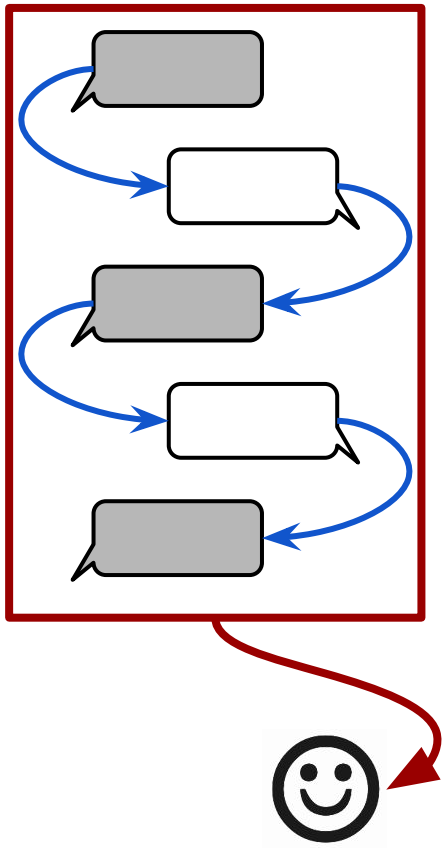
Wishful thinking: Making causal claims about conversational behavior

Cristian Danescu-Niculescu-Mizil
Cornell University

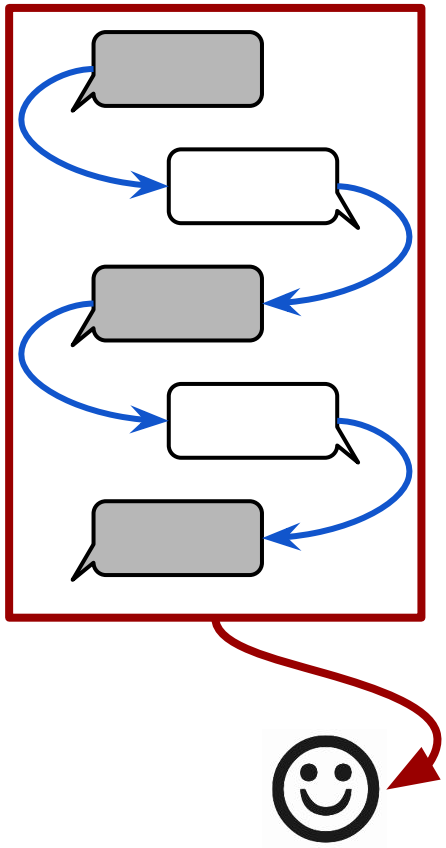
Includes slides prepared by Justine Zhang

conversational behavior

What conversational behavior **leads** to better outcomes?



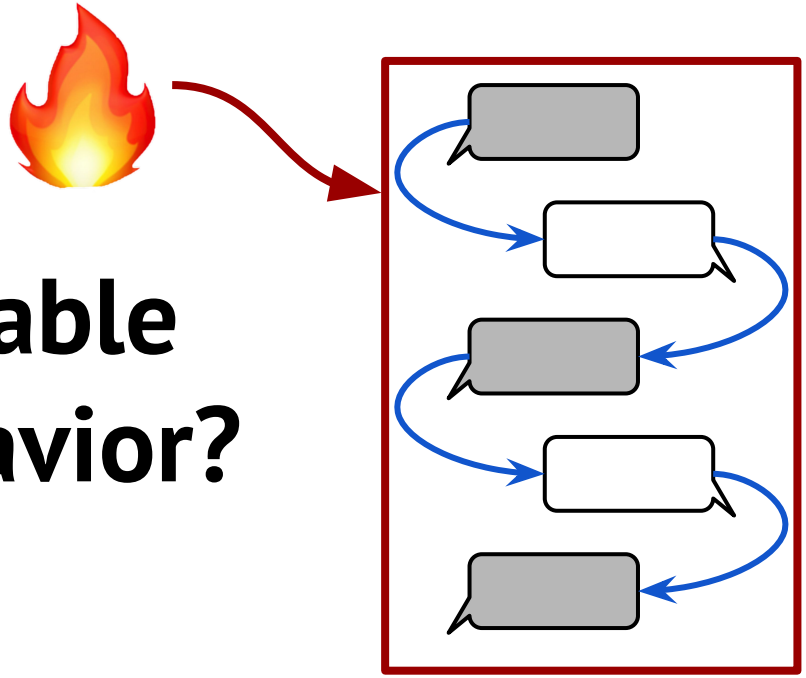
What conversational behavior **leads** to better outcomes?



Potential applications:

- **training**
- **assignment**
- **assistance**

  […]

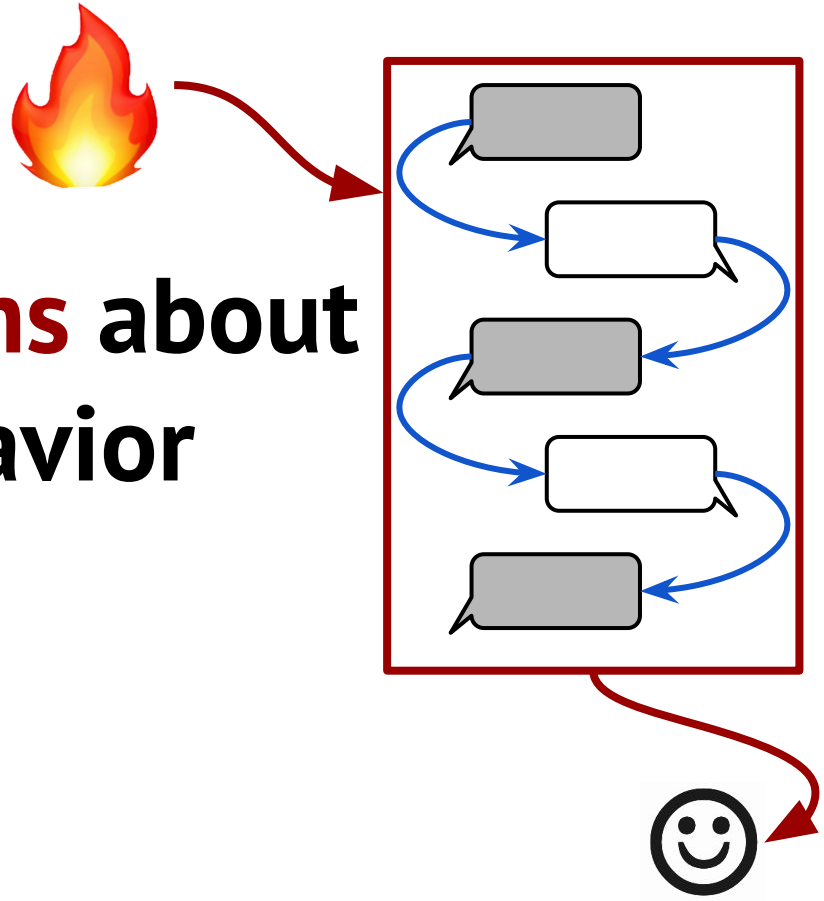


What leads to desirable conversational behavior?

