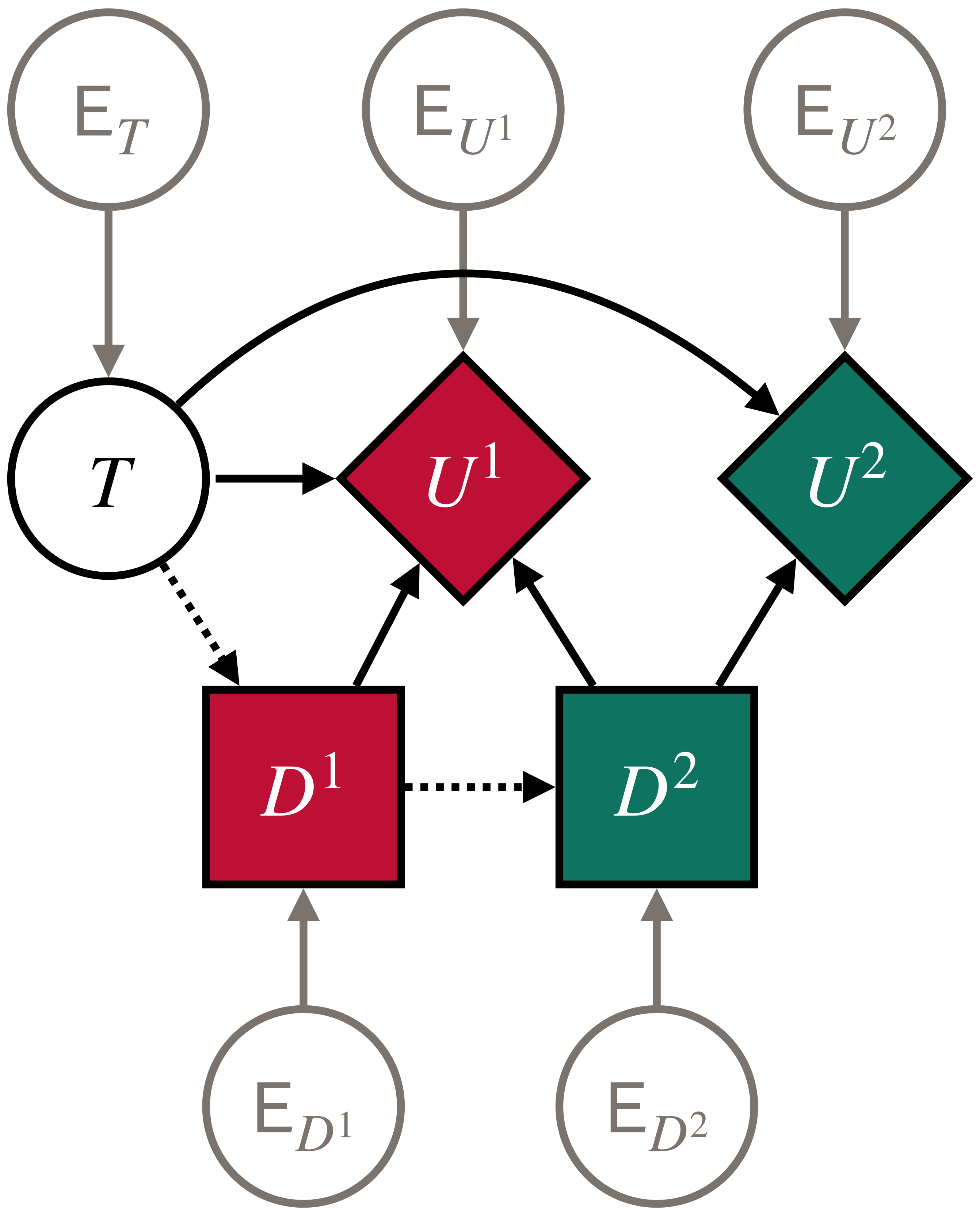
Counterfactuals

- A (Markovian) <u>MASCIM</u> is a MACIM $\mathscr{M} = (\mathscr{G}, \theta)$ where the MAID $\mathscr{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:
  - $\Pr^\pi(\mathbf{e}) = \prod_E \Pr^\pi(\mathsf{e}_V)$
  - The distribution $\Pr^\pi(V \mid \mathbf{Pa}_V)$ is deterministic for every $V \in \mathbf{V}$

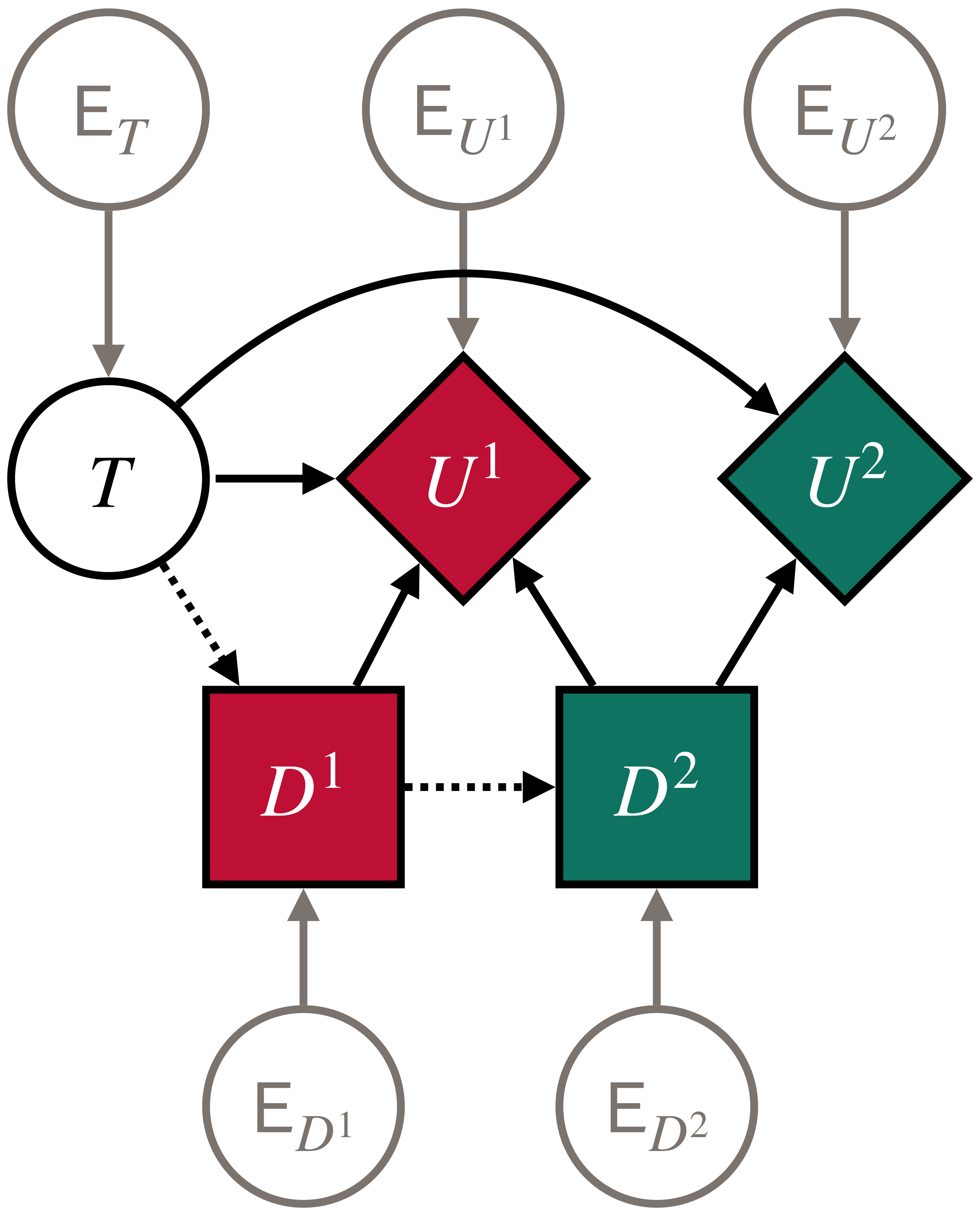


Counterfactuals

- A (Markovian) MASCIM is a MACIM $\mathscr{M} = (\mathscr{G}, \theta)$ where the MAID $\mathscr{G} = (N, \mathbf{V} \cup \mathsf{E}, \mathbb{E} \cup (\mathsf{E}_V, V)_{V \in \mathbf{V}})$ has additional exogenous variables E and edges $(\mathsf{E}_V, V)_{V \in \mathbf{V}}$ and for any π:
  - $\Pr^\pi(\mathbf{e}) = \prod_E \Pr^\pi(\mathsf{e}_V)$
  - The distribution $\Pr^\pi(V \mid \mathbf{Pa}_V)$ is deterministic for every $V \in \mathbf{V}$
- $\mathscr{M}$ is an SCM without parameters $\theta_\mathbf{D}$ or $\theta_{\mathsf{E}_\mathbf{D}}$

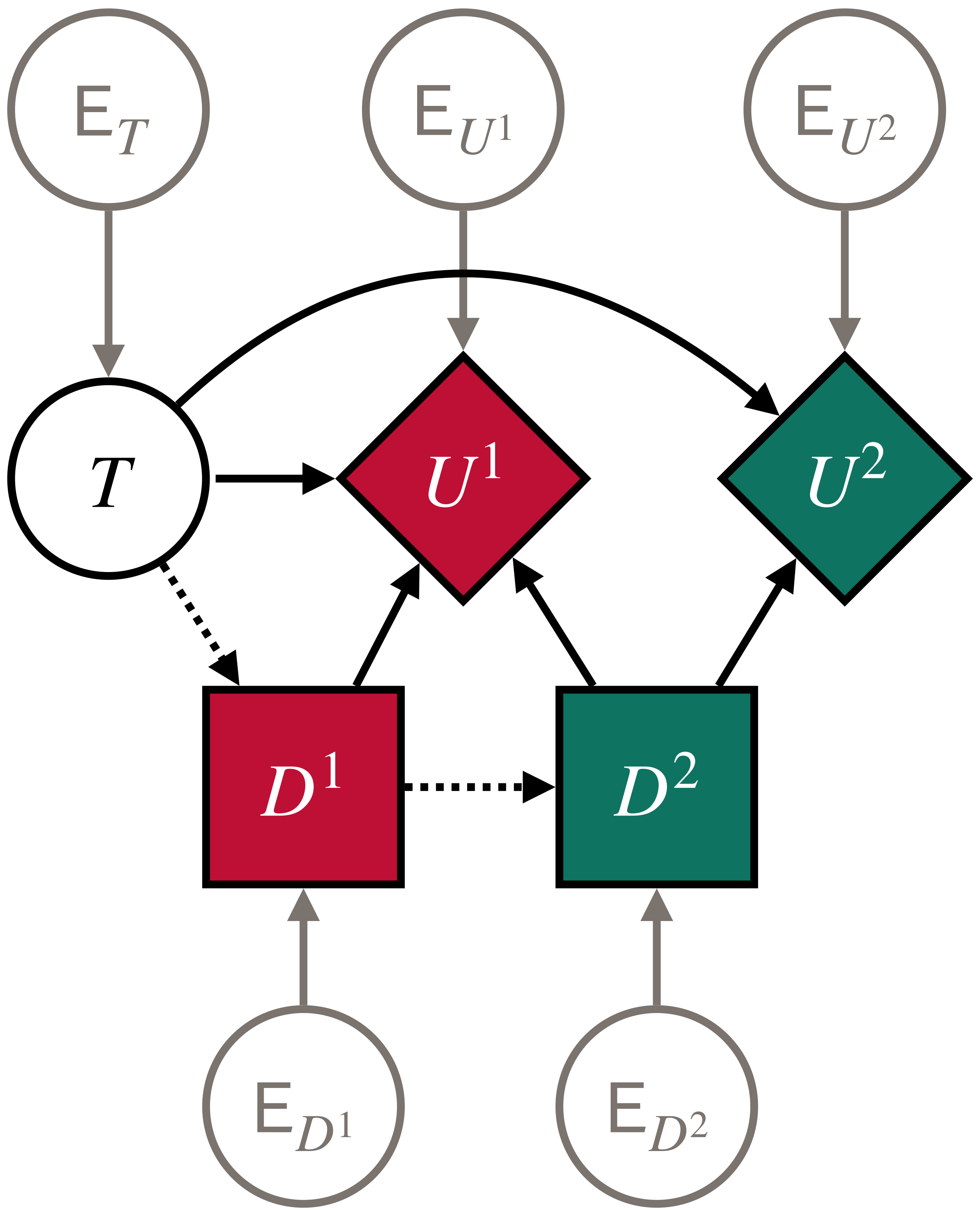


Counterfactuals



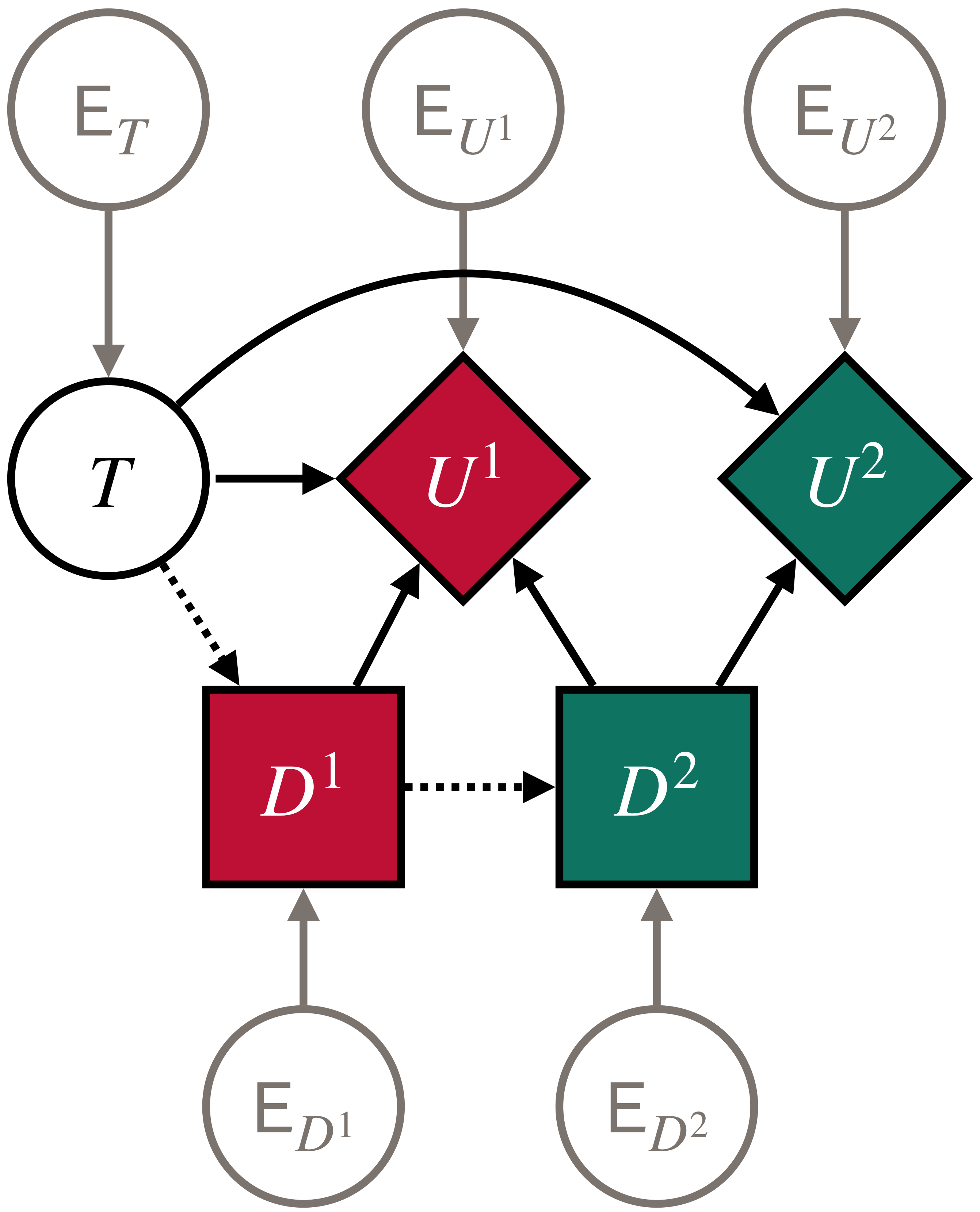
Counterfactuals

- In general, game-theoretic arguments are only sufficient for telling us the (stochastic) CPD $\pi_D(d \mid \mathbf{pa}'_D)$, where $\mathbf{Pa}'_D = \mathbf{Pa}_D \setminus \{\mathsf{E}_D\}$



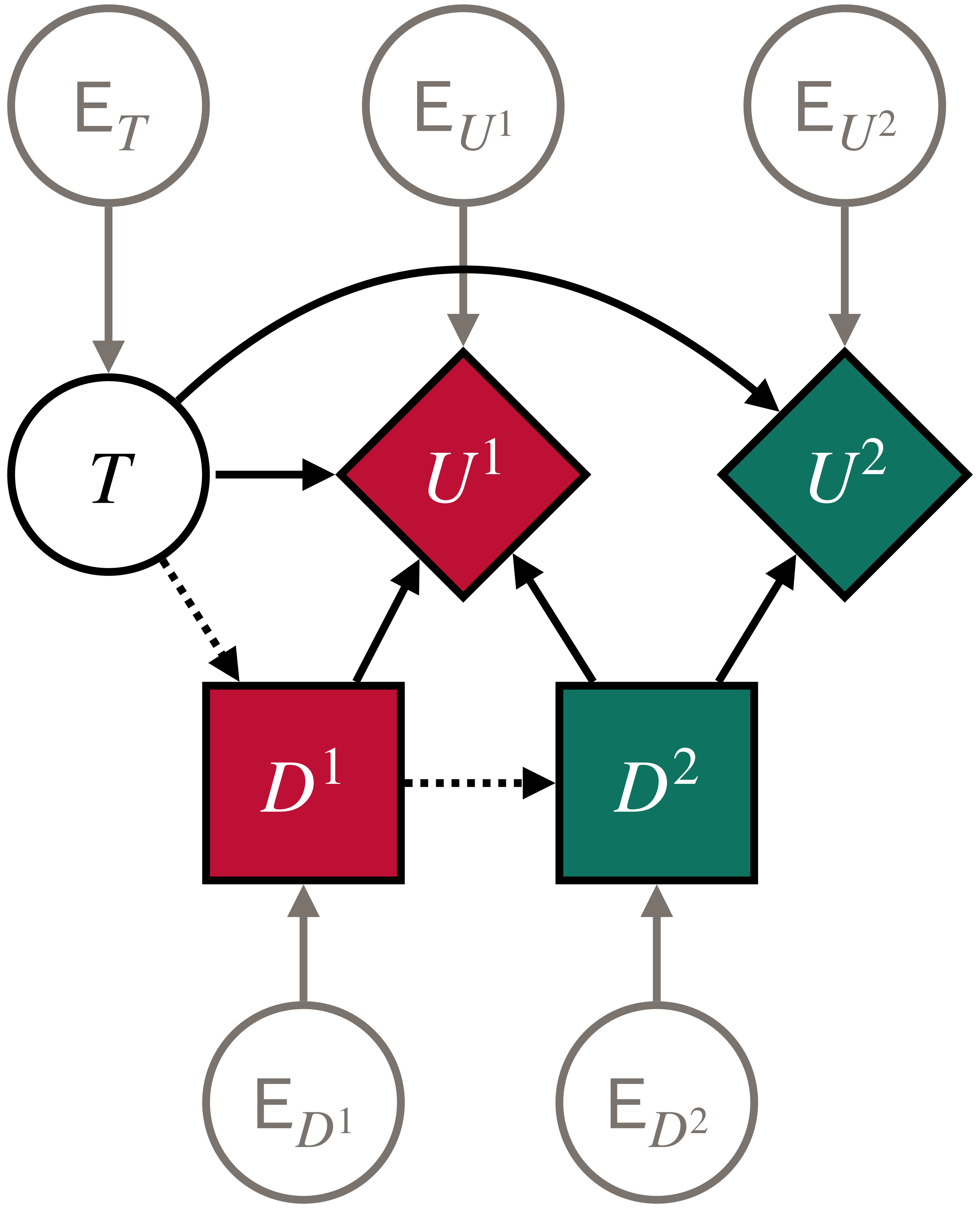
Counterfactuals

- In general, game-theoretic arguments are only sufficient for telling us the (stochastic) CPD $\pi_D(d \mid \mathbf{pa}'_D)$, where $\mathbf{Pa}'_D = \mathbf{Pa}_D \backslash \{\mathsf{E}_D\}$
- How should we express this CPD using a deterministic function and stochastic exogenous variable?