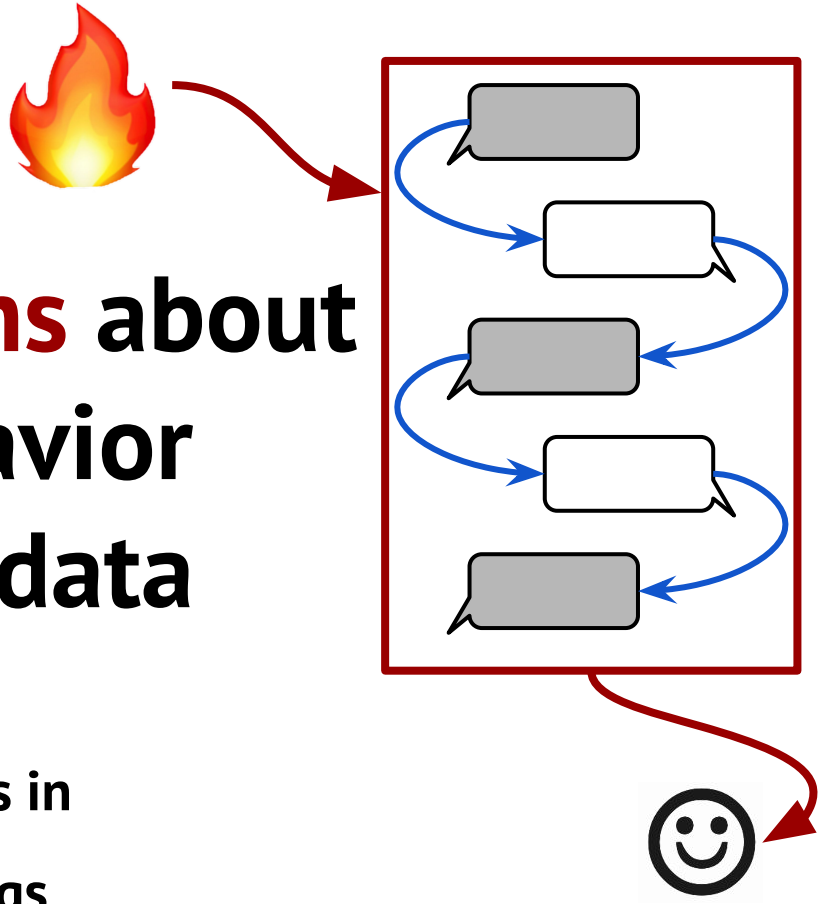
Potential applications:

- **platform design**
- **moderation**

Making causal claims about conversational behavior



Making causal claims about conversational behavior from observational data



This talk: three inference challenges in two conversational settings

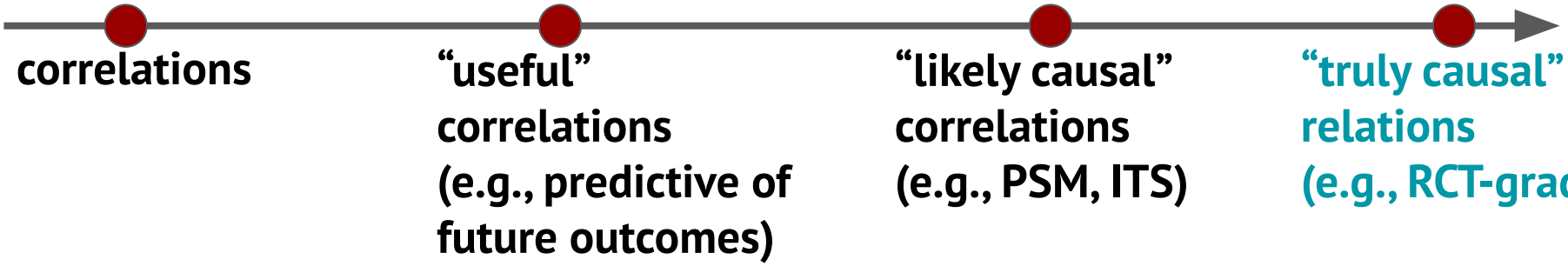
Making causal claims about
[anything]
from observational data

Making causal claims about
[anything]
from observational data

correlations

“truly causal” relations (e.g., RCT-grade)

Making causal claims about [anything] from observational data



Wishful thinking:

Making causal claims about

[anything]

from observational data

Wishful thinking:

correlations

“truly causal” relations (e.g., RCT-grade)

Wishful thinking:

Imagine ideal setting needed for establishing "truly causal" relations

Wishful thinking:

Imagine ideal setting needed for establishing "truly causal" relations

Formally describe what is missing in our non-ideal observational setting

Wishful thinking:

Imagine ideal setting needed for establishing "truly causal" relations

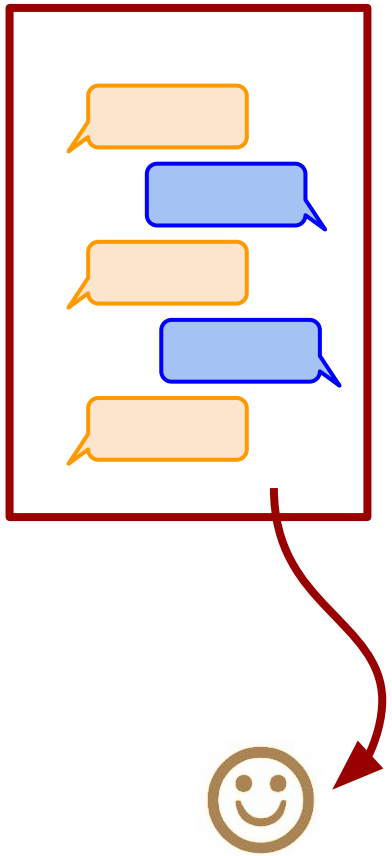
Formally describe what is missing in our non-ideal observational setting

Can we find an alternative that is just as good (or "good enough")?

correlations

"truly causal" relations (e.g., RCT-grade)

What conversational behavior **leads** to better outcomes?



Quantifying the Causal Effects of Conversational Tendencies
Justine Zhang, Sendhil Mullainathan, and
Cristian Danescu-Niculescu-Mizil
Proceedings of CSCW, 2020

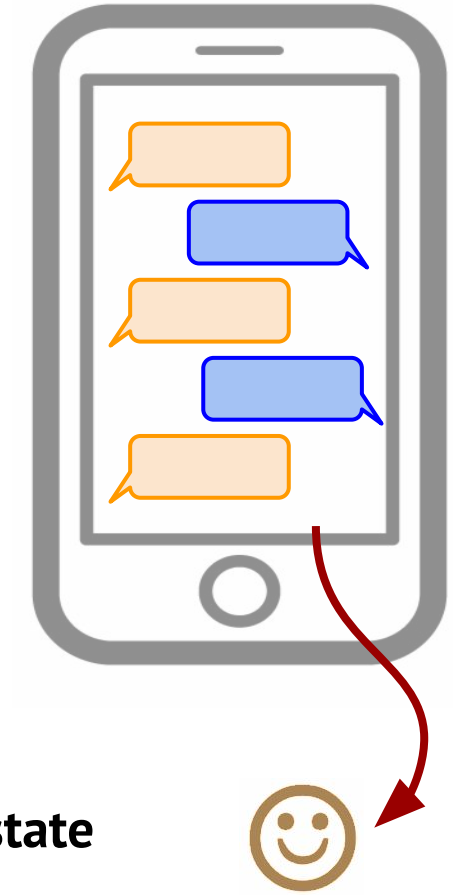
Empirical setting:
Crisis counseling conversations

CRISIS TEXT LINE |

Individuals in mental health crisis
(anxiety, suicidal ideation, …)

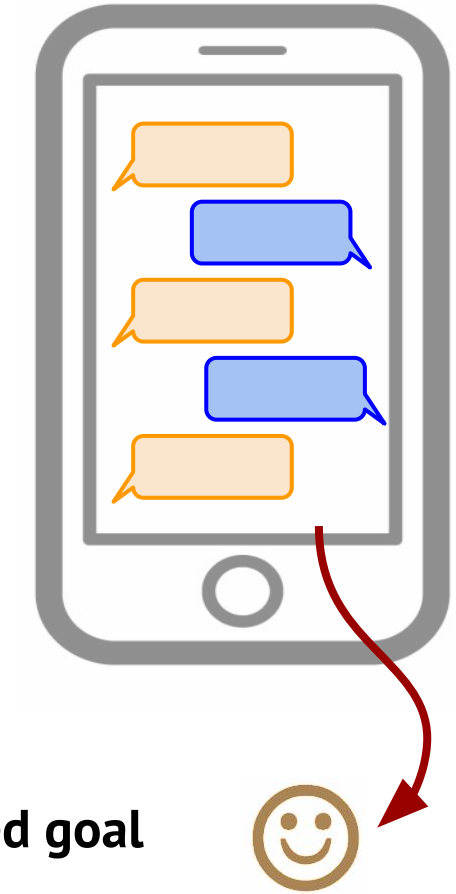
trained counselors

1.5 million conversations



Goal: calmer state

General setting:
goal-oriented conversation platforms

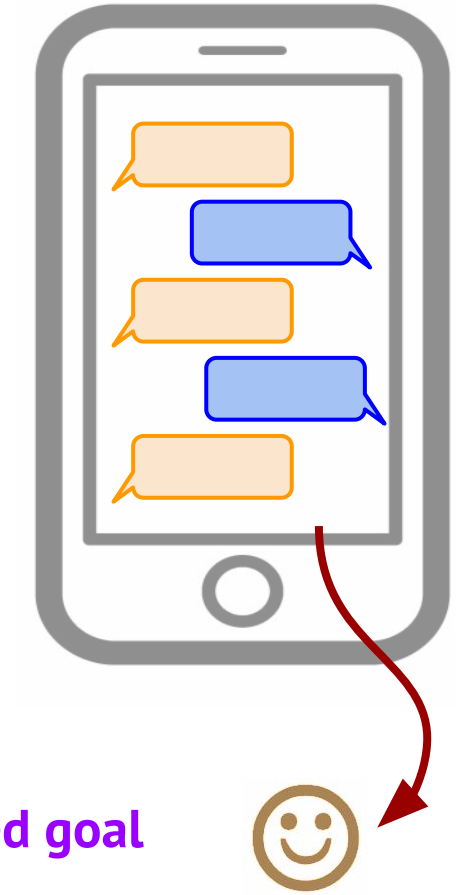


consequential, conversational settings:
crisis counseling, medicine, teaching, therapy, advising, customer service, contact tracing...

General setting:
goal-oriented conversation platforms

What **(agent) conversational behaviour** leads to a **better (client) outcome**?

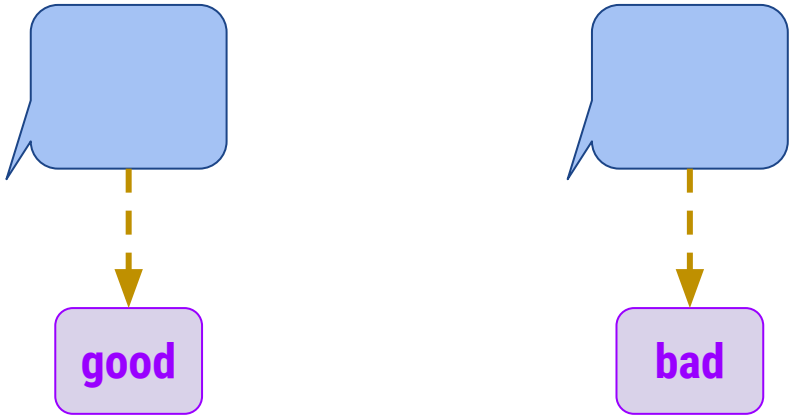
- - -► training or assignment policies



What **(agent) conversational behaviour** leads to a **better (client) outcome**?

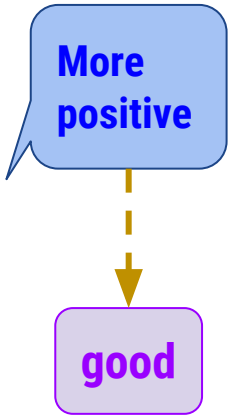
What **(agent) conversational behaviour** leads to a **better (client) outcome**?

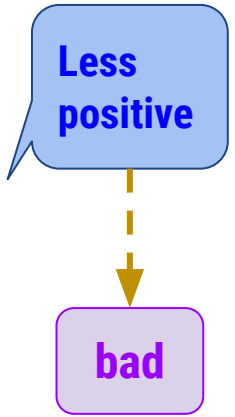
Preliminary idea: compare behaviours in good versus bad conversations



What **(agent) conversational behaviour** leads to a **better (client) outcome**?

Preliminary idea: compare behaviours in good versus bad conversations

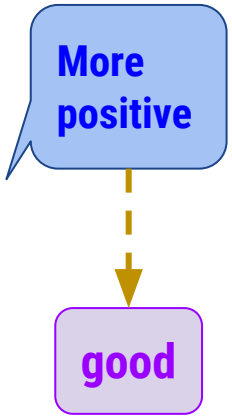


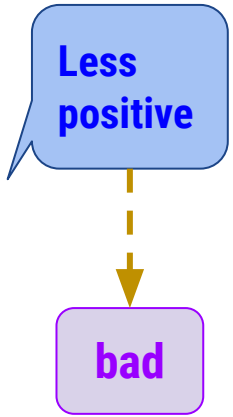


What **(agent) conversational behaviour** leads to a **better (client) outcome**?

- - -► Assign more positive agents?

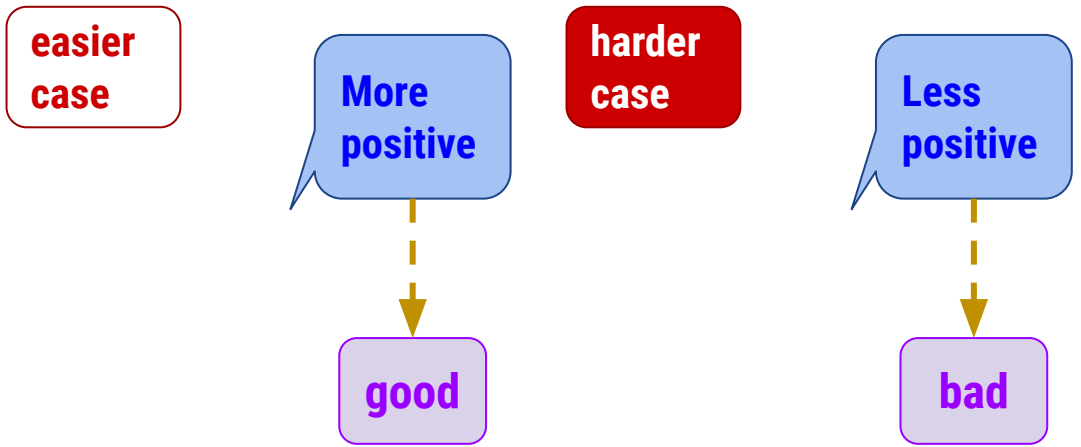
Preliminary idea: compare behaviours in good versus bad conversations





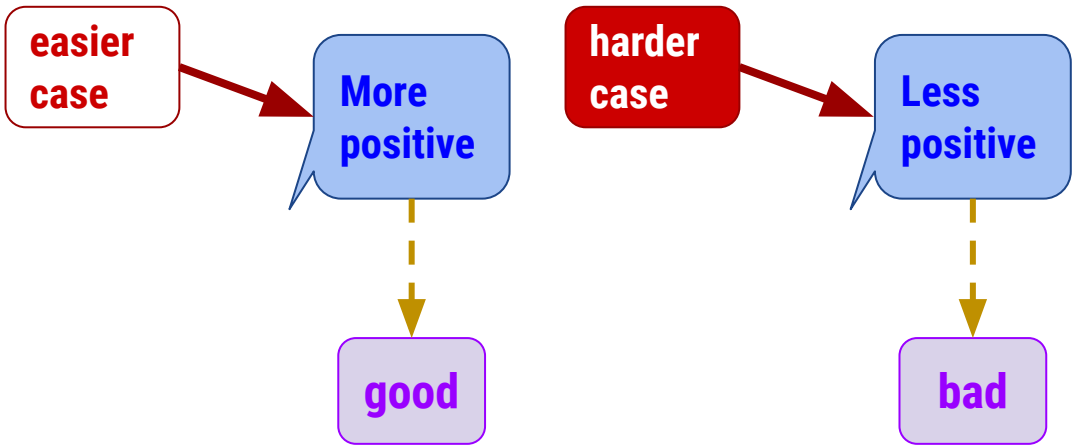
What **(agent) conversational behaviour** leads to a **better (client) outcome**?