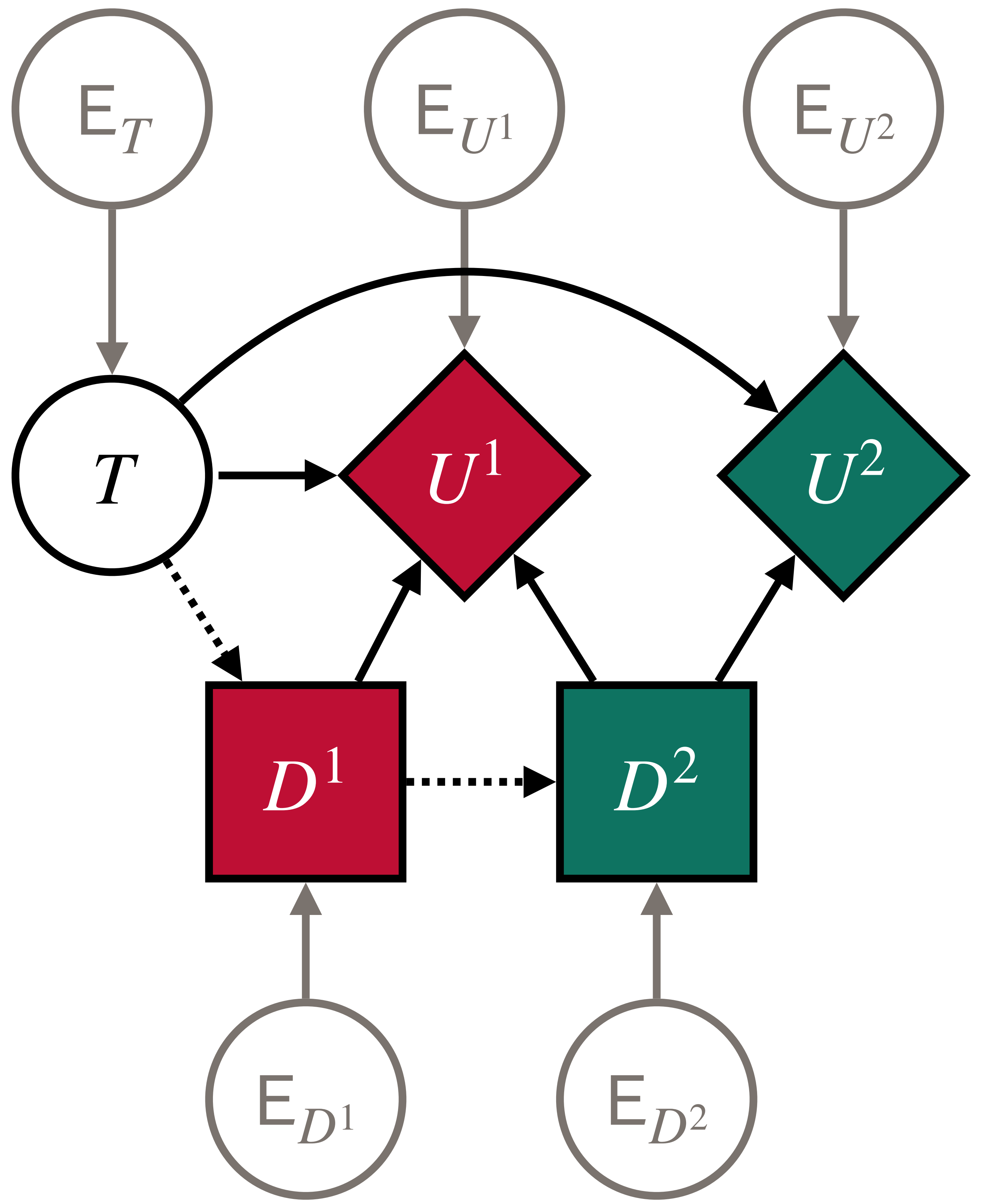
Counterfactuals

- In general, game-theoretic arguments are only sufficient for telling us the (stochastic) CPD $\pi_D(d \mid \mathbf{pa}'_D)$, where $\mathbf{Pa}'_D = \mathbf{Pa}_D \setminus \{\mathsf{E}_D\}$
- How should we express this CPD using a deterministic function and stochastic exogenous variable?
- Our main insight:

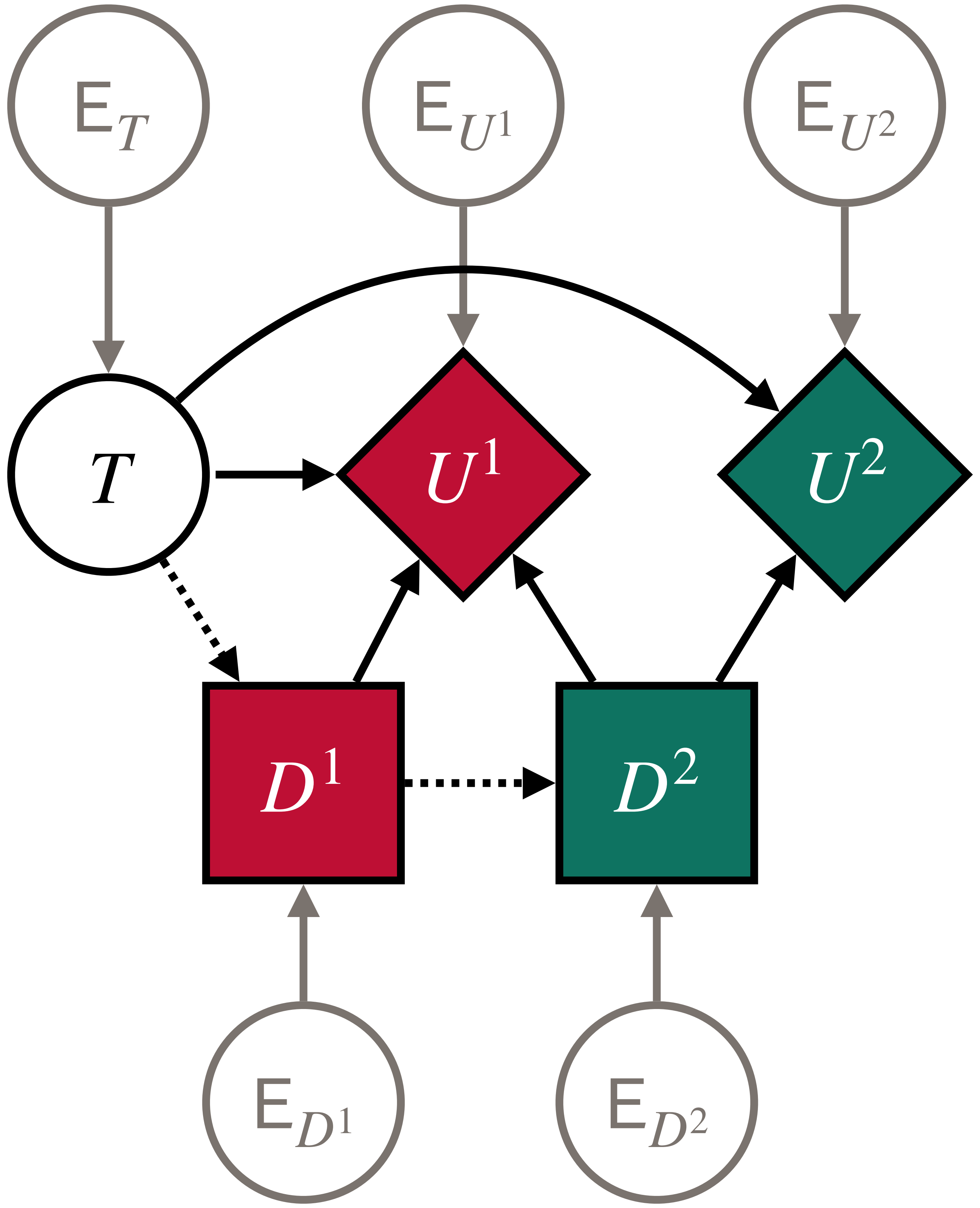


Counterfactuals

- In general, game-theoretic arguments are only sufficient for telling us the (stochastic) CPD $\pi_D(d \mid \mathbf{pa}'_D)$, where $\mathbf{Pa}'_D = \mathbf{Pa}_D \backslash \{\mathsf{E}_D\}$
- How should we express this CPD using a deterministic function and stochastic exogenous variable?
- Our main insight:
  - Without further knowledge about the function/randomisation, it's reasonable to model agents as (stochastically) choosing a decision d *after* seeing $\mathbf{pa}'_D$

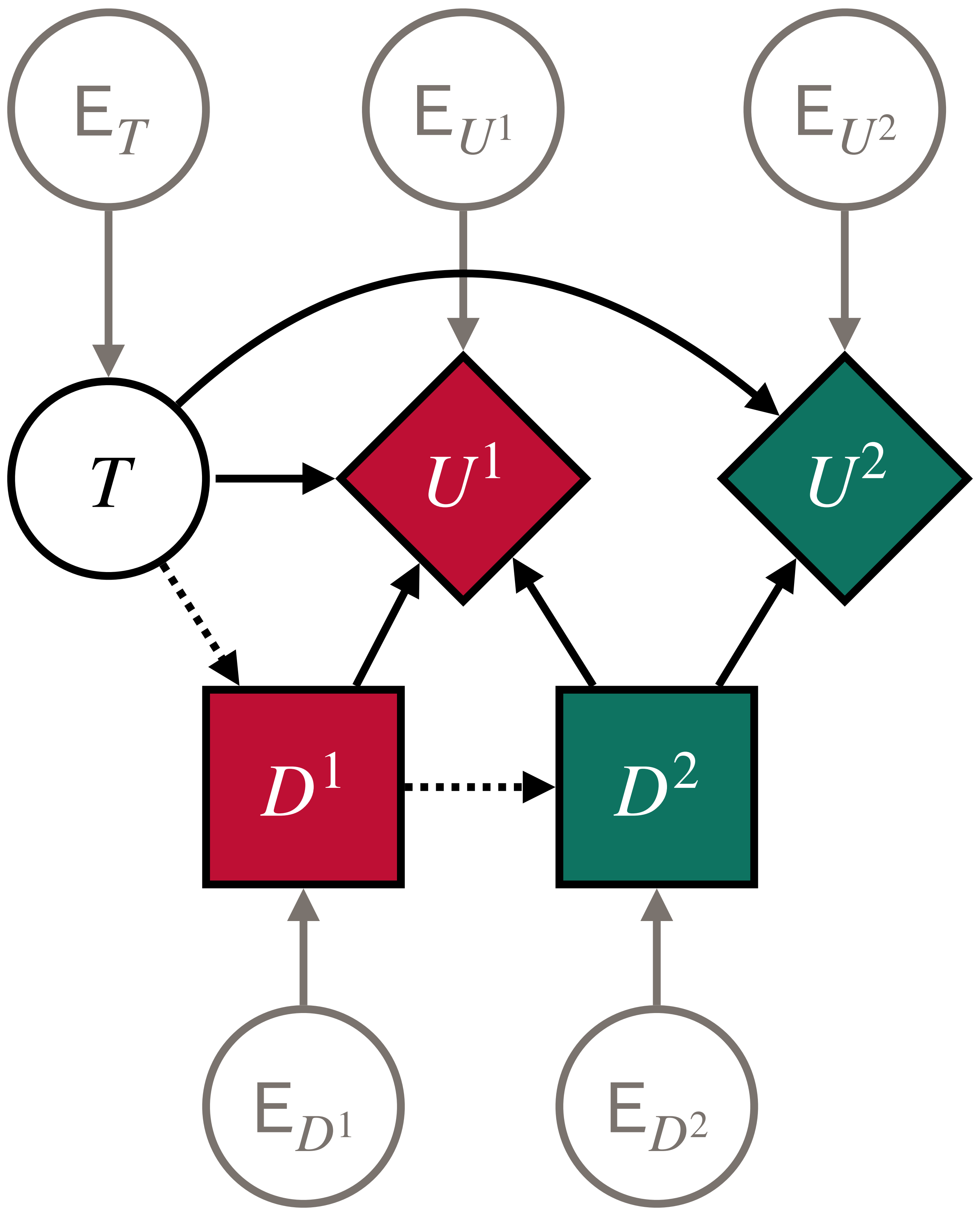


Counterfactuals



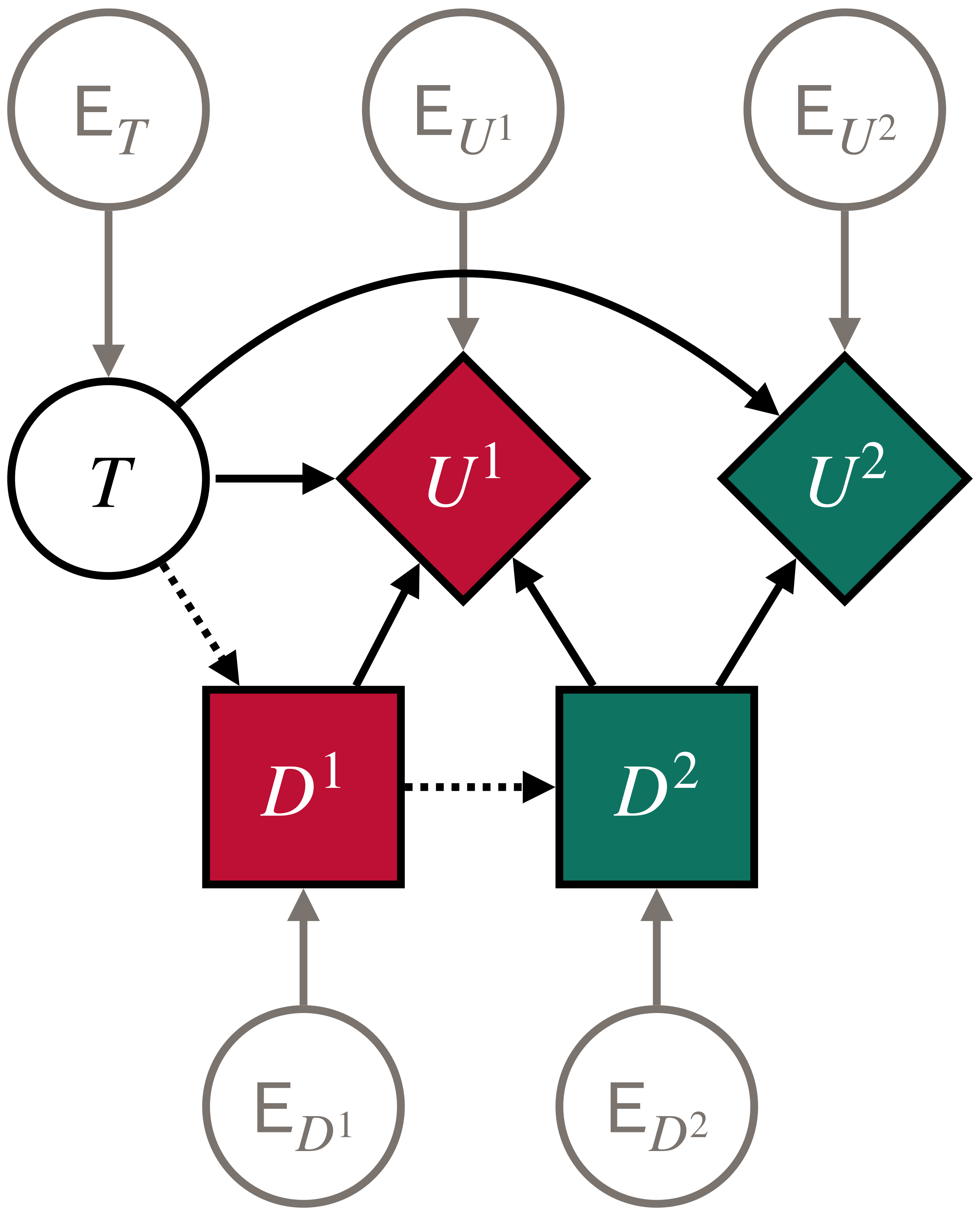
Counterfactuals

- We therefore let $\mathsf{E}_D = (\mathsf{E}_D^1, \ldots, \mathsf{E}_D^m)$ where $m = |\mathrm{dom}(\mathbf{Pa}'_D)|$ and $\Pr^\pi(\mathsf{e}_D) = \prod_k \Pr^\pi(\mathsf{e}_D^k)$



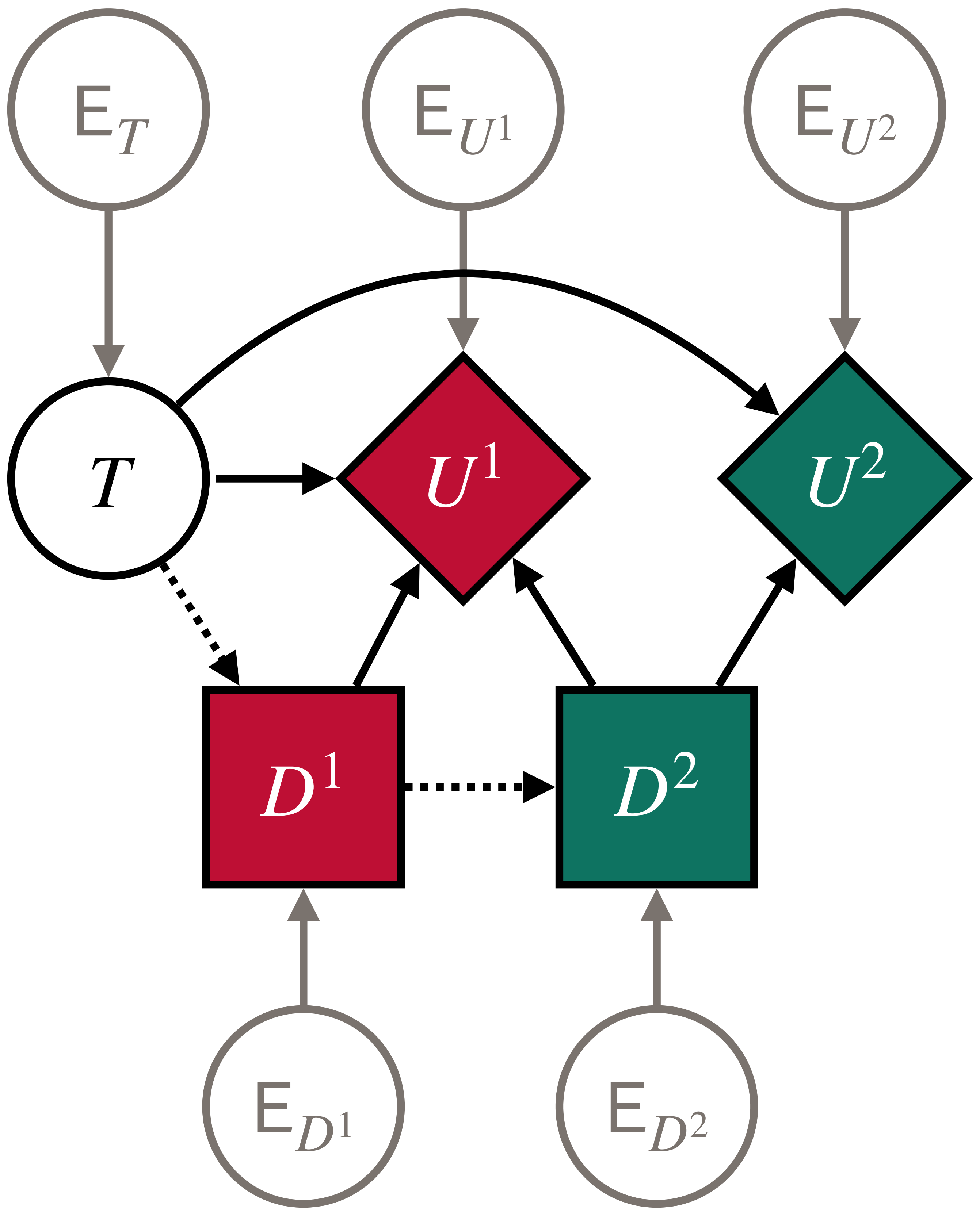
Counterfactuals

- We therefore let $\mathsf{E}_D = (\mathsf{E}_D^1, \ldots, \mathsf{E}_D^m)$ where $m = |\mathrm{dom}(\mathbf{Pa}'_D)|$ and $\Pr^\pi(\mathsf{e}_D) = \prod_k \Pr^\pi(\mathsf{e}_D^k)$
- One choice for a canonical structural representation is then given by:



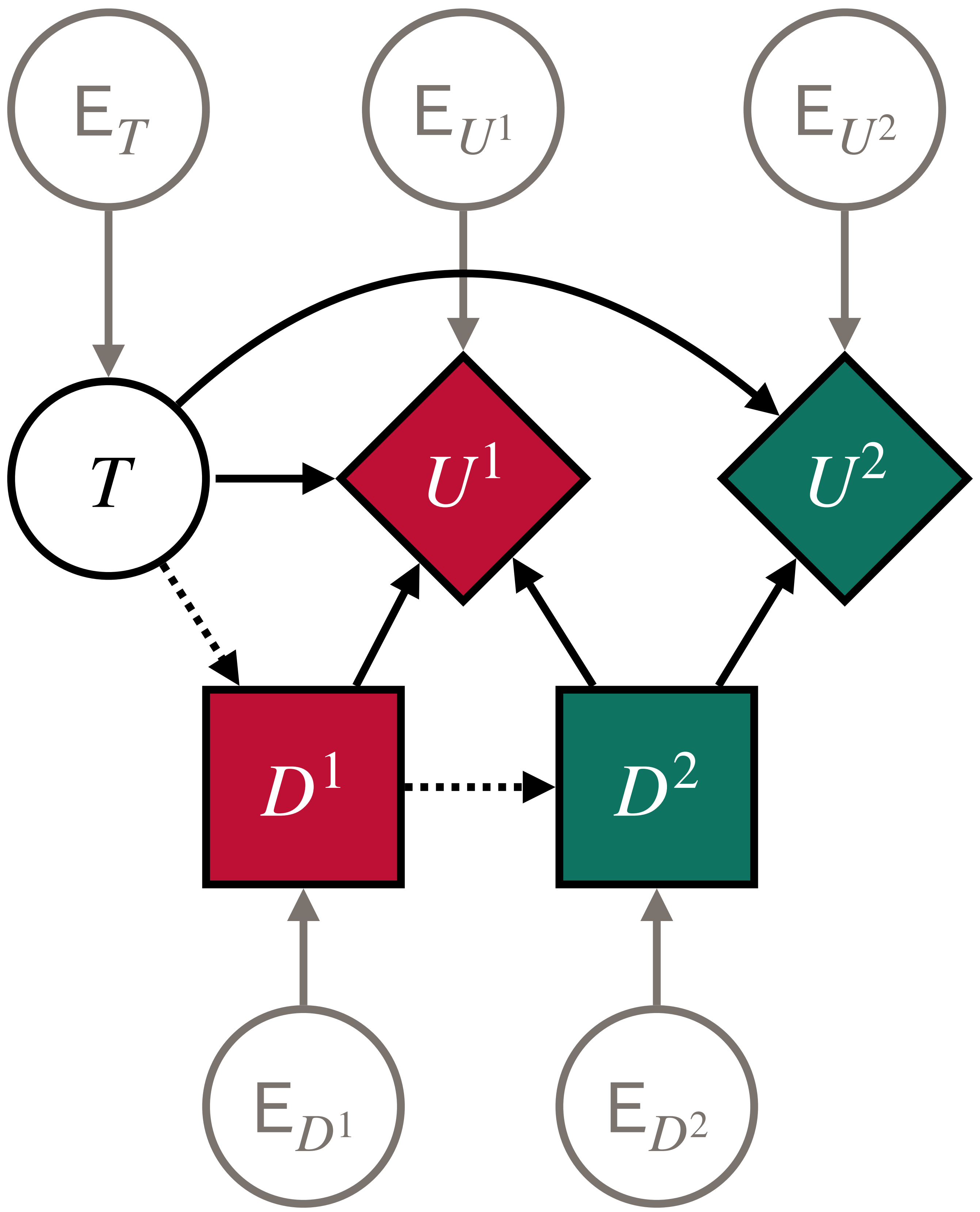
Counterfactuals

- We therefore let $\mathsf{E}_D = (\mathsf{E}^1_D, \ldots, \mathsf{E}^m_D)$ where $m = |\mathrm{dom}(\mathbf{Pa}'_D)|$ and $\Pr^\pi(\mathsf{e}_D) = \prod_k \Pr^\pi(\mathsf{e}^k_D)$
- One choice for a canonical structural representation is then given by:
  - $\Pr^\pi(\mathsf{e}^k_D) = \pi(d \mid \mathbf{pa}'^k_D)$



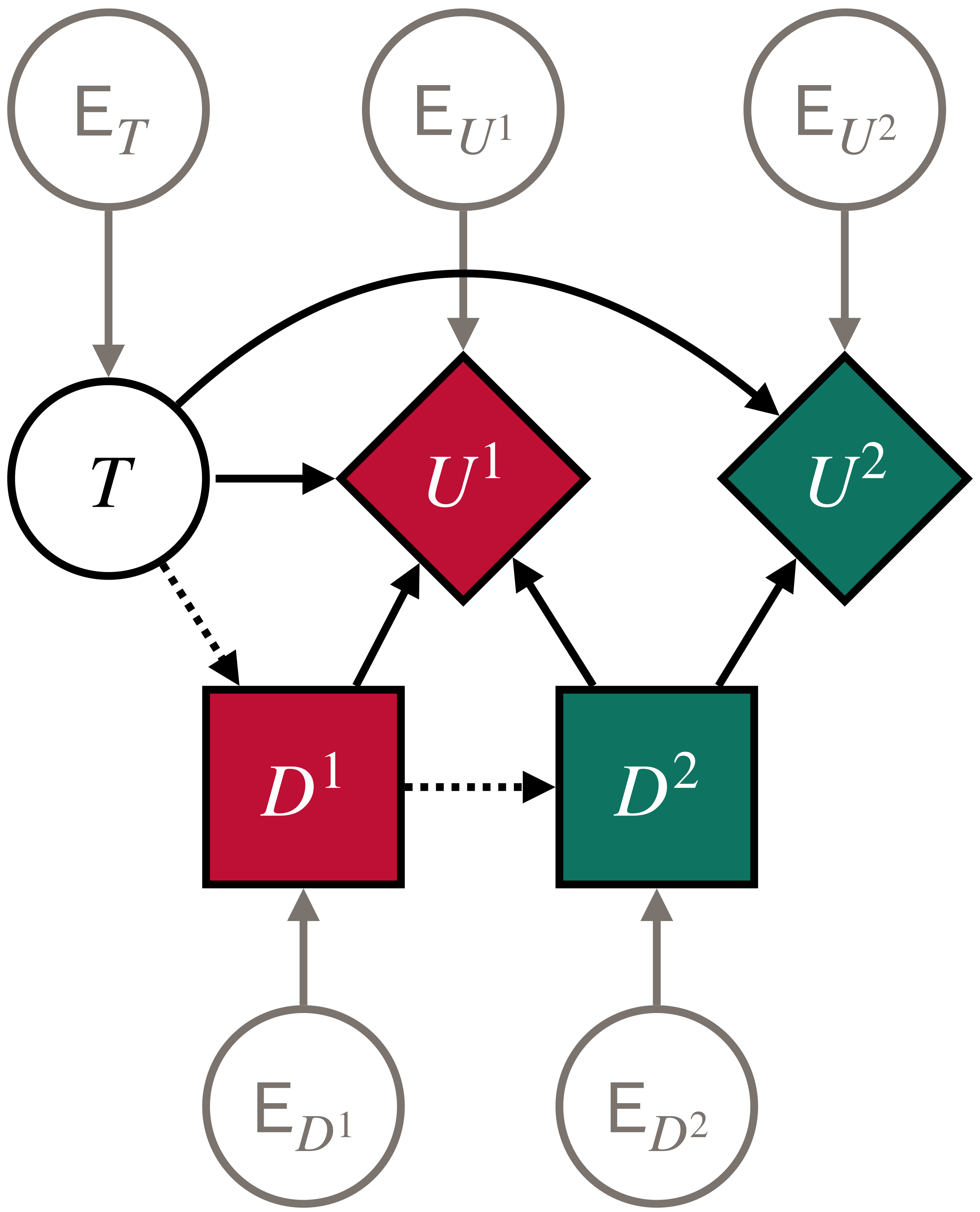
Counterfactuals

- We therefore let $\mathsf{E}_D = (\mathsf{E}_D^1, \ldots, \mathsf{E}_D^m)$ where $m = |\mathrm{dom}(\mathbf{Pa}'_D)|$ and $\Pr^\pi(\mathsf{e}_D) = \prod_k \Pr^\pi(\mathsf{e}_D^k)$
- One choice for a canonical structural representation is then given by:
  - $\Pr^\pi(\mathsf{e}_D^k) = \pi(d \mid \mathbf{pa}'^k_D)$
  - $\Pr^\pi(d \mid \mathbf{pa}'^k_D, \mathsf{e}_D) = \delta(d = e_D^k)$



Counterfactuals

- We therefore let $\mathsf{E}_D = (\mathsf{E}_D^1, \ldots, \mathsf{E}_D^m)$ where $m = |\mathrm{dom}(\mathbf{Pa}'_D)|$ and $\Pr^\pi(\mathsf{e}_D) = \prod_k \Pr^\pi(\mathsf{e}_D^k)$
- One choice for a canonical structural representation is then given by:
  - $\Pr^\pi(\mathsf{e}_D^k) = \pi(d \mid \mathbf{pa}'^k_D)$
  - $\Pr^\pi(d \mid \mathbf{pa}'^k_D, \mathsf{e}_D) = \delta(d = e_D^k)$
- In extended MASCIMs, we *merge* the mechanism variables for D and E_D into a single decision rule variable Π_D



Counterfactuals

Counterfactuals

- <u>Query:</u> Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier

Counterfactuals

- <u>Query:</u> Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier
- The basic idea:

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
    - This turns out to be a *lot* trickier
- The basic idea:
    - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$

Counterfactuals

- <u>Query:</u> Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
    - This turns out to be a *lot* trickier
- The basic idea:
    - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$
    - Apply a modified version of Pearl's three step procedure, where decision rules *not* in $\Pi(\mathbf{y})$ are consistent with $\mathbf{z}$

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier
- The basic idea:
  - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$
  - Apply a modified version of Pearl's three step procedure, where decision rules *not* in $\Pi(\mathbf{y})$ are consistent with $\mathbf{z}$
- Three step procedure

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier
- The basic idea:
  - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$
  - Apply a modified version of Pearl's three step procedure, where decision rules *not* in $\Pi(\mathbf{y})$ are consistent with $\mathbf{z}$

- Three step procedure
  1. For $\pi' \in \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z})$ update $\Pr^{\pi}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \leftarrow \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z})$ where $\mathbf{D}(\mathbf{y}) = \left\{D : \Pi_D \in \Pi(\mathbf{y})\right\}$

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier
- The basic idea:
  - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$
  - Apply a modified version of Pearl's three step procedure, where decision rules *not* in $\Pi(\mathbf{y})$ are consistent with $\mathbf{z}$

- Three step procedure
  1. For $\pi' \in \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z})$ update $\Pr^{\pi}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \leftarrow \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z})$ where $\mathbf{D}(\mathbf{y}) = \{D : \Pi_D \in \Pi(\mathbf{y})\}$
  2. Intervene on $\mathbf{Y} \cap \mathbf{M}$ to find each $\pi \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}})$ such that $\pi_{-\mathbf{D}(\mathbf{y})} = \pi'_{-\mathbf{D}(\mathbf{y})}$ then intervene on $\mathbf{Y} \cap \mathbf{V}$ as normal

Counterfactuals

- Query: Given an observation $\mathbf{z}$, what would be the probability of $\mathbf{x}$ if we had $\mathbf{Y} \leftarrow \mathbf{y}$?
  - This turns out to be a *lot* trickier
- The basic idea:
  - Find the set of variables $\Pi(\mathbf{y}) \subseteq \Pi$ that are affected by $\mathbf{Y} \leftarrow \mathbf{y}$
  - Apply a modified version of Pearl's three step procedure, where decision rules *not* in $\Pi(\mathbf{y})$ are consistent with $\mathbf{z}$

- Three step procedure
  1. For $\pi' \in \mathcal{R}(\mathsf{x}\mathcal{M} \mid \mathbf{z})$ update $\Pr^{\pi}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \leftarrow \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z})$ where $\mathbf{D}(\mathbf{y}) = \{D : \Pi_D \in \Pi(\mathbf{y})\}$
  2. Intervene on $\mathbf{Y} \cap \mathbf{M}$ to find each $\pi \in \mathcal{R}(\mathsf{x}\mathcal{M}_{\mathbf{y}})$ such that $\pi_{-\mathbf{D}(\mathbf{y})} = \pi'_{-\mathbf{D}(\mathbf{y})}$ then intervene on $\mathbf{Y} \cap \mathbf{V}$ as normal
  3. Return the updated distribution $\Pr^{\pi}(\mathbf{x})$ for each π

Counterfactuals

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x_y} \mid \mathbf{e_{D(y)}}, \mathbf{e_{-D(y)}}) \Pr^{\pi}(\mathbf{e_{D(y)}}) \Pr^{\pi'}(\mathbf{e_{-D(y)}} \mid \mathbf{z}) \right\}_{(\pi,\pi')\in\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}}|\mathbf{z})}$$

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x}_{\mathbf{y}} \mid \mathbf{e}_{\mathbf{D}(\mathbf{y})}, \mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \Pr^{\pi}(\mathbf{e}_{\mathbf{D}(\mathbf{y})}) \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z}) \right\}_{(\pi,\pi') \in \mathscr{R}(\mathsf{x}\mathscr{M}_{\mathbf{y}} | \mathbf{z})}$$

- For each counterfactual-actual rational outcome $(\pi, \pi') \in \mathscr{R}(\mathsf{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})$ where:

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x_y} \mid \mathbf{e}_{\mathbf{D(y)}}, \mathbf{e}_{-\mathbf{D(y)}}) \Pr^{\pi}(\mathbf{e}_{\mathbf{D(y)}}) \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D(y)}} \mid \mathbf{z}) \right\}_{(\pi,\pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})}$$

- For each counterfactual-actual rational outcome $(\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})$ where:

$$\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z}) := \left\{(\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}}) \times \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z}) : \pi_{-\mathbf{D(y)}} = \pi'_{-\mathbf{D(y)}}\right\}$$

Counterfactuals

- More concretely, the <u>answer to a counterfactual query</u> that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x}_{\mathbf{y}} \mid \mathbf{e}_{\mathbf{D(y)}}, \mathbf{e}_{-\mathbf{D(y)}}) \Pr^{\pi}(\mathbf{e}_{\mathbf{D(y)}}) \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D(y)}} \mid \mathbf{z}) \right\}_{(\pi,\pi') \in \mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}} \mid \mathbf{z})}$$

- For each <u>counterfactual-actual rational outcome</u> $(\pi, \pi') \in \mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}} \mid \mathbf{z})$ where:

$$\mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}} \mid \mathbf{z}) := \left\{ (\pi, \pi') \in \mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}}) \times \mathcal{R}(\mathrm{x}\mathcal{M} \mid \mathbf{z}) : \pi_{-\mathbf{D(y)}} = \pi'_{-\mathbf{D(y)}} \right\}$$

- <u>Counterfactual joint policies</u> π are members of $\mathcal{R}(\mathrm{x}\mathcal{M}_{\mathbf{y}})$ such that π_D is consistent with the observation $\mathbf{z}$ whenever Π_D is *not* affected by $\mathbf{Y} \leftarrow \mathbf{y}$, i.e. $\Pi_D \notin \Pi(\mathbf{y})$

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x_y} \mid \mathbf{e}_{\mathbf{D}(\mathbf{y})}, \mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \Pr^{\pi}(\mathbf{e}_{\mathbf{D}(\mathbf{y})}) \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z}) \right\}_{(\pi,\pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})}$$

- For each counterfactual-actual rational outcome $(\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})$ where:

$$\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z}) := \left\{(\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}}) \times \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z}) : \pi_{-\mathbf{D}(\mathbf{y})} = \pi'_{-\mathbf{D}(\mathbf{y})}\right\}$$

- Counterfactual joint policies π are members of $\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}})$ such that π_D is consistent with the observation $\mathbf{z}$ whenever Π_D is *not* affected by $\mathbf{Y} \leftarrow \mathbf{y}$, i.e. $\Pi_D \notin \Pi(\mathbf{y})$
  - We then sample from $\Pr^{\pi}(\mathbf{e}_{\mathbf{D}(\mathbf{y})})$ according to the new joint policy

Counterfactuals

- More concretely, the answer to a counterfactual query that we return is:

$$\left\{ \sum_{\mathbf{e}} \Pr^{\pi}(\mathbf{x_y} \mid \mathbf{e}_{\mathbf{D}(\mathbf{y})}, \mathbf{e}_{-\mathbf{D}(\mathbf{y})}) \Pr^{\pi}(\mathbf{e}_{\mathbf{D}(\mathbf{y})}) \Pr^{\pi'}(\mathbf{e}_{-\mathbf{D}(\mathbf{y})} \mid \mathbf{z}) \right\}_{(\pi,\pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})}$$

- For each counterfactual-actual rational outcome $(\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z})$ where:

$$\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}} \mid \mathbf{z}) := \left\{ (\pi, \pi') \in \mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}}) \times \mathscr{R}(\mathrm{x}\mathscr{M} \mid \mathbf{z}) : \pi_{-\mathbf{D}(\mathbf{y})} = \pi'_{-\mathbf{D}(\mathbf{y})} \right\}$$

- Counterfactual joint policies π are members of $\mathscr{R}(\mathrm{x}\mathscr{M}_{\mathbf{y}})$ such that π_D is consistent with the observation $\mathbf{z}$ whenever Π_D is *not* affected by $\mathbf{Y} \leftarrow \mathbf{y}$, i.e. $\Pi_D \notin \Pi(\mathbf{y})$
  - We then sample from $\Pr^{\pi}(\mathbf{e}_{\mathbf{D}(\mathbf{y})})$ according to the new joint policy
- But when we learn about $\mathbf{e}_{-\mathbf{D}(\mathbf{y})}$ based on $\mathbf{z}$ we do so under the actual joint policy π'

Counterfactuals

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?