- - -► Assign more positive agents?



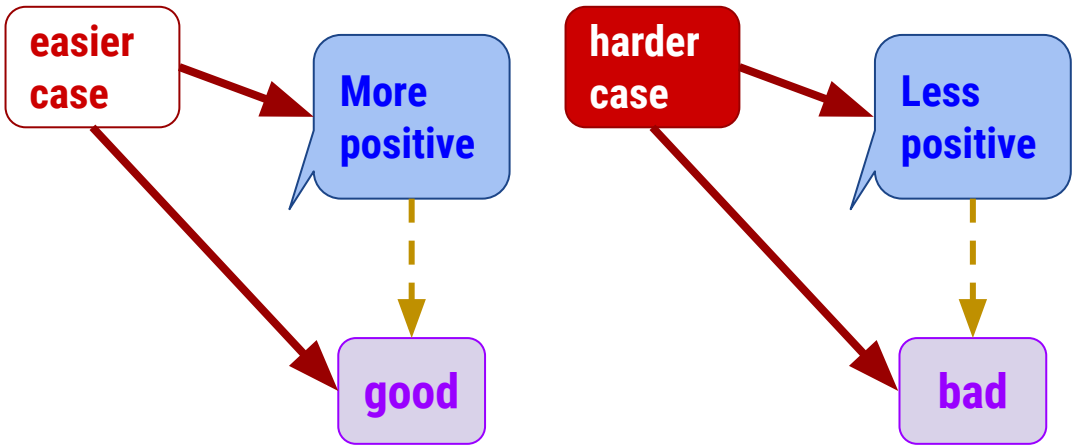
What **(agent) conversational behaviour**
leads to a **better (client) outcome**?

- - -► Assign more positive agents?



What **(agent) conversational behaviour** leads to a **better (client) outcome**?

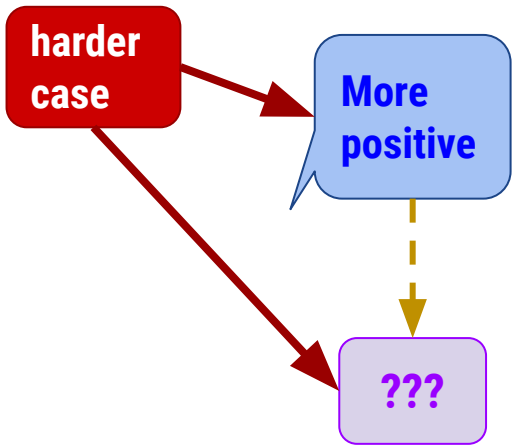
- - -► Assign more positive agents?



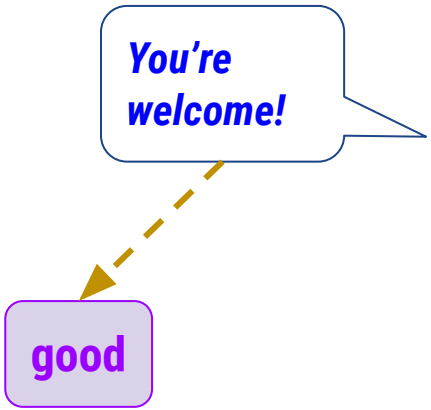
What **(agent) conversational behaviour** leads to a **better (client) outcome**?

- - -► Assign more positive agents?

We cannot address this **counterfactual**:

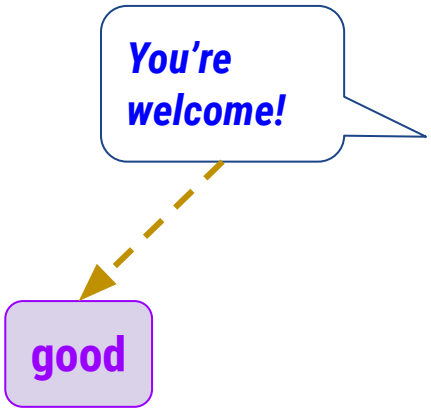


What **(agent) conversational behaviour** leads to a **better (client) outcome**?



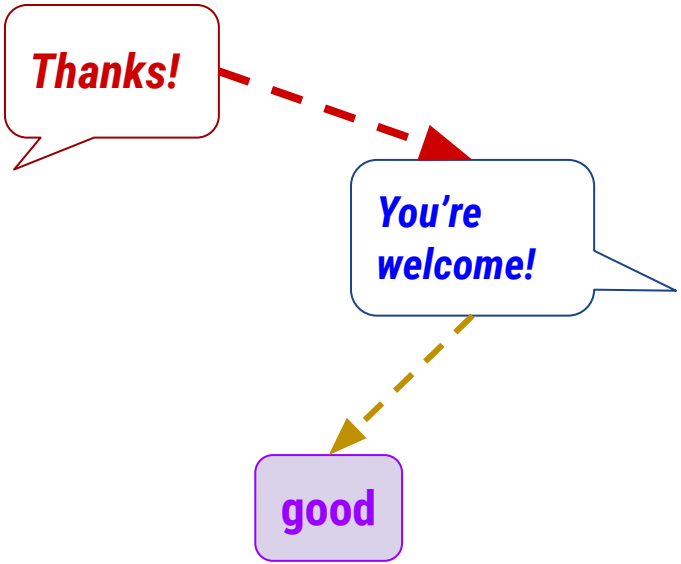
What **(agent) conversational behaviour**
leads to a **better (client) outcome**?

- - -► Teach agents to say “you’re welcome” more?



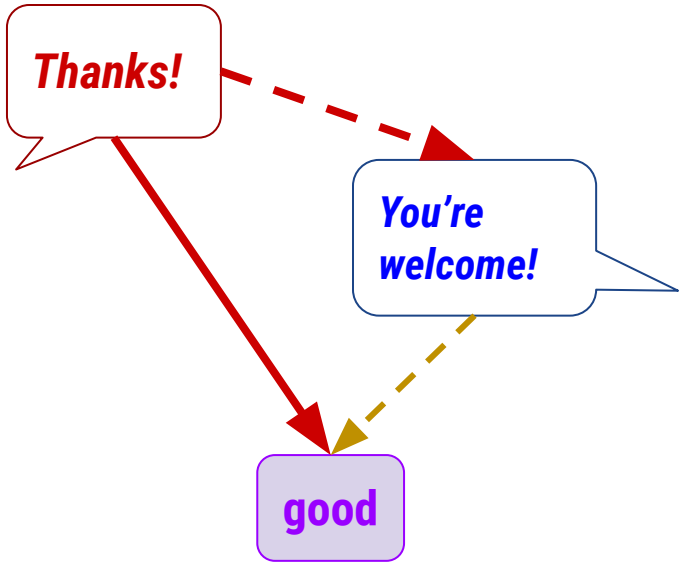
What **(agent) conversational behaviour**
leads to a **better (client) outcome**?