Counterfactuals

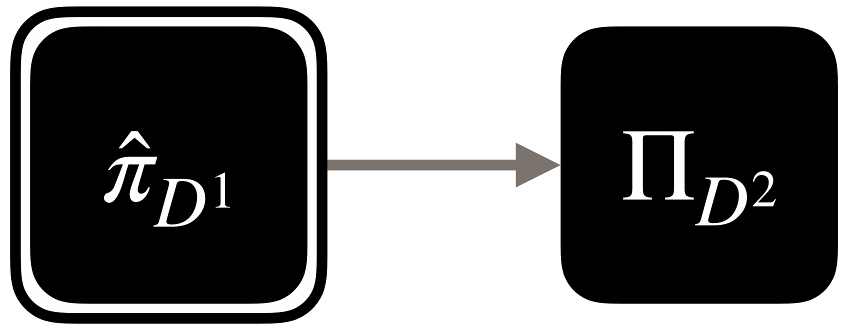
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:

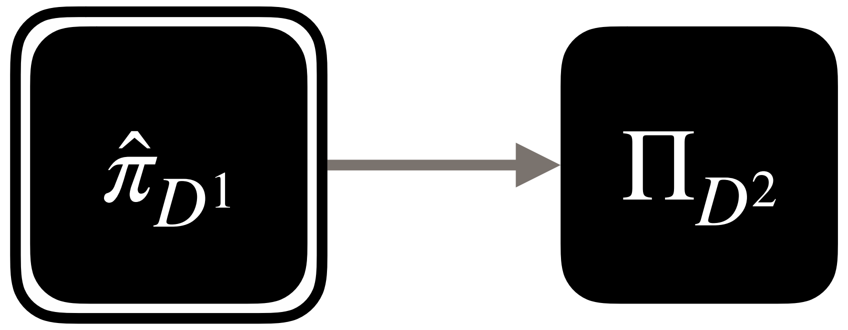
Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



Counterfactuals

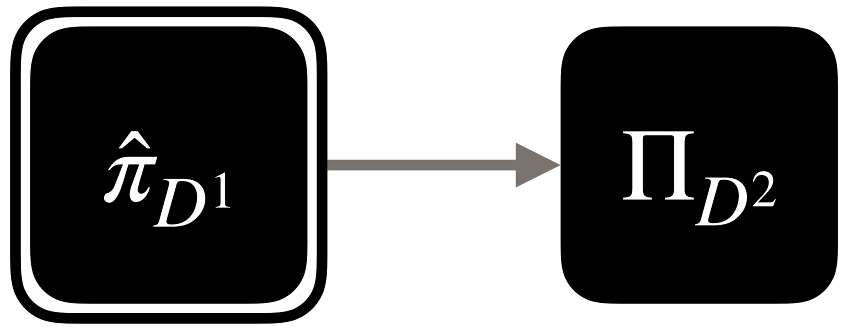
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$
  - Counterexample:

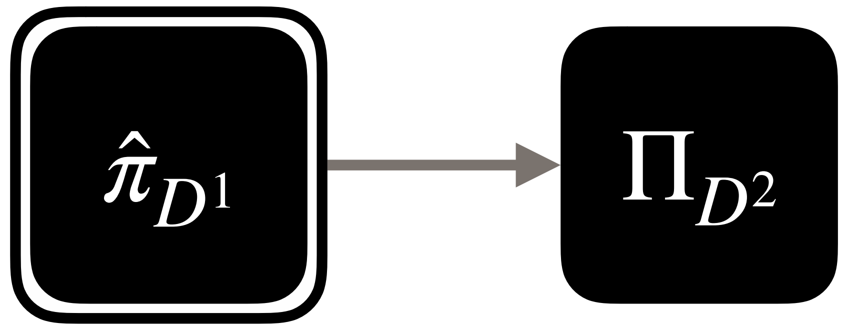


$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{ r_{D^2}(\pi_{D^1}) \right\}_{\pi_{D^1} \in r_{D^1}()}$$

- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



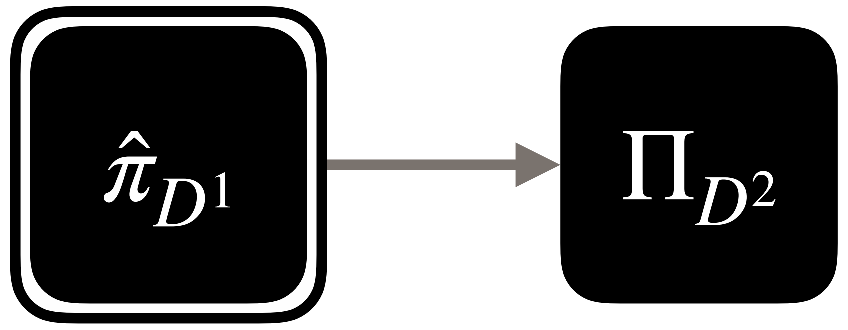
$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



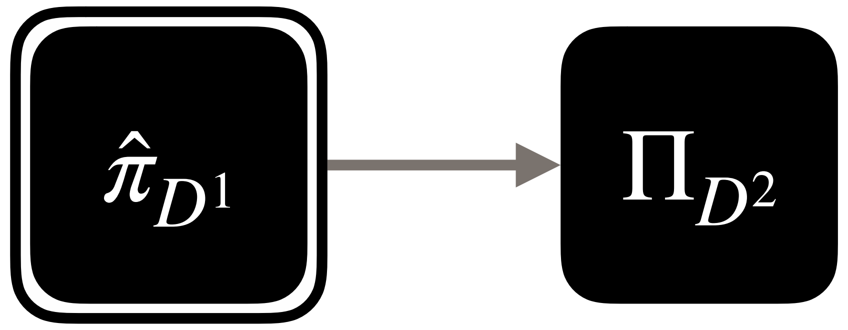
$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?
  - We can have cycles between decision rule variables!

Counterfactuals

- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$
  - Counterexample:

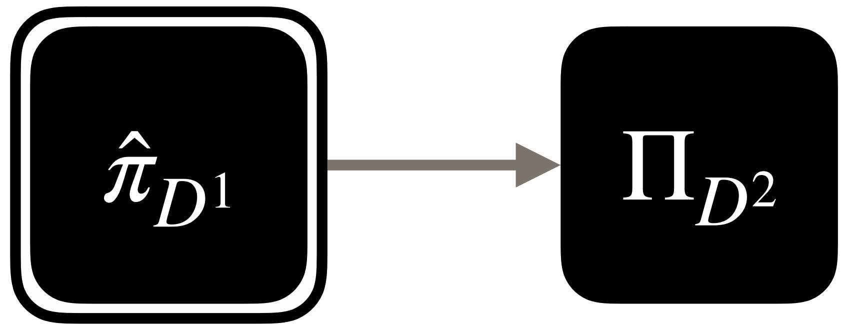


$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{ r_{D^2}(\pi_{D^1}) \right\}_{\pi_{D^1} \in r_{D^1}()}$$

- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$
- What's the problem with this?
  - We can have cycles between decision rule variables!
- Solution: form the condensation

Counterfactuals

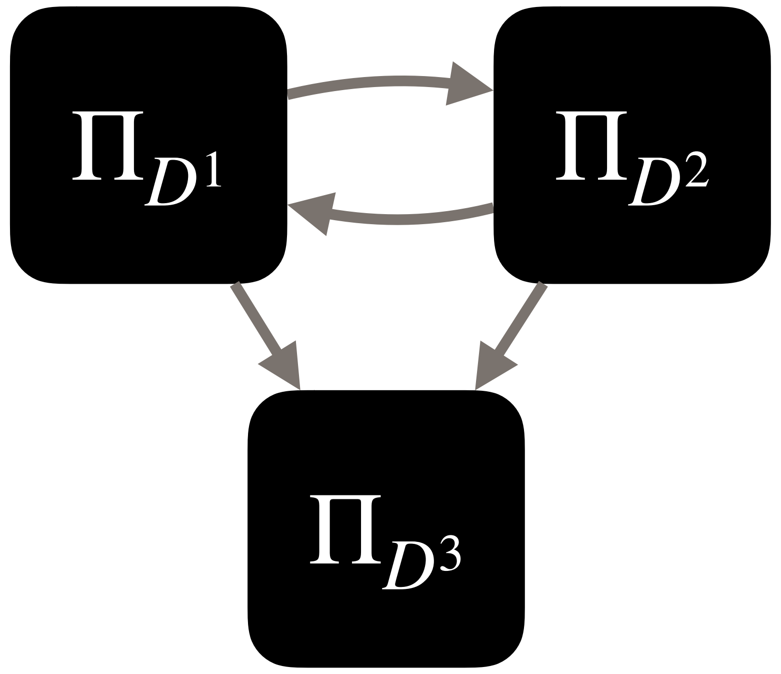
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{ r_{D^2}(\pi_{D^1}) \right\}_{\pi_{D^1} \in r_{D^1}()}$$

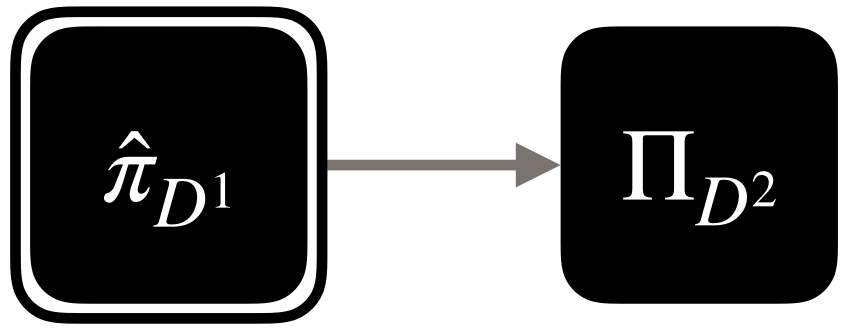
- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?
  - We can have cycles between decision rule variables!
- Solution: form the condensation



Counterfactuals

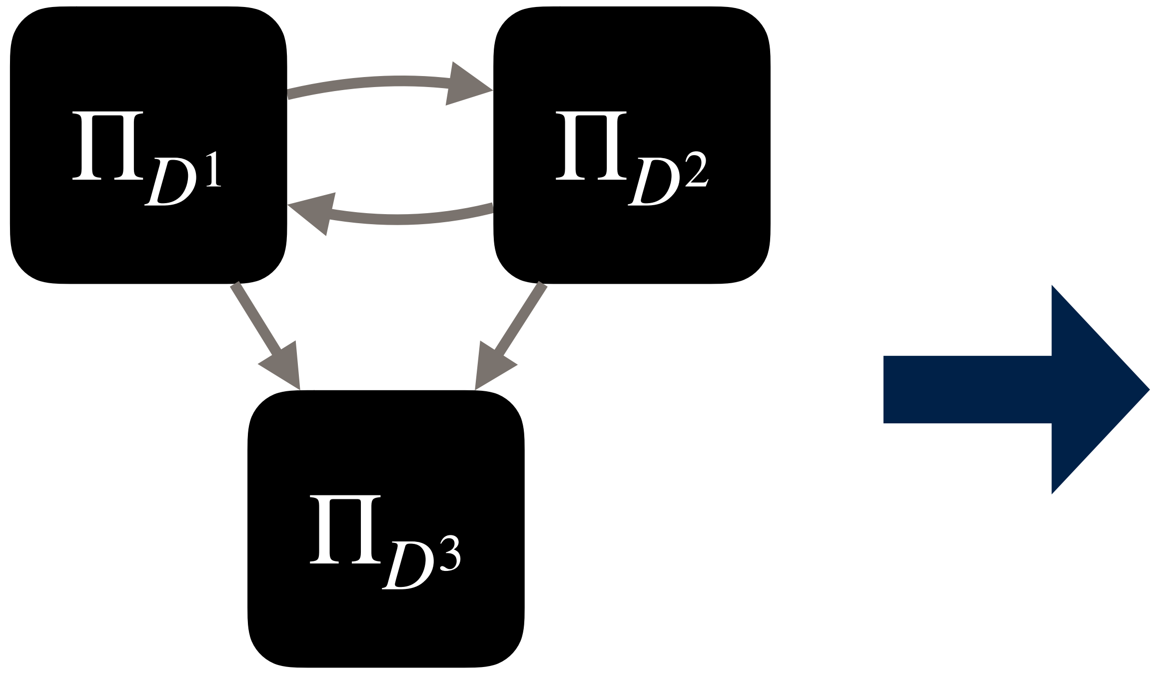
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$
  - Counterexample:



$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

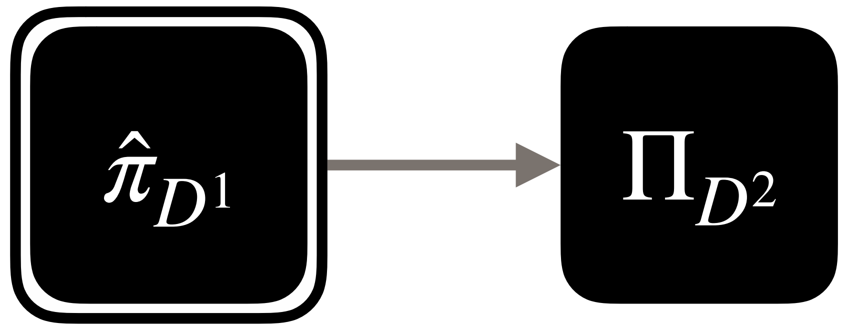
- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?
  - We can have cycles between decision rule variables!
- Solution: form the condensation



Counterfactuals

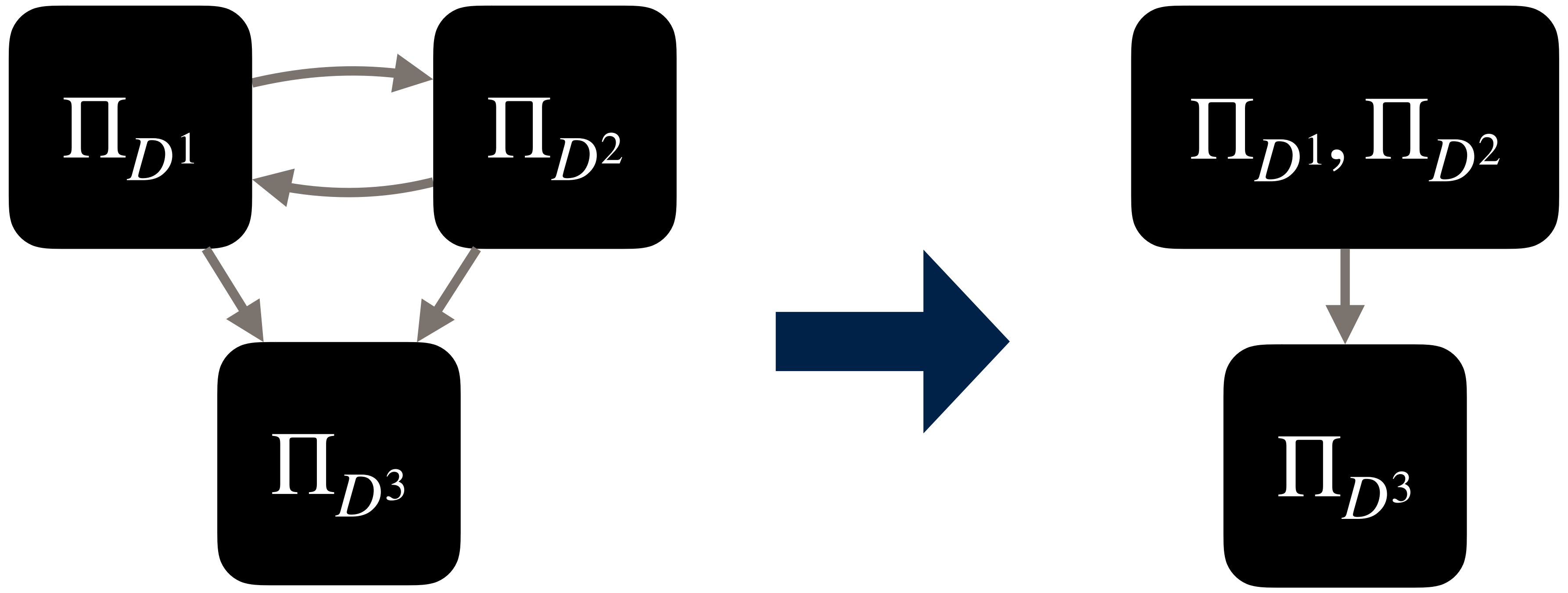
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$
  - Counterexample:



$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

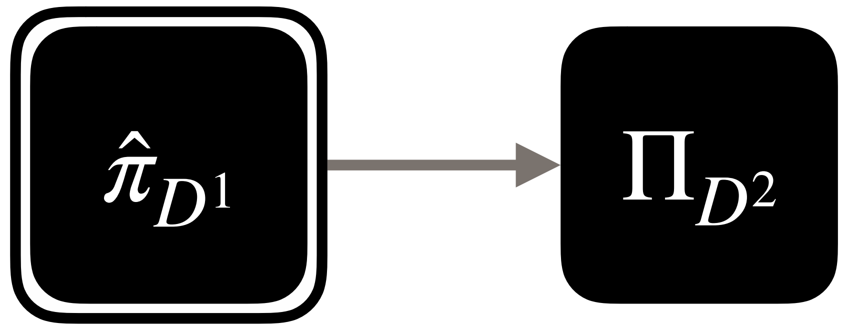
- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc_Y}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?
  - We can have cycles between decision rule variables!
- Solution: form the condensation



Counterfactuals

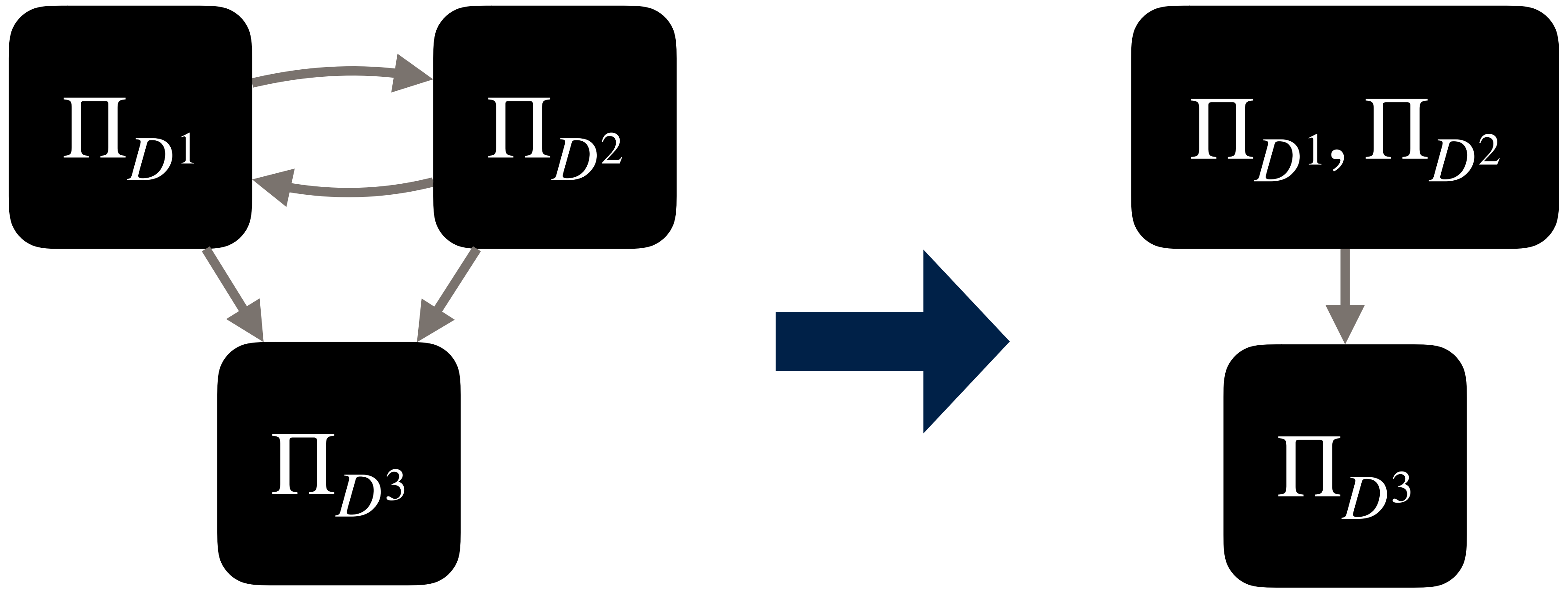
- How do we find $\Pi(\mathbf{y})$?
- First attempt: $\Pi(\mathbf{y}) := \Pi \cap \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$
  - Counterexample:



$$r_{D^2}(\hat{\pi}_{D^1}) = \left\{r_{D^2}(\pi_{D^1})\right\}_{\pi_{D^1} \in r_{D^1}()}$$

- Instead, we let $\Pi \backslash \left(\mathbf{Y} \cup \mathbf{Desc}_{\mathbf{Y}}\right)$ be consistent with $\mathbf{z}$, then recursively compute $r_D(\mathbf{pa}_D)$ to find $\Pi(\mathbf{y})$

- What's the problem with this?
  - We can have cycles between decision rule variables!
- Solution: form the condensation



- We have $\Pi_D \in \Pi(\mathbf{y})$ if and only if the rational responses for Π_D are invariant

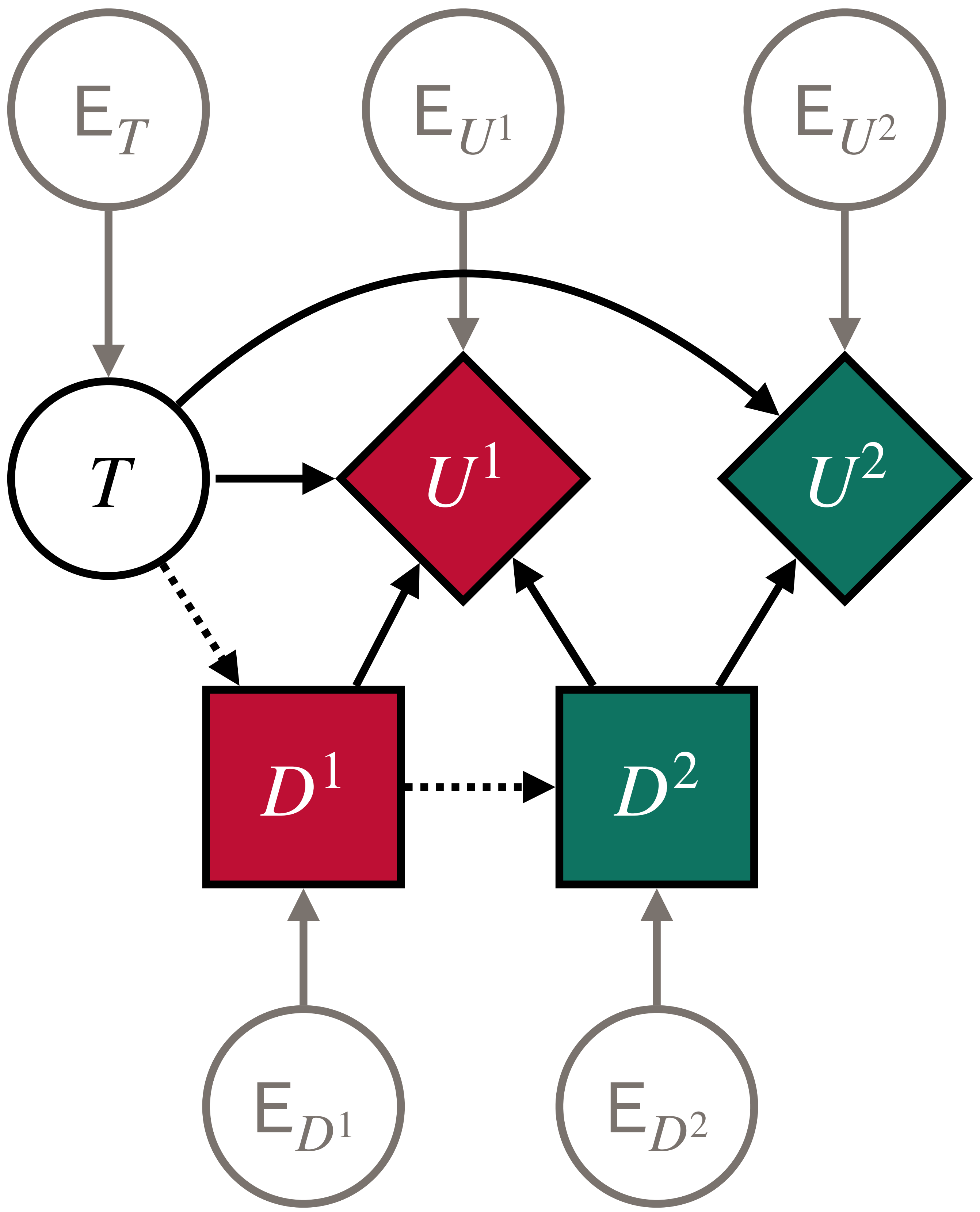
Counterfactuals

Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

Counterfactuals

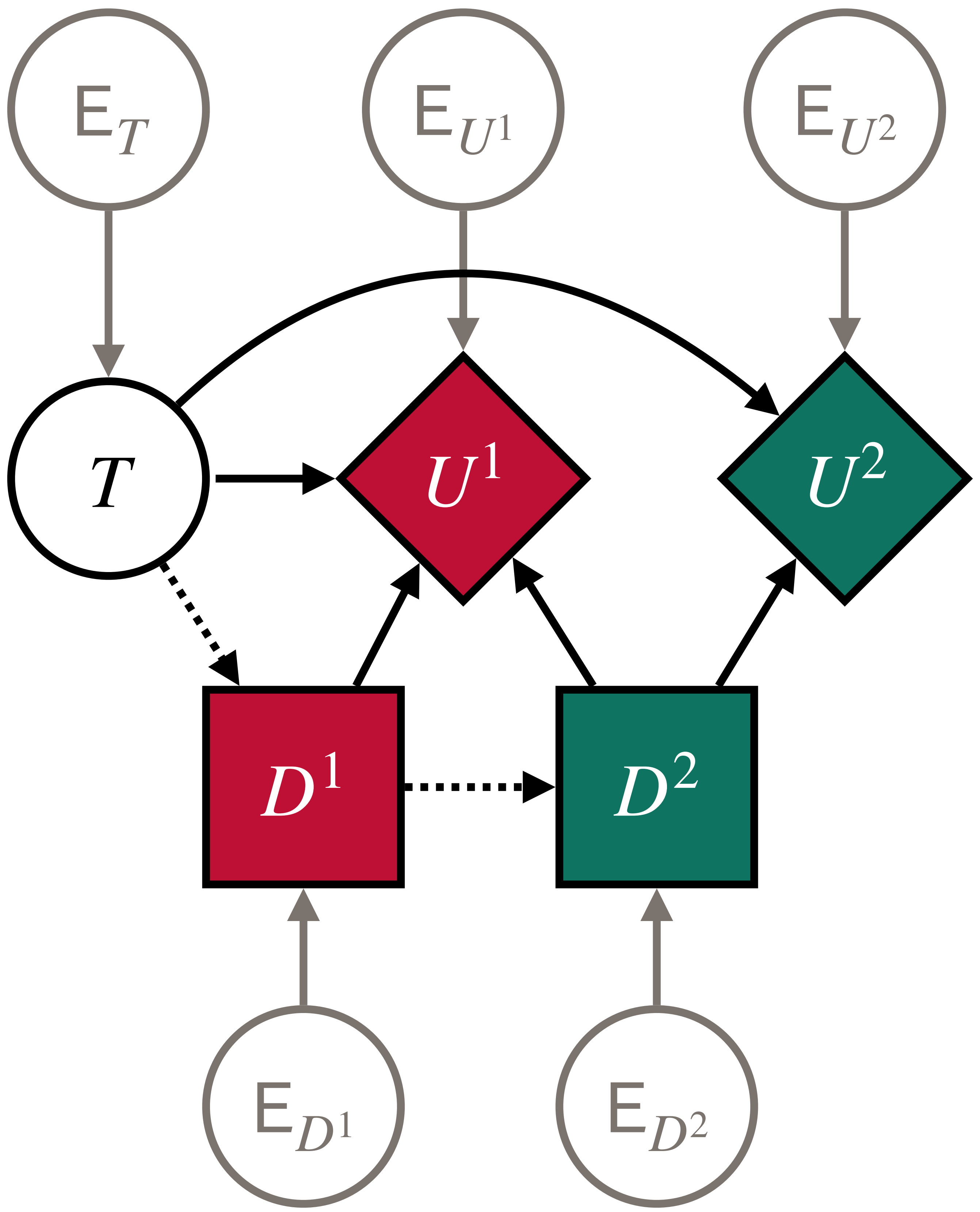
3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

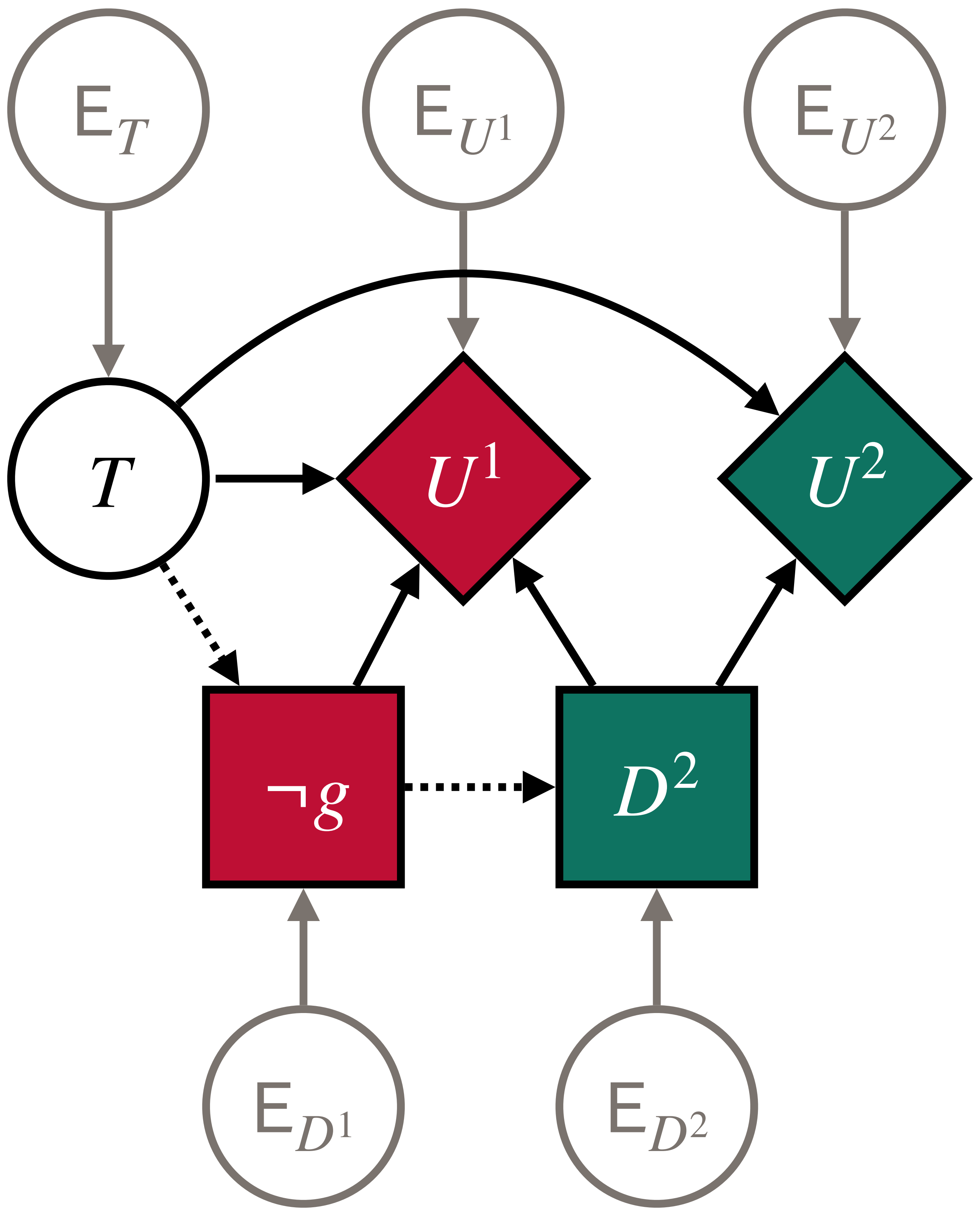
- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

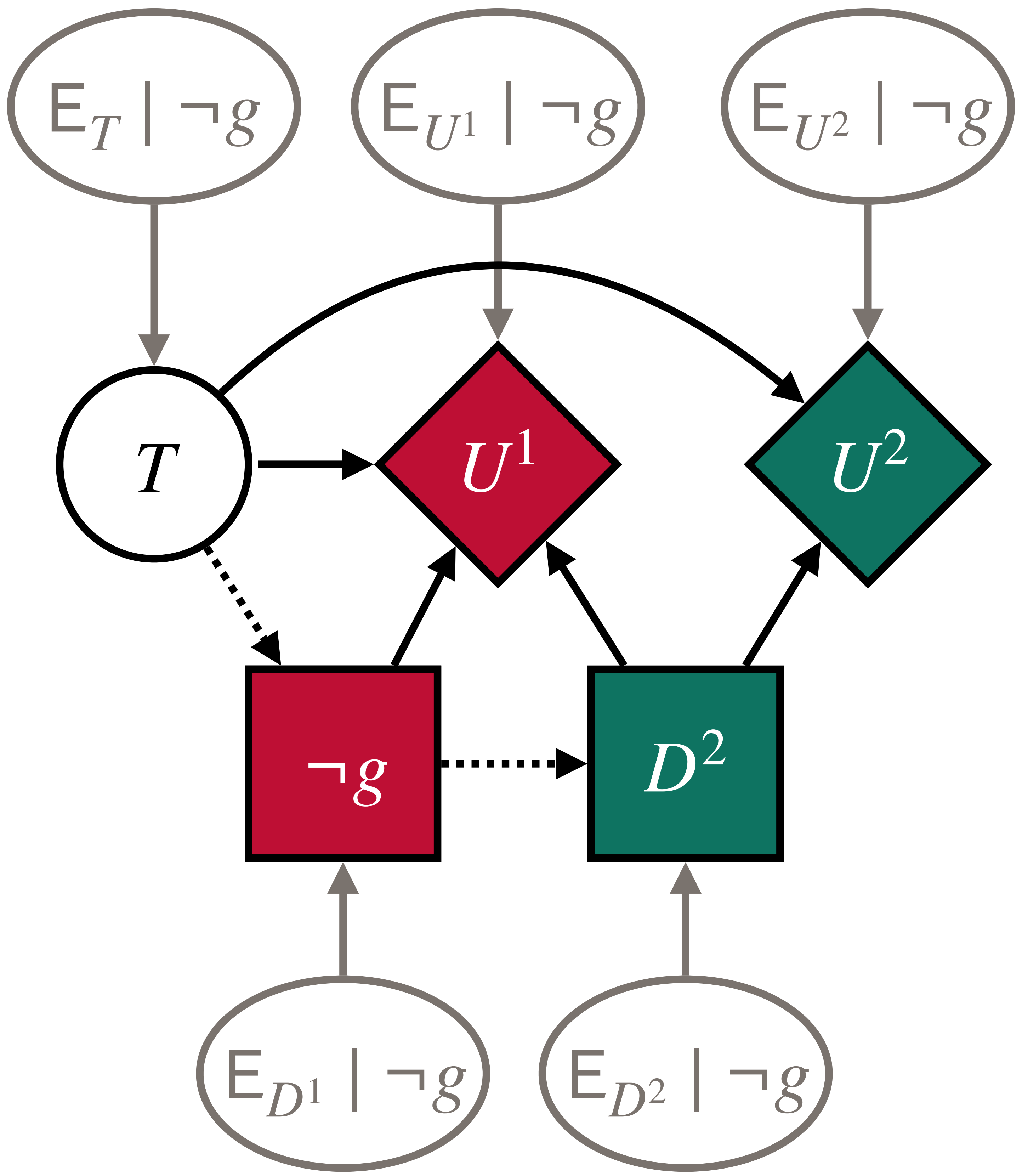
- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

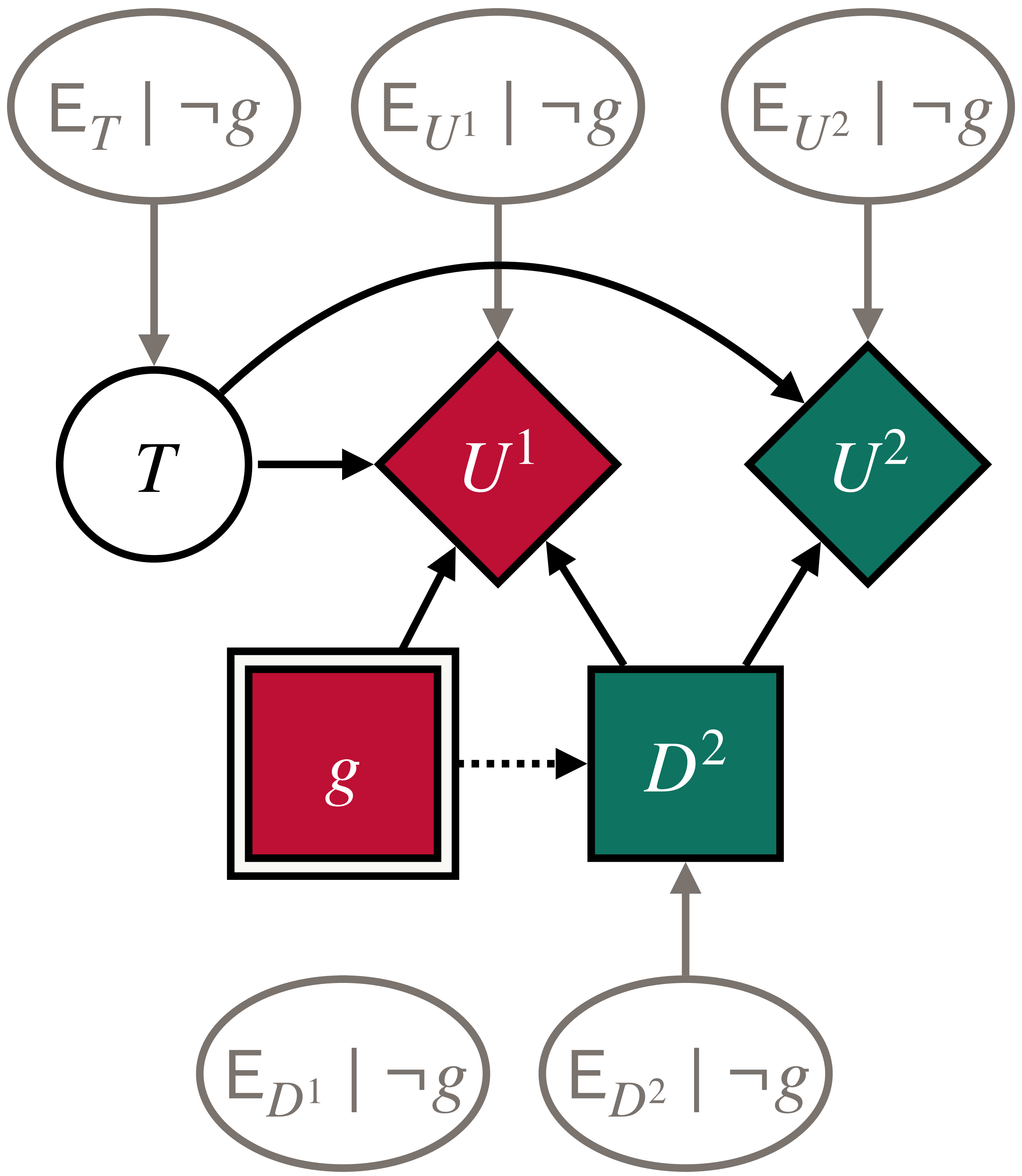
- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