- - -► Teach agents to say “you’re welcome” more?

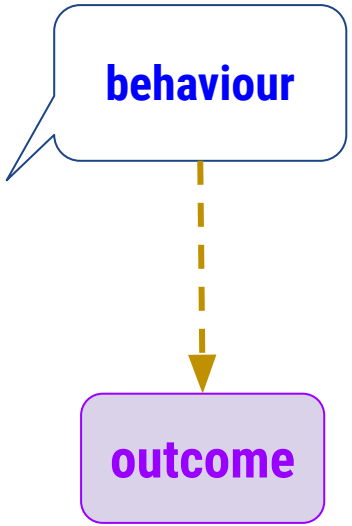


What **(agent) conversational behaviour**
leads to a **better (client) outcome**?

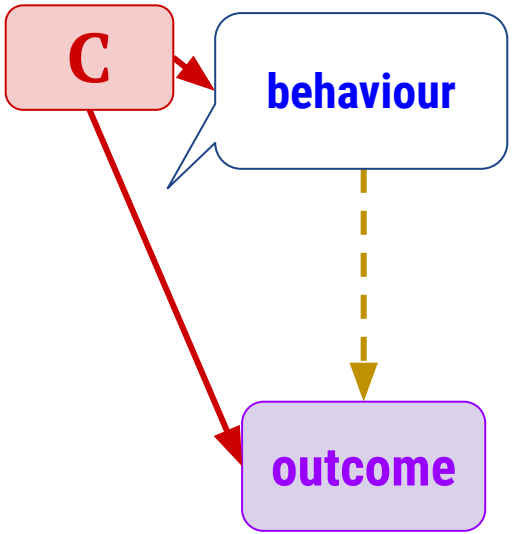
– – –► Teach agents to say “you’re welcome” more?



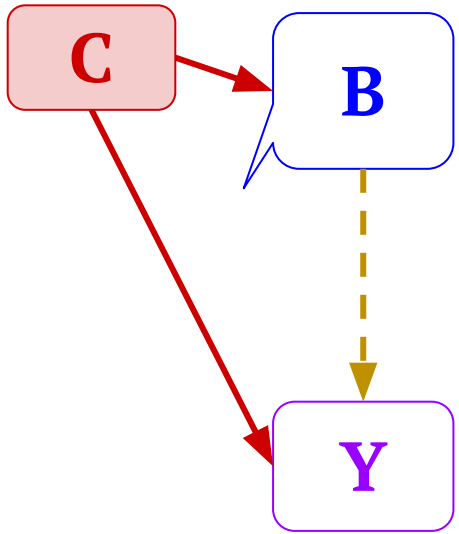
Problem:



Problem: **circumstances** of the conversation

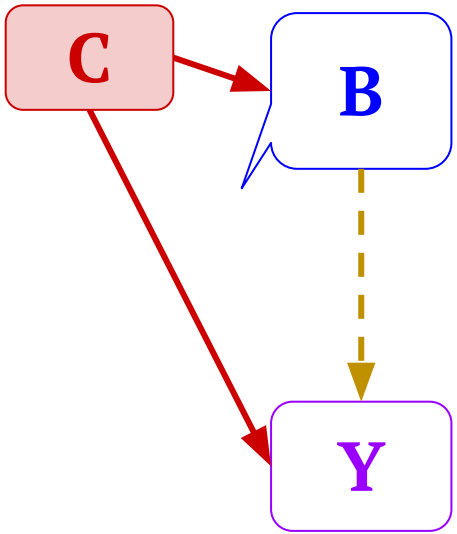


Problem: **circumstances** of the conversation



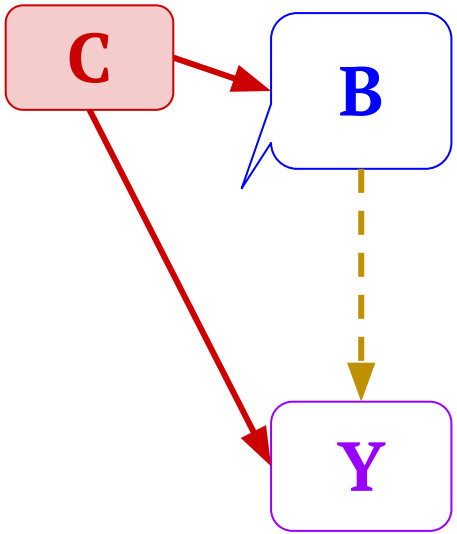
Problem: **circumstances** of the conversation

Drawing on the
causal inference literature,



Problem: **circumstances** of the conversation

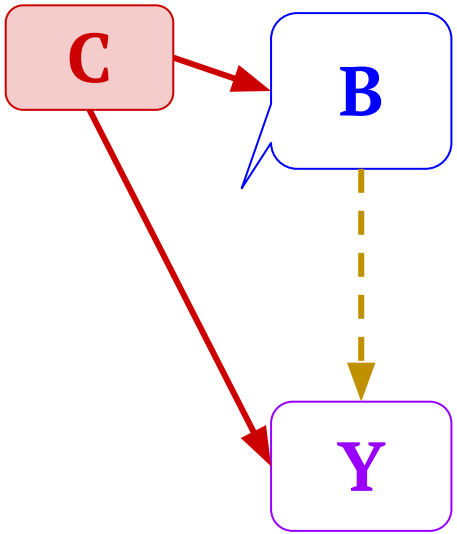
Drawing on the causal inference literature, we formally describe the problem of drawing **causal relationships** between **behaviour** and **outcome**



Problem: **circumstances** of the conversation

Drawing on the causal inference literature, we formally describe the problem of drawing **causal relationships** between **behaviour** and **outcome**,

and highlight **challenges** that arise for conversational settings



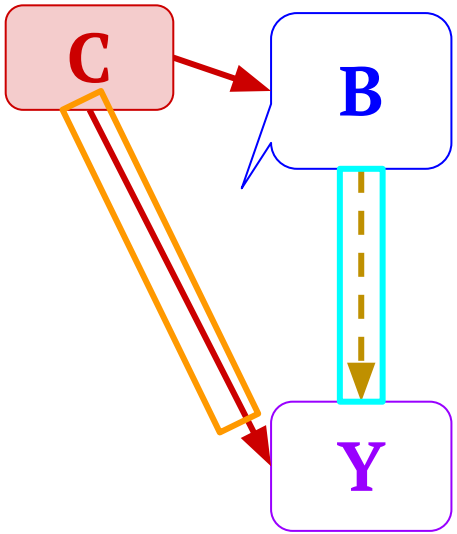
Problem: **circumstances** of the conversation

In this talk: **graphical models** of dependencies between **circumstance**, **behaviour** and **outcome**

Pearl, 1995

Drawing on the causal inference literature, we formally describe the problem of drawing **causal relationships** between **behaviour** and **outcome**,

and highlight **challenges** that arise for conversational settings



Probem: **circumstances** of the conversation

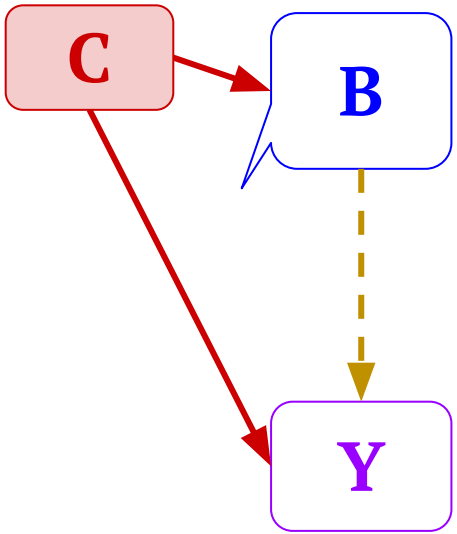
In this talk: **graphical models** of dependencies between **circumstance**, **behaviour** and **outcome**