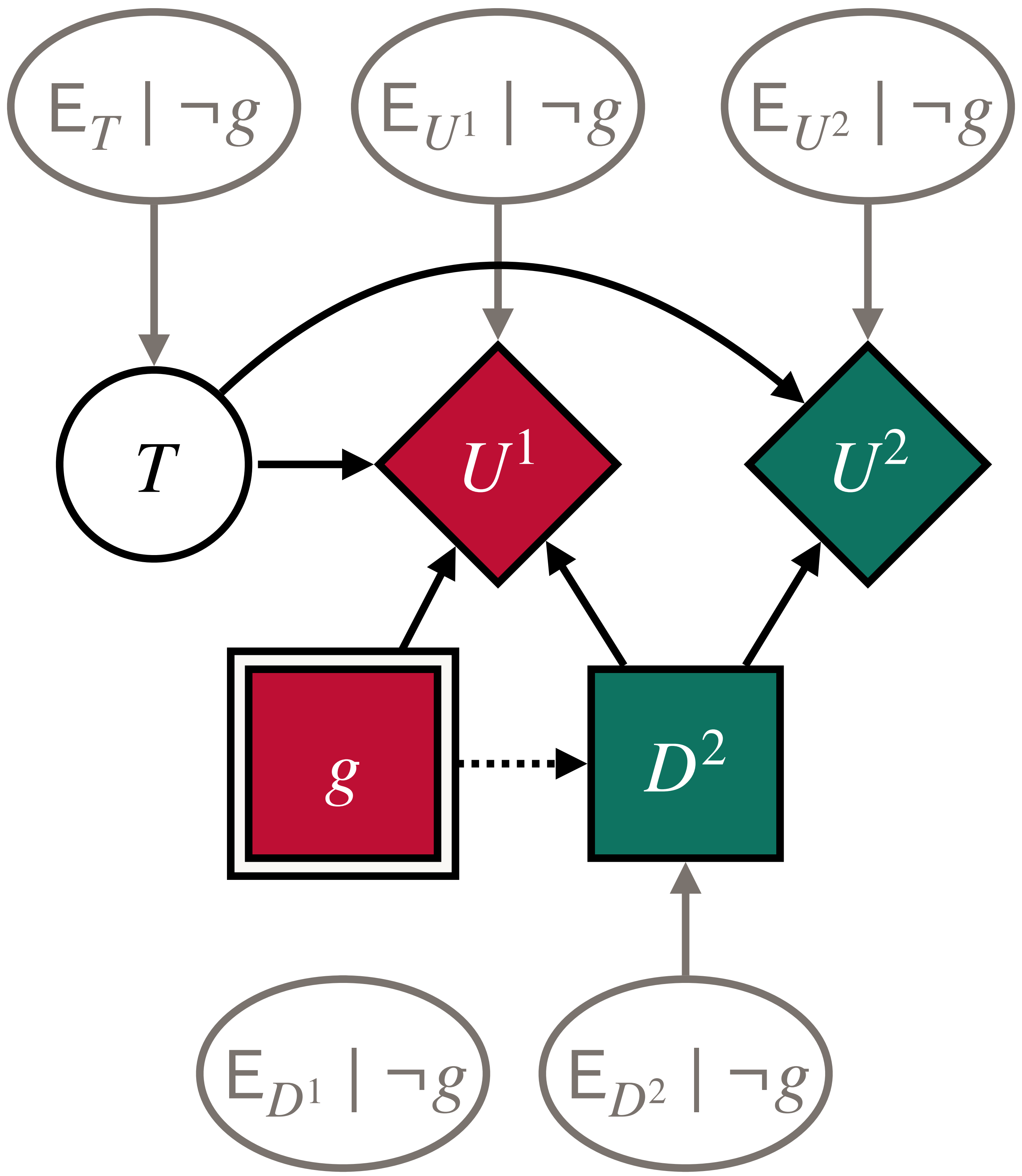
- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

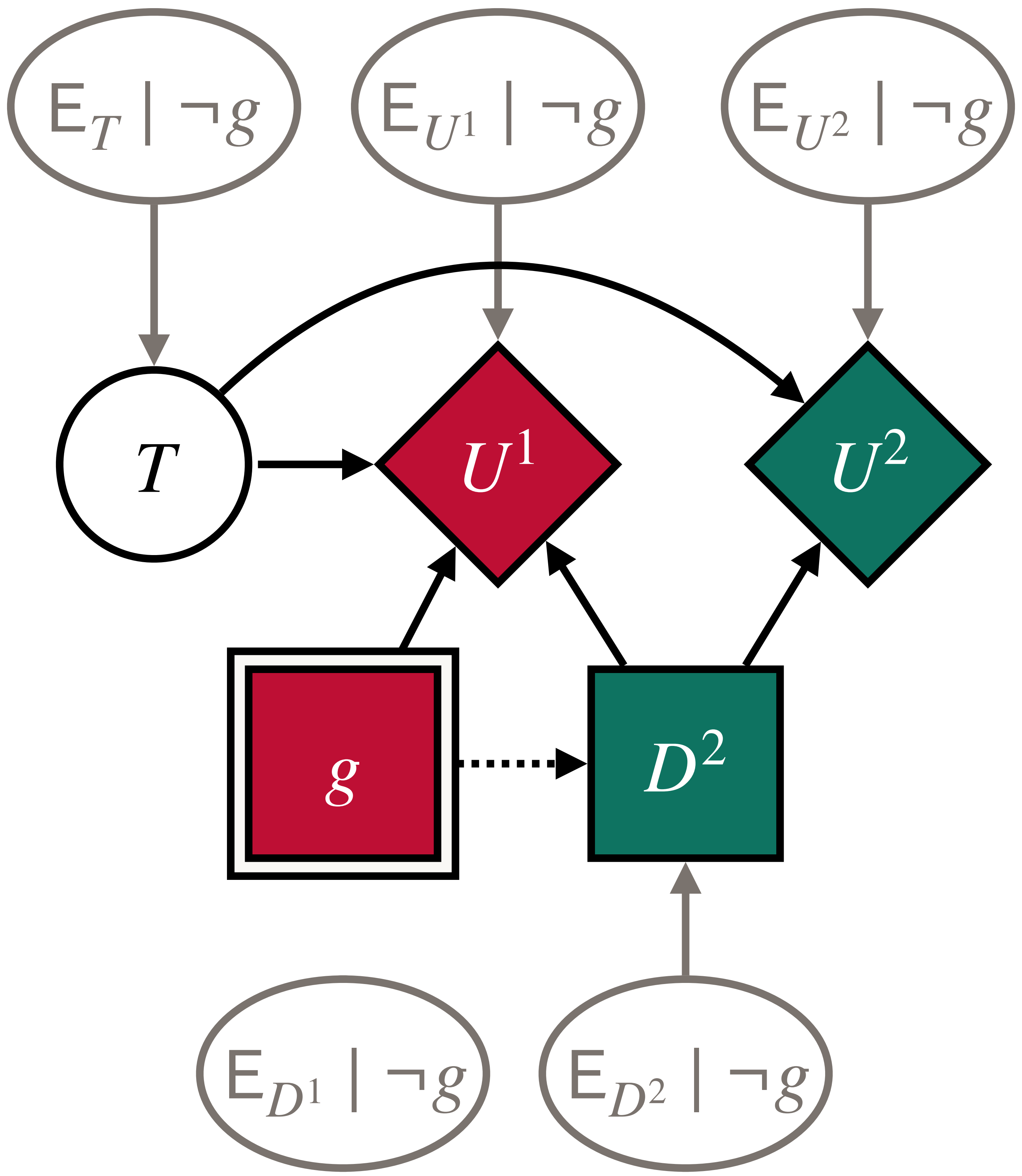
- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1
- As $\{D^1\} \cap \mathrm{M} = \emptyset$ then all the difficulties of the previous slides can be ignored



Counterfactuals

3.a) Given that the worker didn't go to university, what would be their wellbeing if they had?

- Observe $\neg g$, set $D^1 \leftarrow g$, and then predict u^1
- As $\{D^1\} \cap \mathsf{M} = \emptyset$ then all the difficulties of the previous slides can be ignored
- So just compute $\Pr^\pi(u^1_g \mid \neg g)$ for each $\pi \in \mathscr{R}(\mathsf{x}\mathscr{M} \mid \neg g)$



Additional Topics

Game-Theoretic Reasoning

Game-Theoretic Reasoning

- Part of the motivation for introducing these models is that they allow for both causal *and* game-theoretic reasoning

Game-Theoretic Reasoning

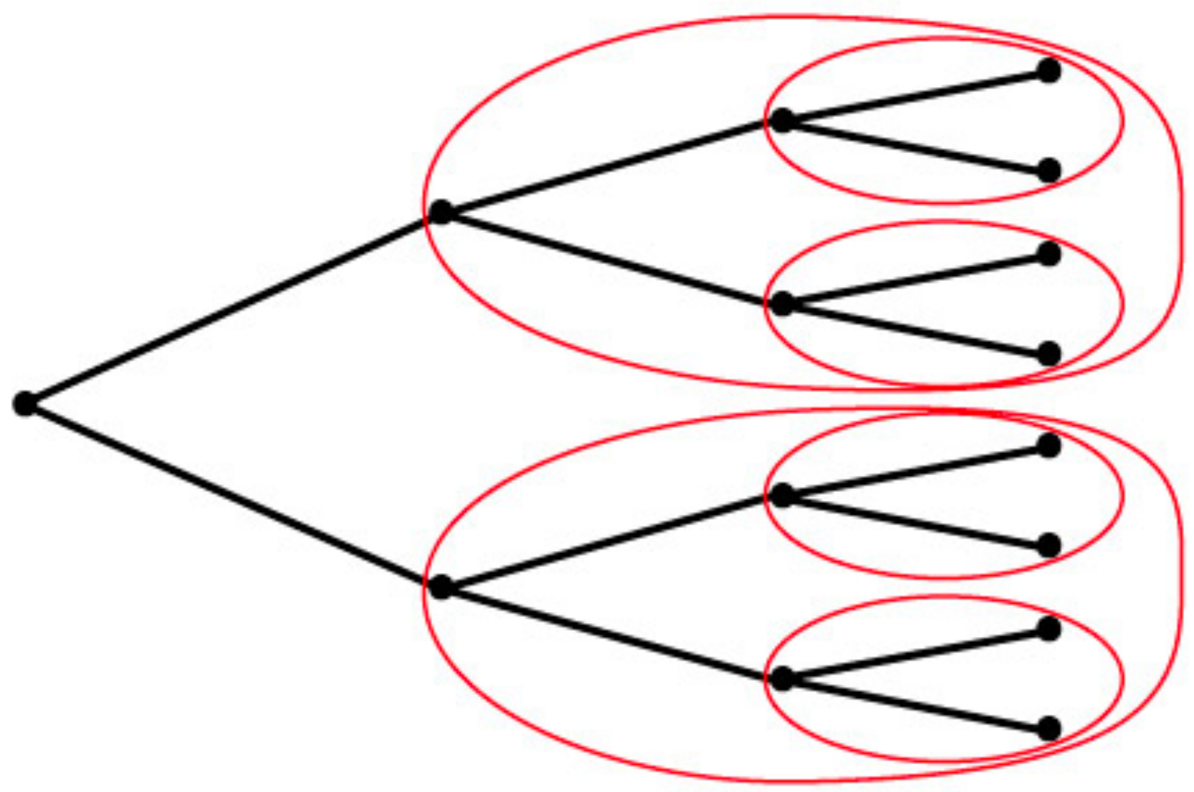
- Part of the motivation for introducing these models is that they allow for both causal *and* game-theoretic reasoning
- In earlier work [6], we study:

Game-Theoretic Reasoning

- Part of the motivation for introducing these models is that they allow for both causal *and* game-theoretic reasoning
- In earlier work [6], we study:
  - Equilibrium refinements (NE [10], SPE [15], THPE [14])

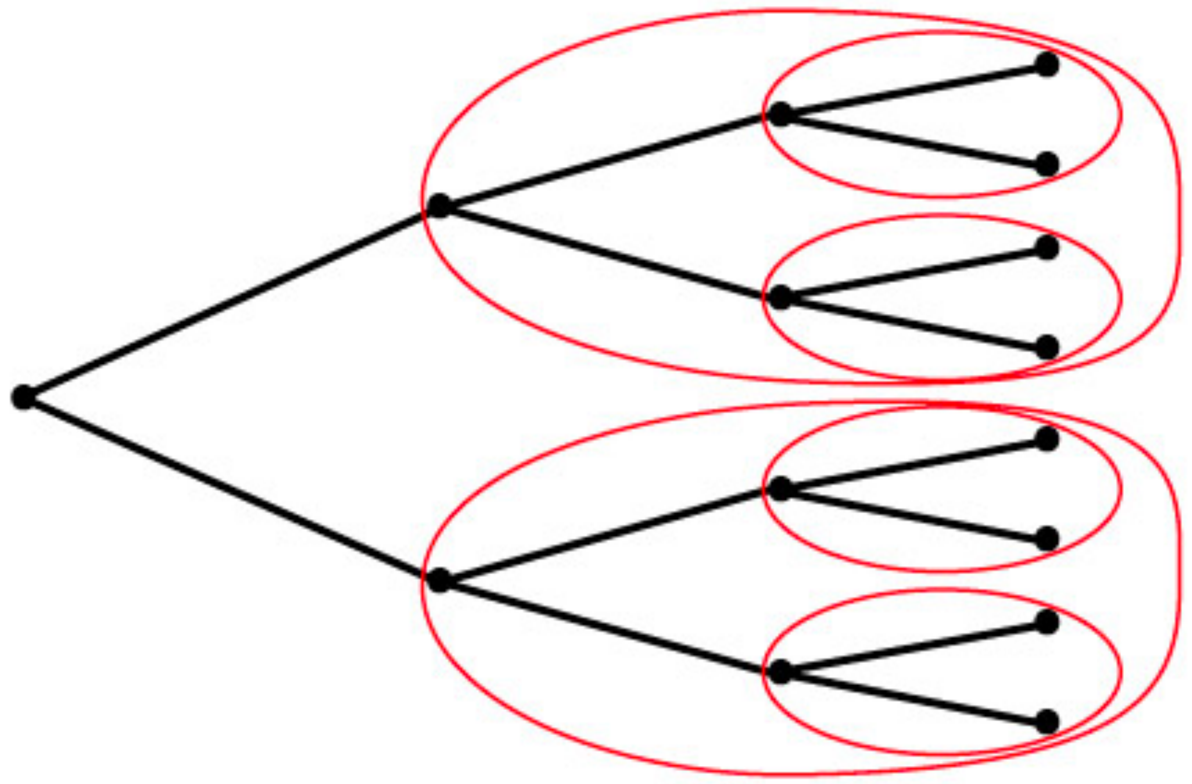
Game-Theoretic Reasoning

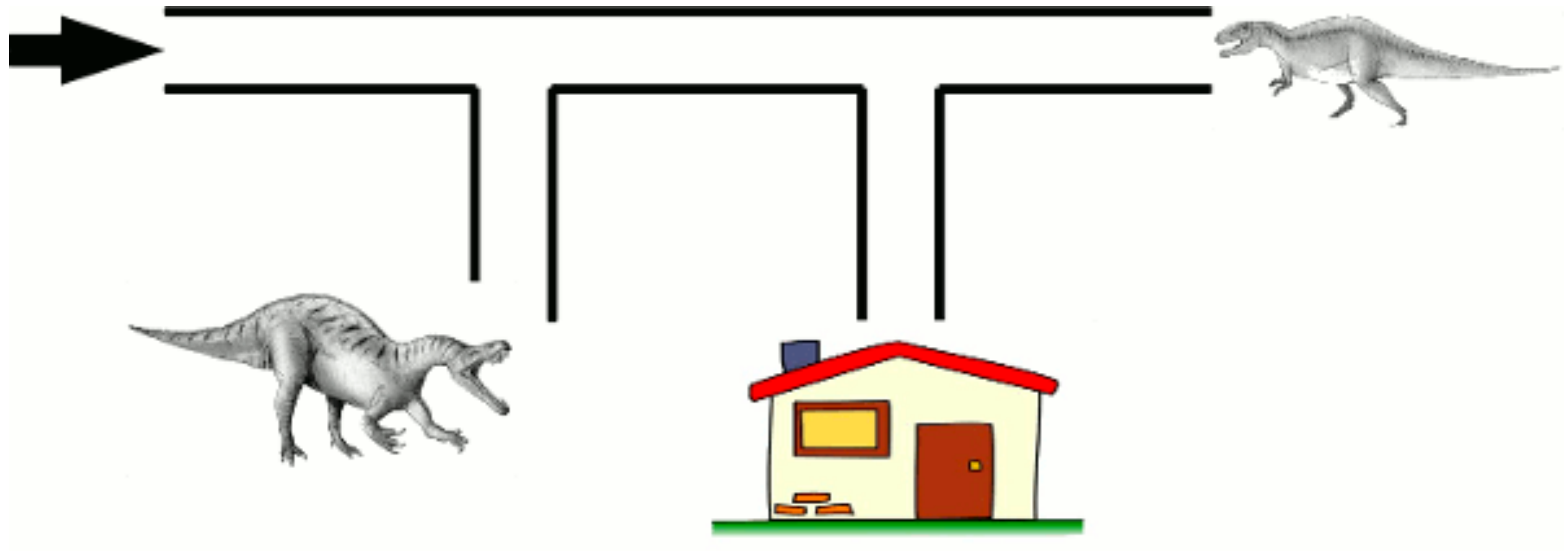
- Part of the motivation for introducing these models is that they allow for both causal *and* game-theoretic reasoning
- In earlier work [6], we study:
  - Equilibrium refinements (NE [10], SPE [15], THPE [14])
  - Subgames



Game-Theoretic Reasoning

- Part of the motivation for introducing these models is that they allow for both causal *and* game-theoretic reasoning
- In earlier work [6], we study:
    - Equilibrium refinements (NE [10], SPE [15], THPE [14])
    - Subgames
    - Information and absentmindedness [12]