Pearl, 1995

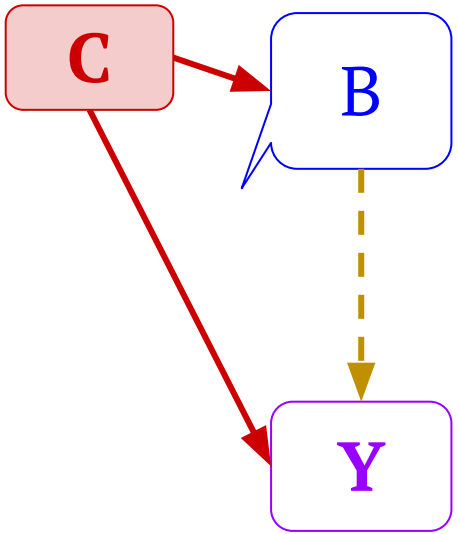
Also in paper: potential outcomes framework

Rosenbaum, 2010

Drawing on the causal inference literature, we formally describe the problem of drawing **causal relationships** between **behaviour** and **outcome**, and highlight **challenges** that arise for conversational settings

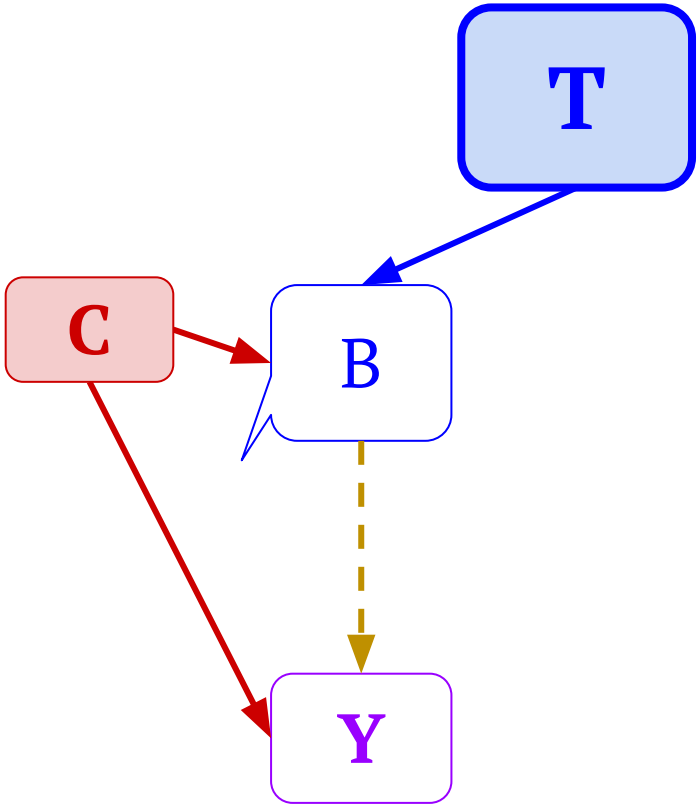


Our motivating task:



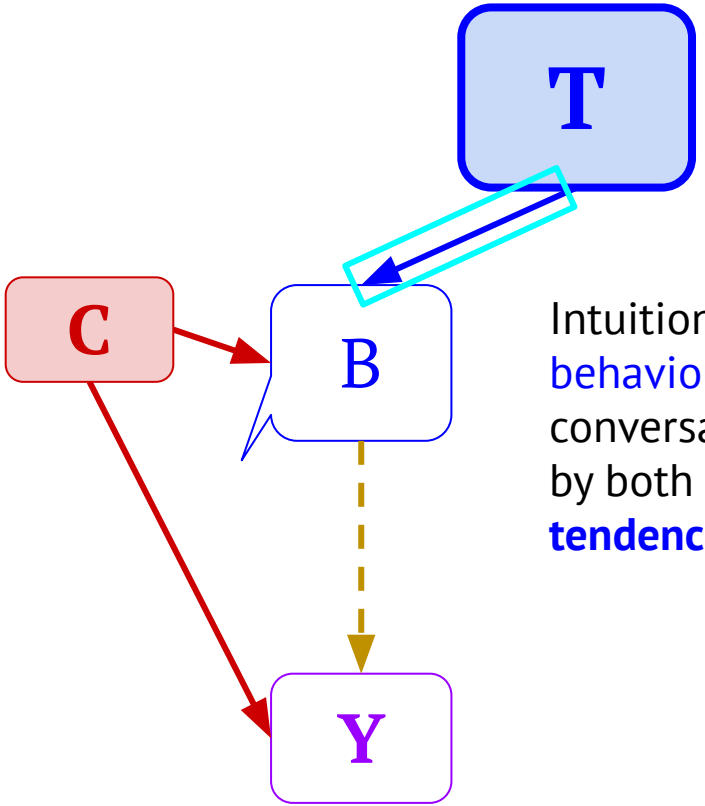
Our motivating task:

Assigning agents based on their **behavioural tendency**



Our motivating task:

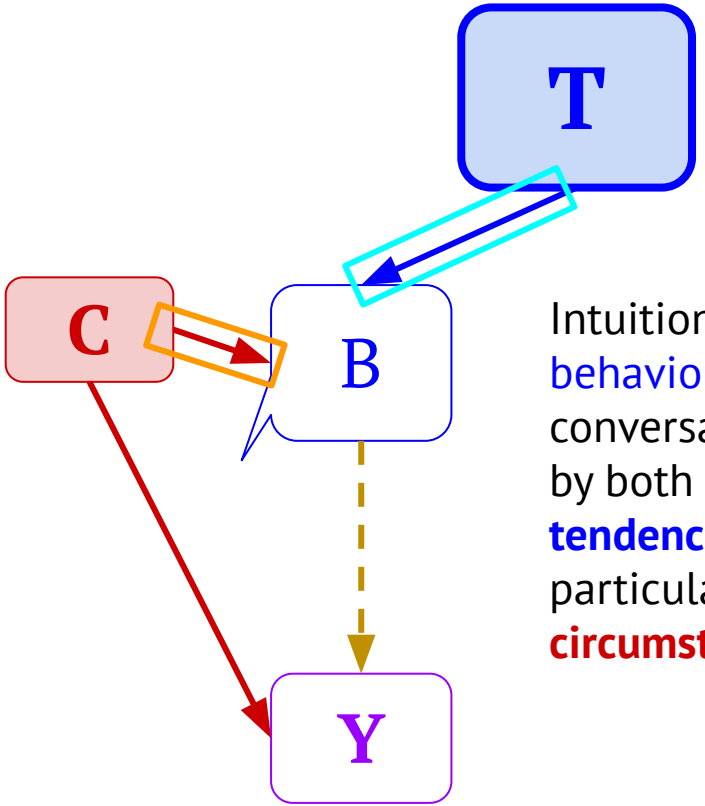
Assigning agents based on their **behavioural tendency**



Intuition: an agent's behaviour in a conversation is shaped by both their **tendencies**

Our motivating task:

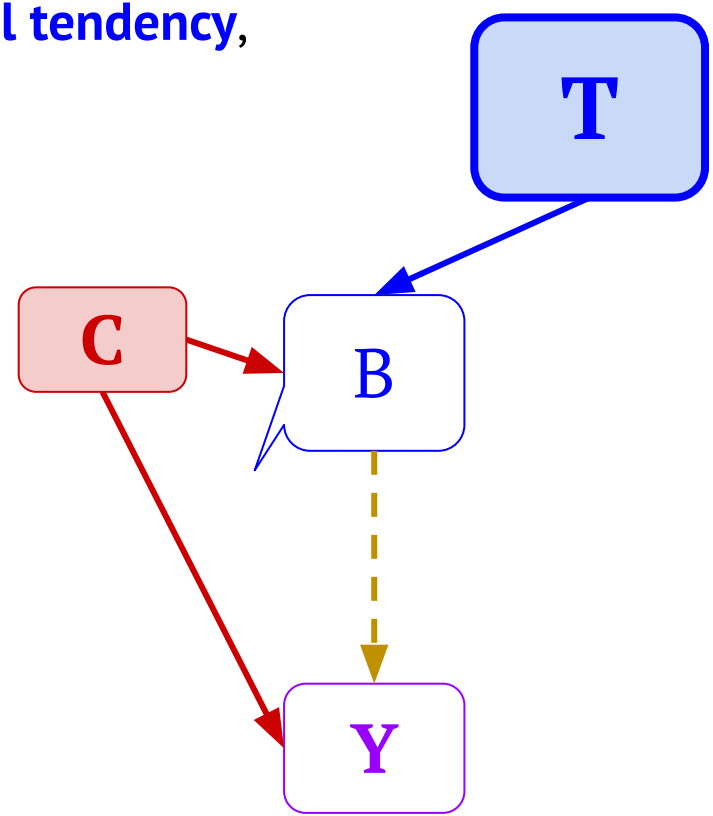
Assigning agents based on their **behavioural tendency**



Intuition: an agent's behaviour in a conversation is shaped by both their **tendencies** and the particular **circumstances**

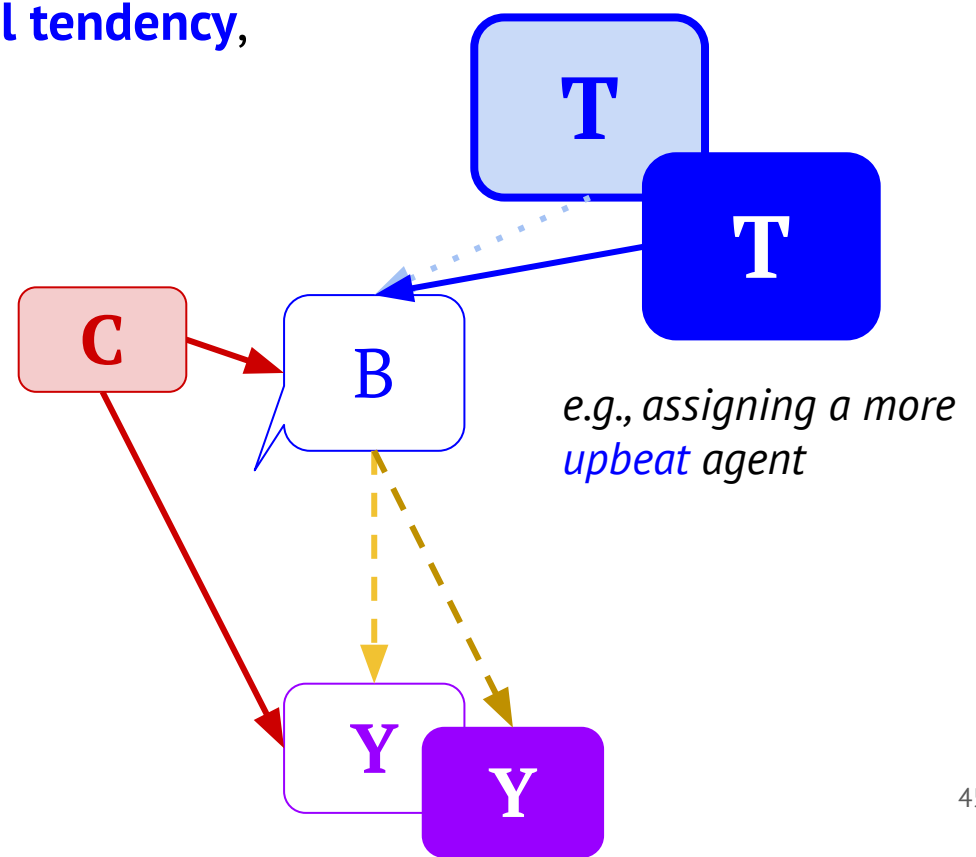
Our motivating task: **tendency-based assignment**

Causal inference problem: estimate **effect of assigning** an agent based on their **behavioural tendency**, as inferred from past conversations



Our motivating task: **tendency-based assignment**

Causal inference problem: estimate **effect of assigning** an agent based on their **behavioural tendency**, as inferred from past conversations

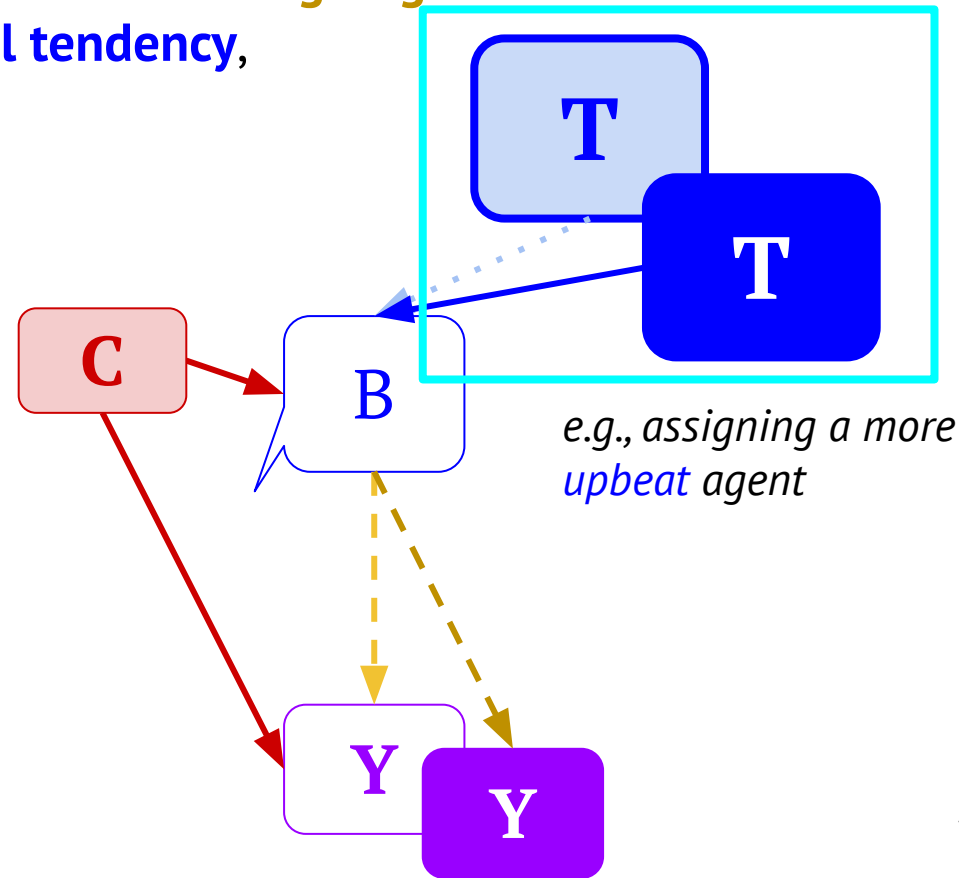


e.g., assigning a more upbeat agent

Our motivating task: **tendency-based assignment**

Causal inference problem: estimate **effect of assigning** an agent based on their **behavioural tendency**, as inferred from past conversations

We want to say that two agents get different **outcomes** because their **tendencies** are different