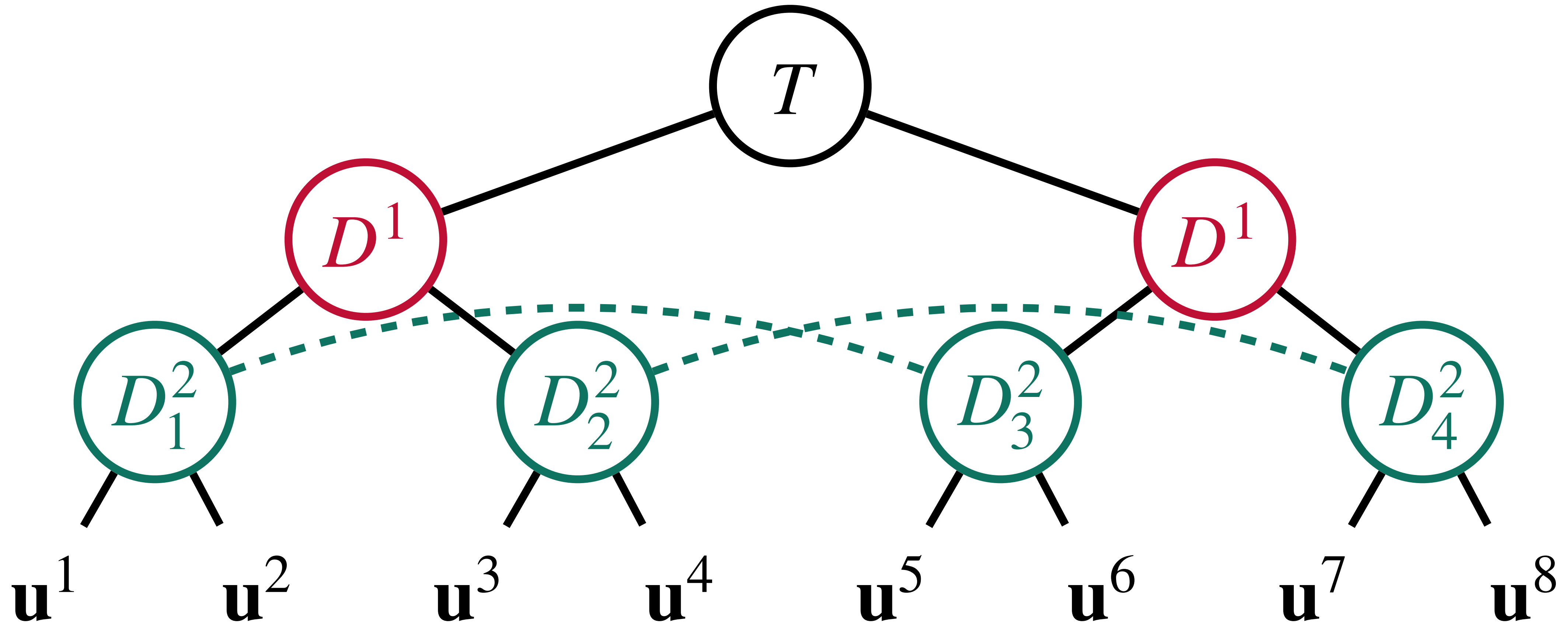
Other Models

Other Models

- Dynamic strategic decision-making most often modelled using EFGs

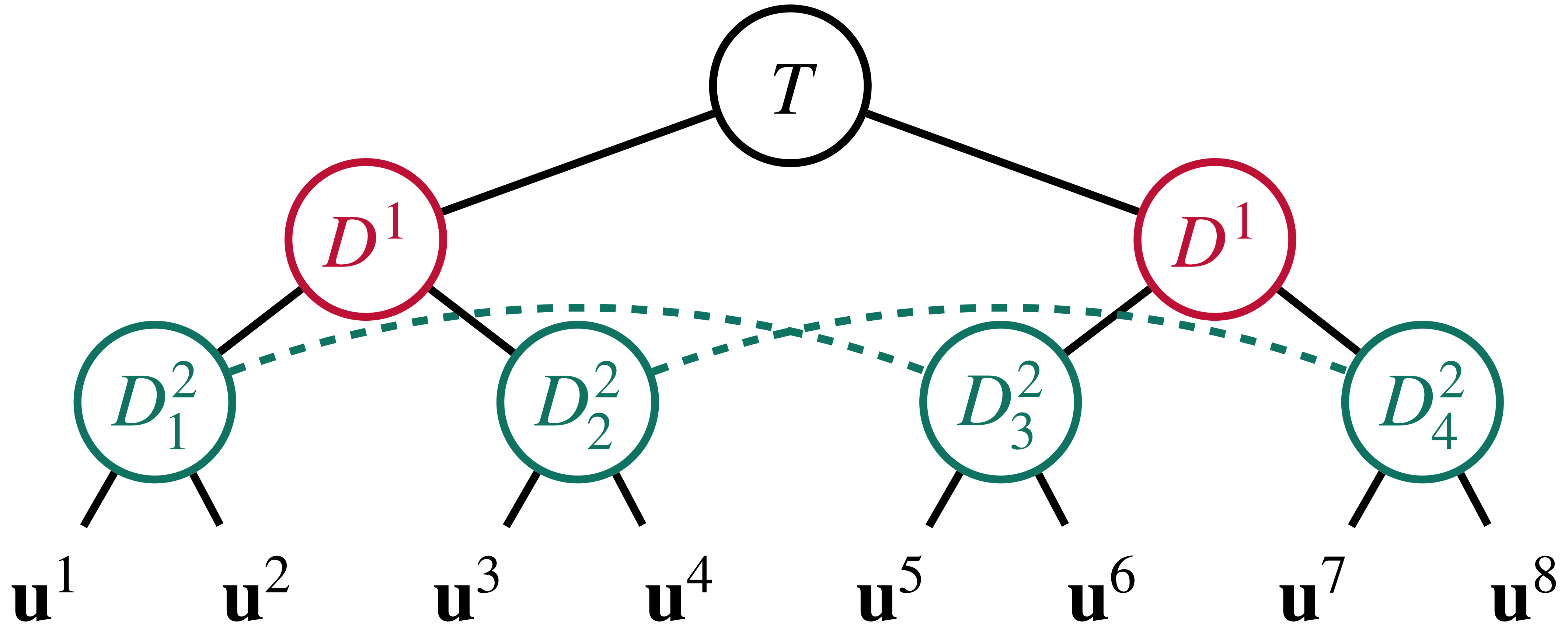
Other Models

- Dynamic strategic decision-making most often modelled using EFGs



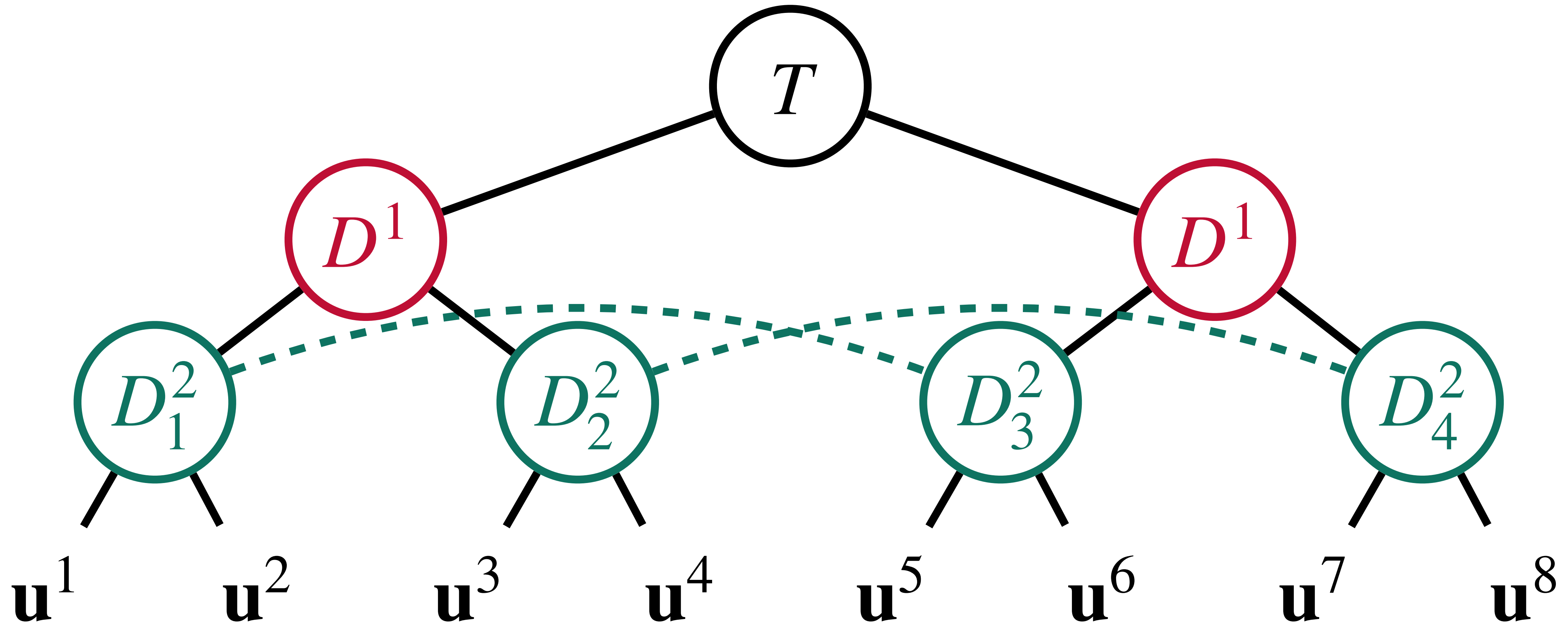
Other Models

- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things



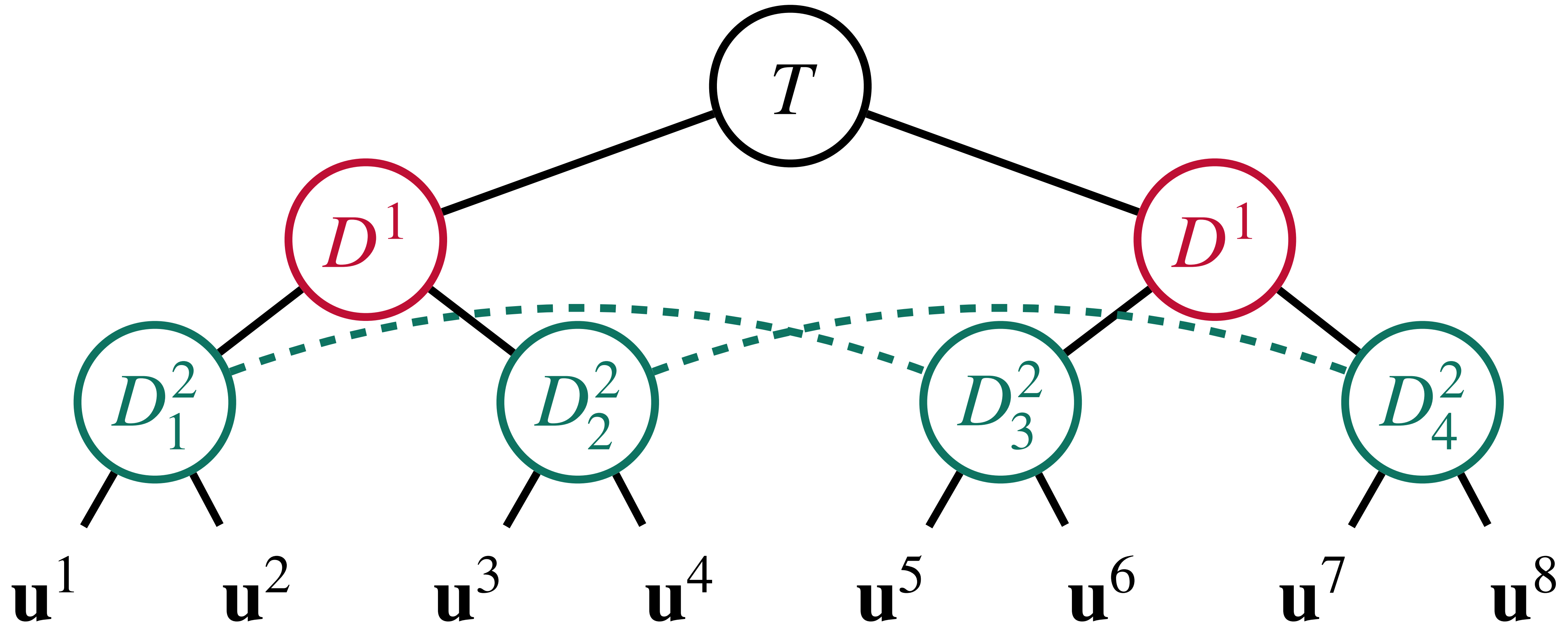
Other Models

- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things
  - Worse for reasoning about causality



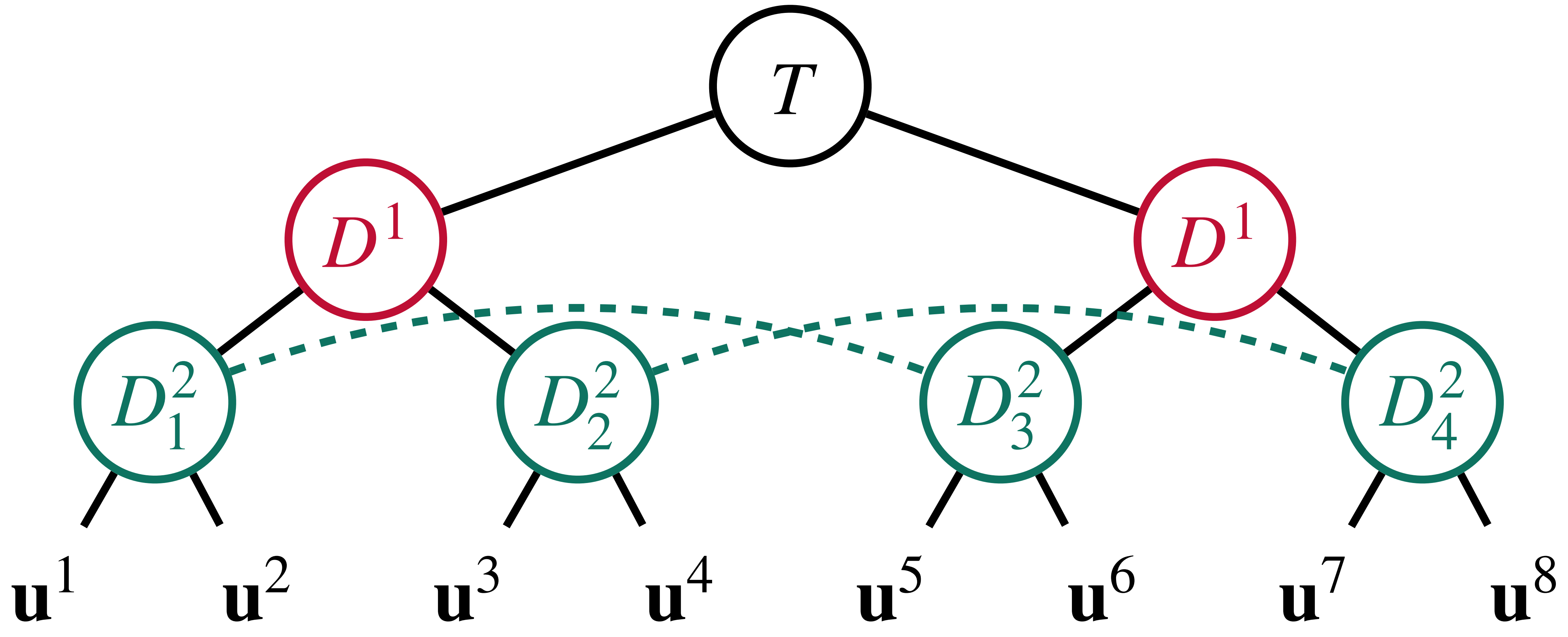
Other Models

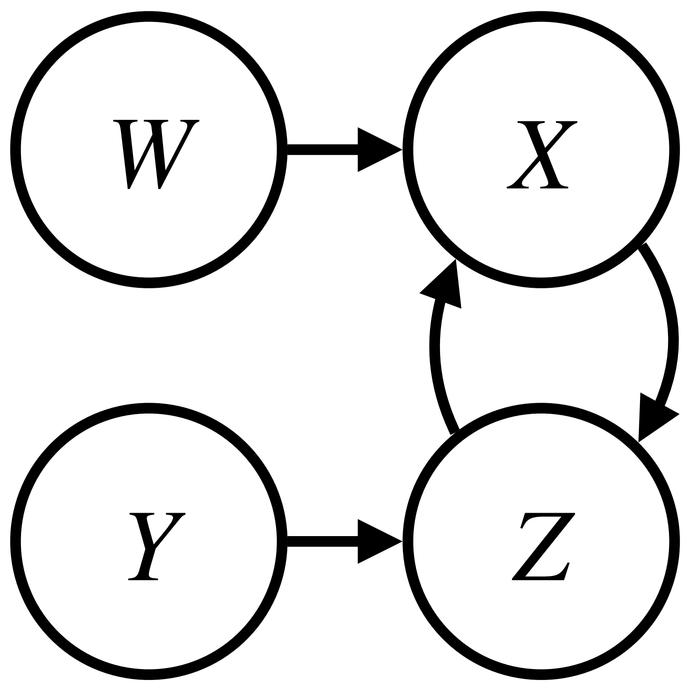
- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things
  - Worse for reasoning about causality
- Other causal models capturing equilibria or optimisation, but no emphasis on strategic reasoning



Other Models

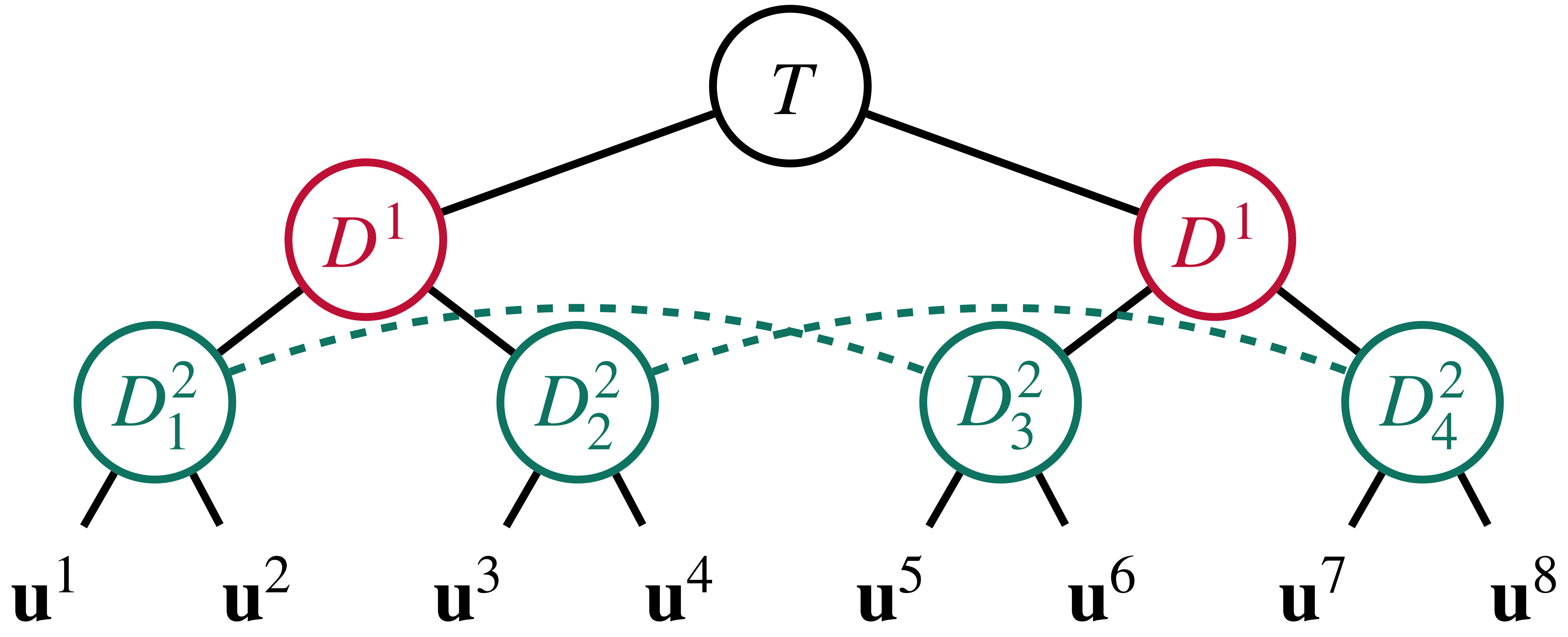
- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things
  - Worse for reasoning about causality
- Other causal models capturing equilibria or optimisation, but no emphasis on strategic reasoning
  - Cyclic causal models [1]

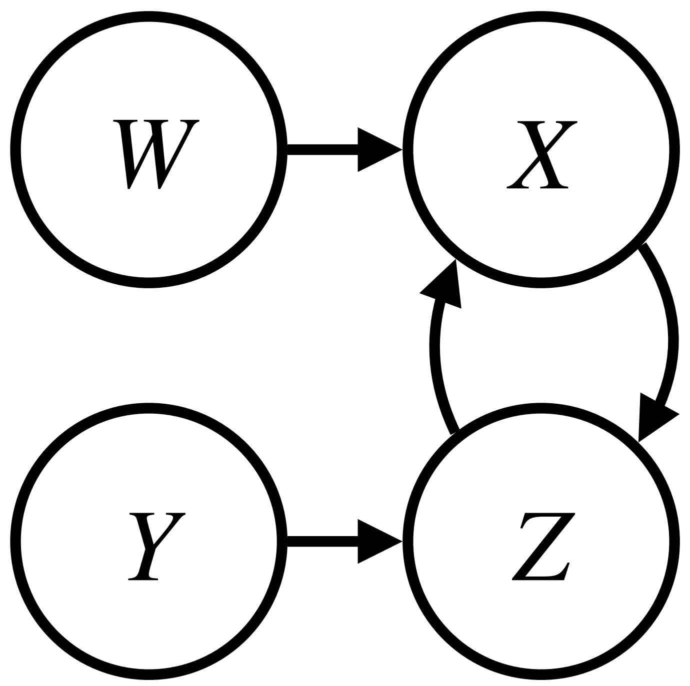


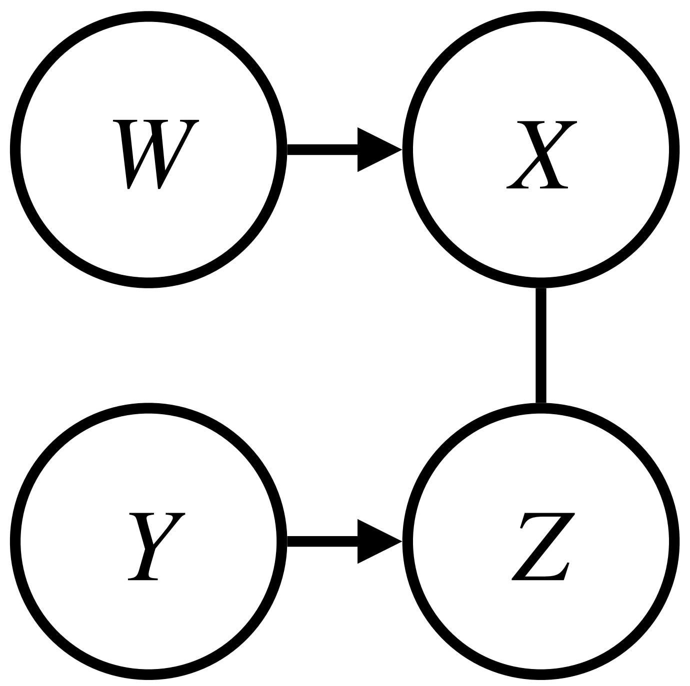


Other Models

- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things
  - Worse for reasoning about causality
- Other causal models capturing equilibria or optimisation, but no emphasis on strategic reasoning
  - Cyclic causal models [1]
  - Chain graphs [9]

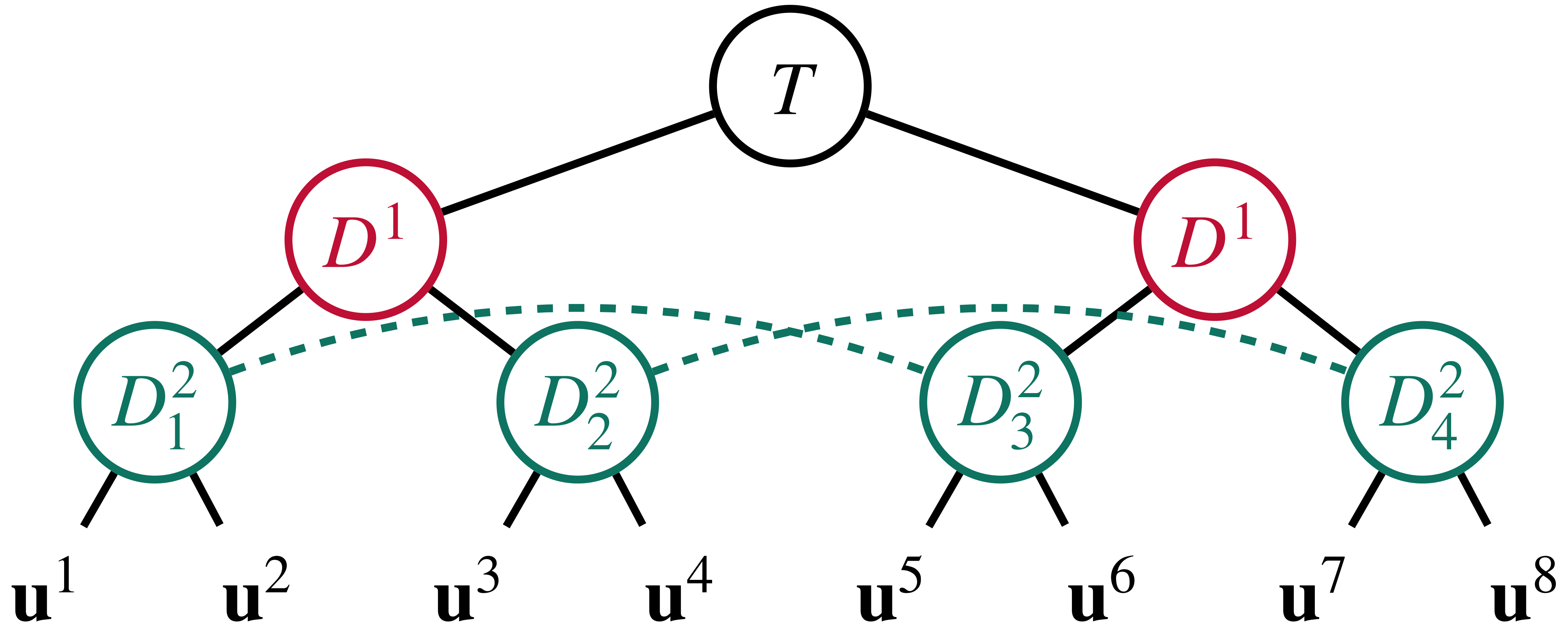


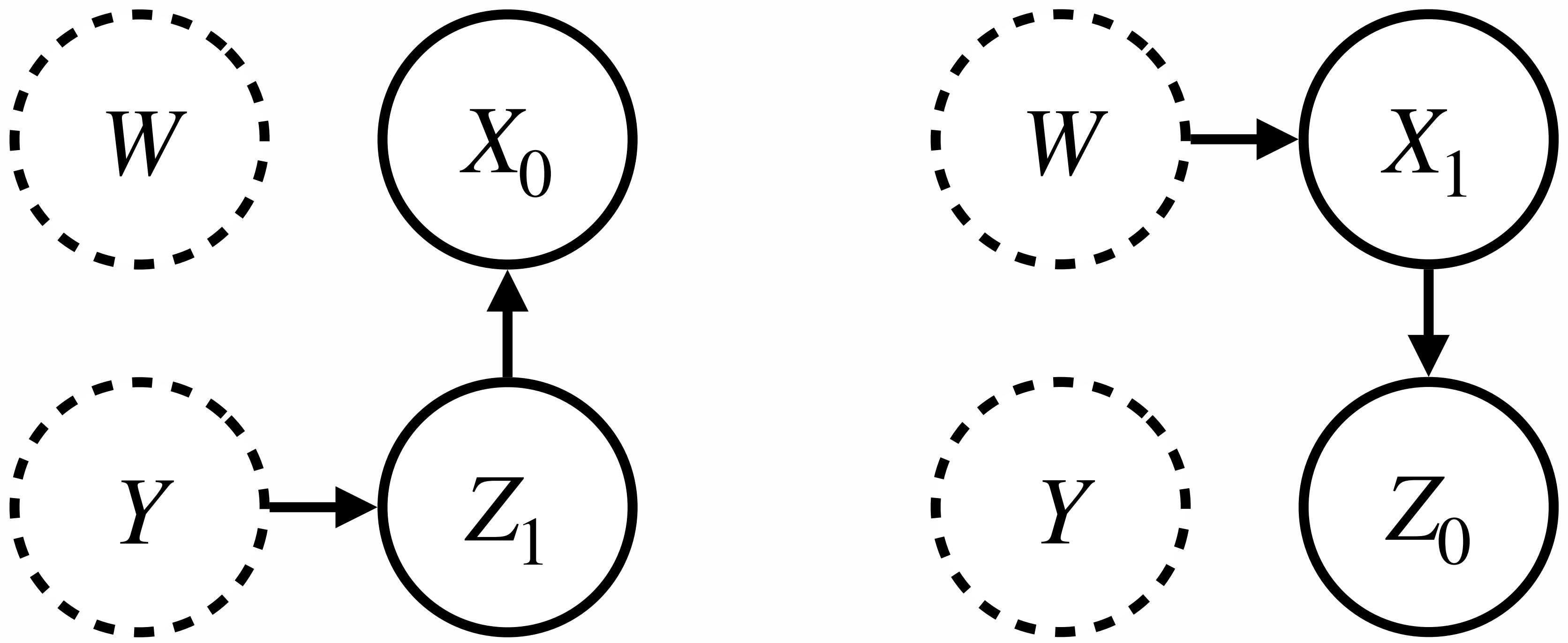




Other Models

- Dynamic strategic decision-making most often modelled using EFGs
  - Better for some things
  - Worse for reasoning about causality
- Other causal models capturing equilibria or optimisation, but no emphasis on strategic reasoning
  - Cyclic causal models [1]
  - Chain graphs [9]
  - Settable systems [17]





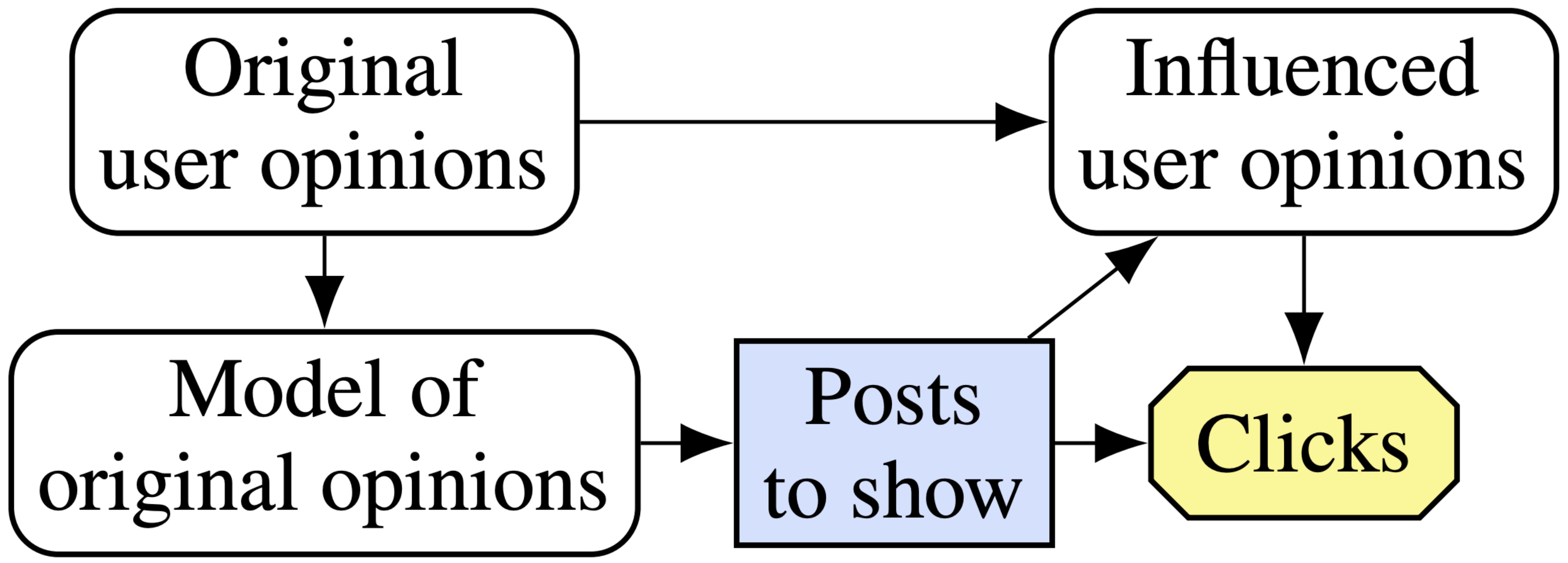
Applications

Applications

- Our main interest is in making AI systems safer, fairer, and better at cooperating

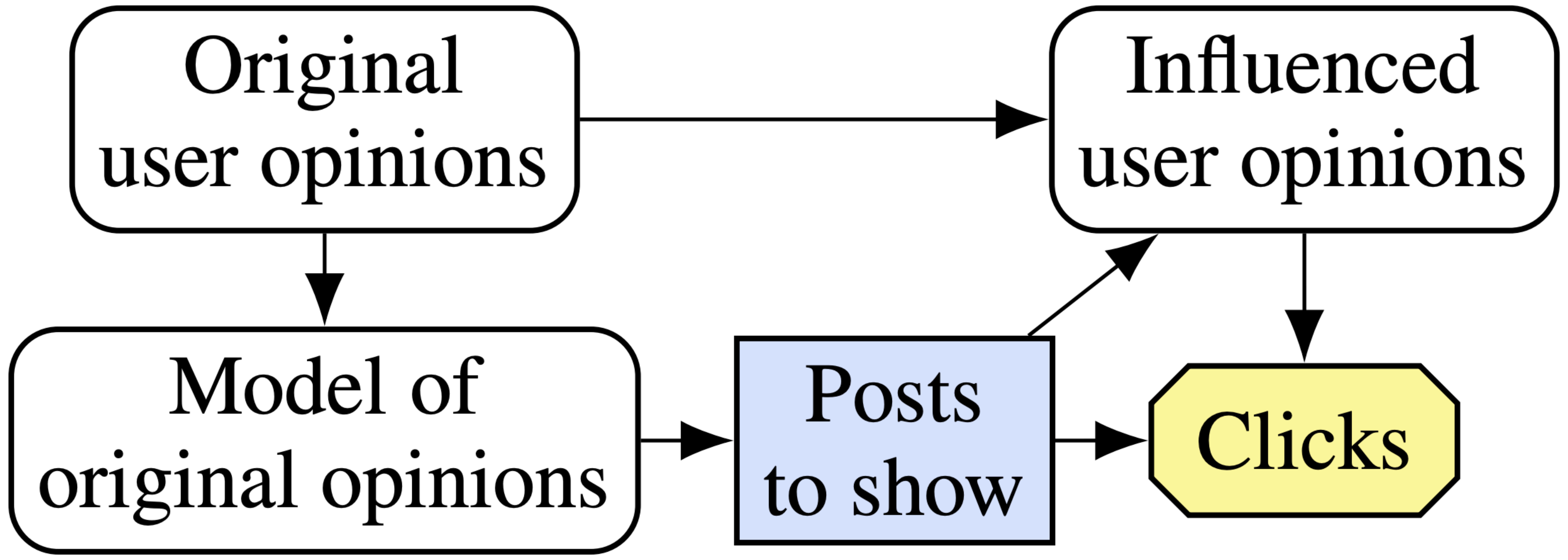
Applications

- Our main interest is in making AI systems safer, fairer, and better at cooperating
  - To ensure safety, we want guarantees that AI systems won't have incentives to do bad things [4]



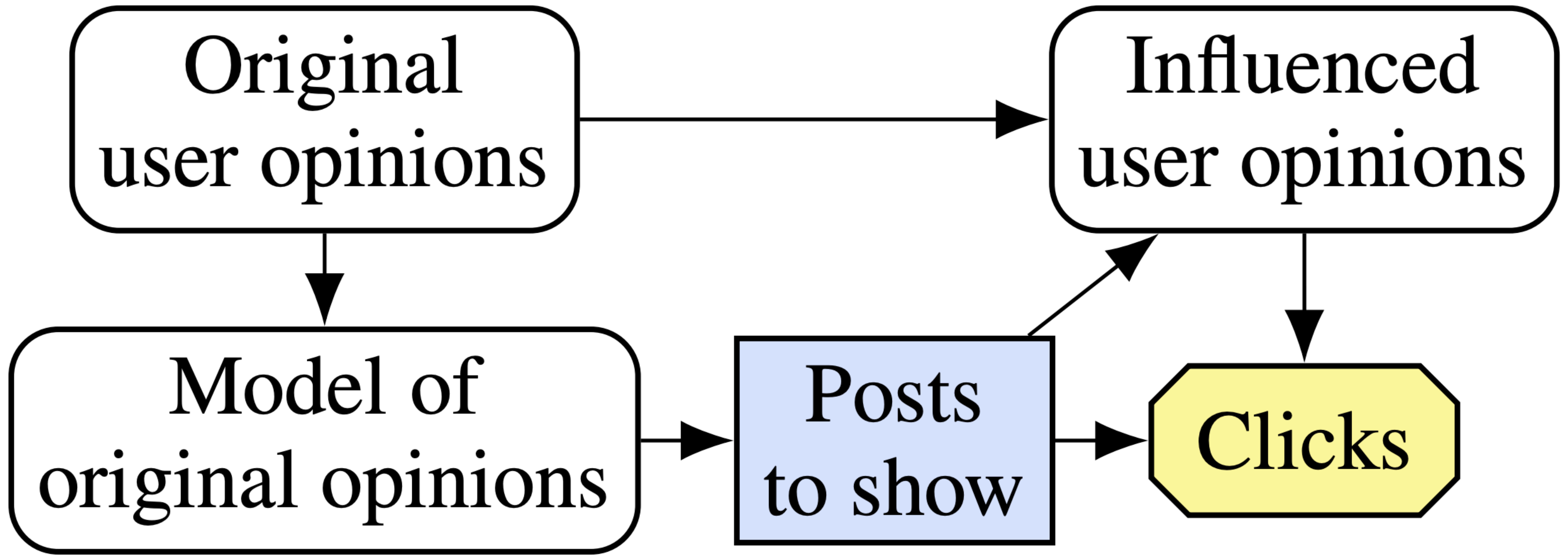
Applications

- Our main interest is in making AI systems safer, fairer, and better at cooperating
  - To ensure safety, we want guarantees that AI systems won't have incentives to do bad things [4]
  - If they do bad things, we want ways to assess blame and intention [5]



Applications

- Our main interest is in making AI systems safer, fairer, and better at cooperating
  - To ensure safety, we want guarantees that AI systems won't have incentives to do bad things [4]
  - If they do bad things, we want ways to assess blame and intention [5]
  - We also want to allow AI systems to harness these notions in order to learn to cooperate [7]



So What?

So What?

- Being able to reason causally about strategic interactions is important for understanding and predicting agents

So What?

- Being able to reason causally about strategic interactions is important for understanding and predicting agents
  - Causality is intrinsic to incentives, fairness, blame, intent, explanations, threats/offers, social influence, etc.

So What?

- Being able to reason causally about strategic interactions is important for understanding and predicting agents
  - Causality is intrinsic to incentives, fairness, blame, intent, explanations, threats/offers, social influence, etc.
- Previously we had causal models without game-theoretic concepts (and vice versa)

So What?

- Being able to reason causally about strategic interactions is important for understanding and predicting agents
  - Causality is intrinsic to incentives, fairness, blame, intent, explanations, threats/offers, social influence, etc.
- Previously we had causal models without game-theoretic concepts (and vice versa)
- Now we have both combined in (what I claim is) a general, formal, and rich framework that subsumes precursors

So What?

- Being able to reason causally about strategic interactions is important for understanding and predicting agents
  - Causality is intrinsic to incentives, fairness, blame, intent, explanations, threats/offers, social influence, etc.
- Previously we had causal models without game-theoretic concepts (and vice versa)
- Now we have both combined in (what I claim is) a general, formal, and rich framework that subsumes precursors
- But there's much more to be done!

Thanks for listening!
Any questions?

Full paper coming soon, watch this space!
Find out more: causalincentives.com

lewis.hammond@cs.ox.ac.uk
lewishammond.com
@lrhammond

References

1. S. Bongers, P. Forré, J. Peters, B. Schölkopf, and J. M. Mooij, "Foundations of Structural Causal Models with Cycles and Latent Variables," *arXiv: 1611.06221*. 2016.
2. A. P. Dawid, "Influence Diagrams for Causal Modelling and Inference," *International Statistical Review (70:2)*, Pages 161–189. International Statistical Institute (ISI). 2002.
3. T. Everitt, R. Kumar, V. Krakovna and S. Legg, "Modeling AGI Safety Frameworks with Causal Influence Diagrams," *arXiv:1906.08663*. 2019.
4. T. Everitt, R. Carey, E. Langlois, P. Ortega and S. Legg, "Agent Incentives: a Causal Perspective," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, Pages 11487–11495. 2021.
5. J. Y. Halpern and M. Kleiman-Weiner, "Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, Pages 1853–1860. 2018.
6. L. Hammond, J. Fox, T. Everitt, A. Abate, and M. Wooldridge, "Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS-21)*, Pages 574–582. 2021.
7. N. Jaques, A. Lazaridou, E. Hughes, Ç. Gülçehre, P. A. Ortega, D. Strouse, J. Z. Leibo and N. de Freitas, "Social Influence As Intrinsic Motivation for Multi-agent Deep Reinforcement Learning," in *Proceedings of the 36th International Conference on Machine Learning (ICML-19)*, Pages 3040–3049. 2019.
8. D. Koller and B. Milch, "Multi-agent Influence Diagrams for Representing and Solving Games," *Games and Economic Behavior (45:1)*, Pages 181–221. Elsevier. 2003.

References

9. S. L. Lauritzen and T. S. Richardson, "Chain graph models and their causal interpretations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64.3*, Pages 321–348. 2002.

10. J. F. Nash, "Equilibrium Points in N-person Games," *Proceedings of the National Academy of Sciences (36:1)*, Pages 48–49. 1950.

11. J. Pearl, *Causality*. Cambridge University Press. 2009.

12. M. Piccione and A. Rubinstein, "The Absent-Minded Driver's Paradox: Synthesis and Responses," *Games and Economic Behavior (20:1)*, Pages 121–130. Elsevier. 1997.

13. A. Pfeffer and Y. Gal, "On the Reasoning Patterns of Agents in Games," in *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 102–109. 2007.

14. R. Selten, "Reexamination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory (4:1)*, Pages 25–55. Springer. 1975.

15. R. Selten, "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit: Teil i: Bestimmung des Dynamischen Preisgleichgewichts". *Journal of Institutional and Theoretical Economics H.2*, Pages 301–324. 1965.

16. M. Spence, "Job Market Signaling,"*The Quarterly Journal of Economics (87:3)*, Pages 355–374. Oxford University Press. 1973.

17. H. White and K. Chalak, "Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning," *Journal of Machine Learning Research* 10, Pages 1759–1799. 2009.