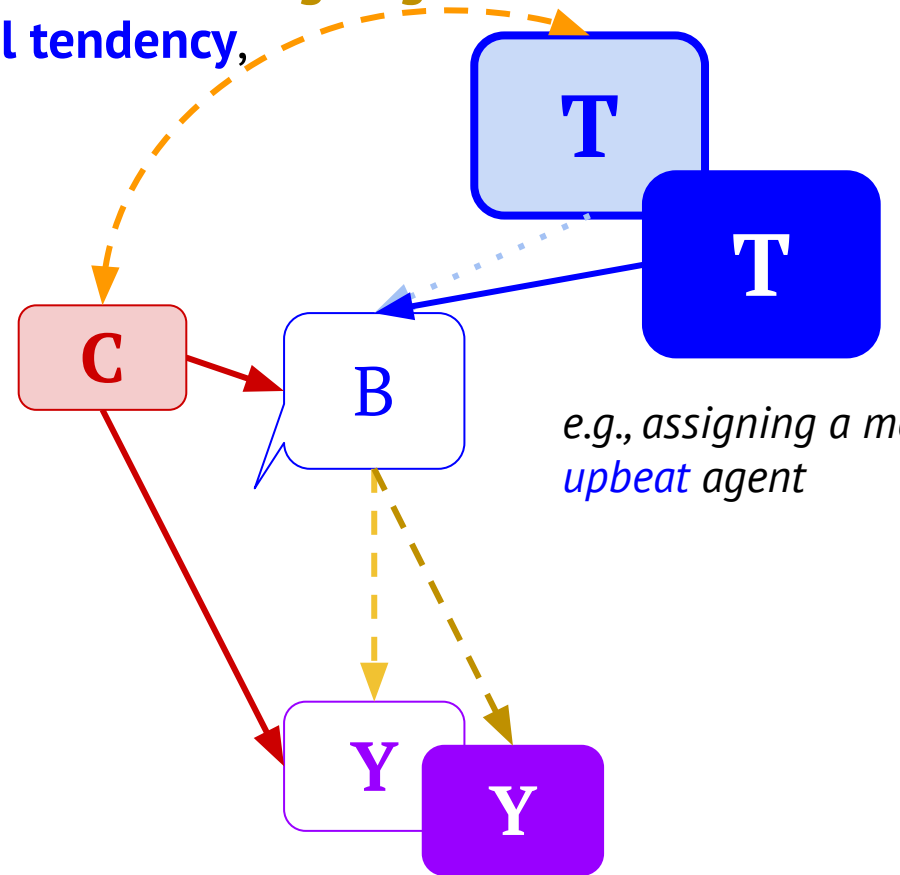
e.g., assigning a more upbeat agent

Our motivating task: **tendency-based assignment**

Causal inference problem: estimate **effect of assigning** an agent based on their **behavioural tendency**, as inferred from past conversations

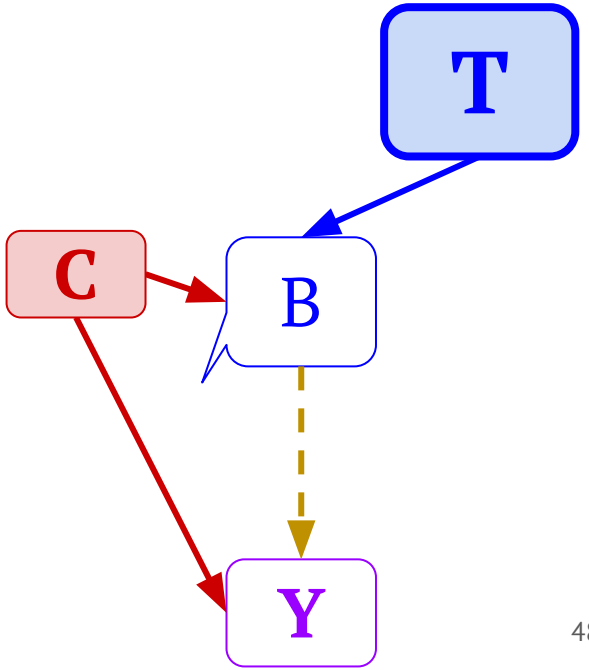
We want to say that two agents get different **outcomes** because their **tendencies** are different,

not because the **circumstances** of their conversations were different



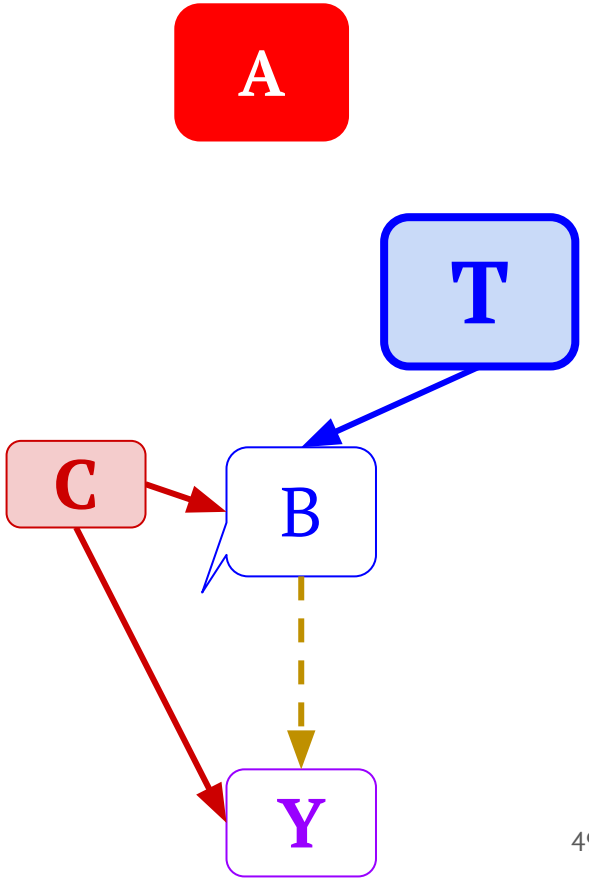
e.g., assigning a more upbeat agent

Challenge 1: **observed assignment**



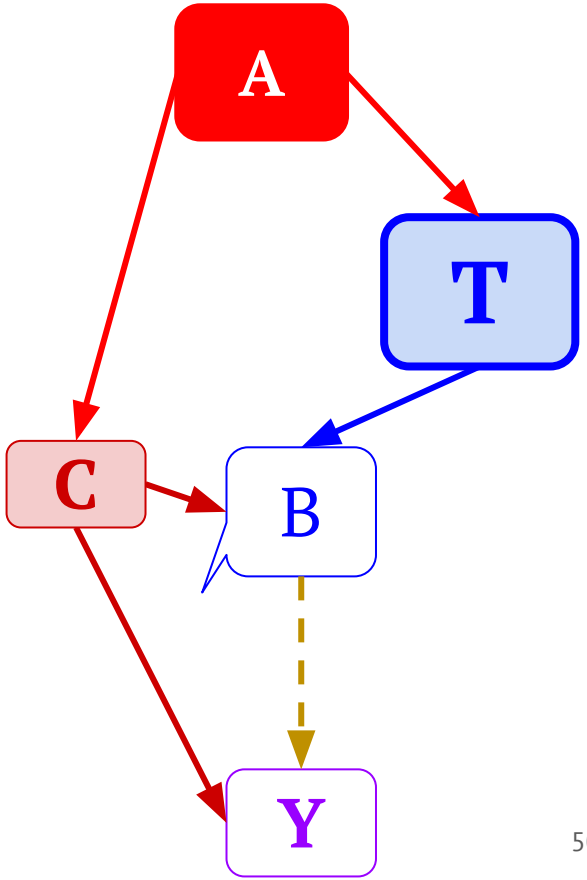
Challenge 1: **observed assignment**

Intuition: we can only observe outcomes in conversations that an agent is *assigned* to



Challenge 1: **observed assignment**

Intuition: we can only observe outcomes in conversations that an agent is *assigned* to

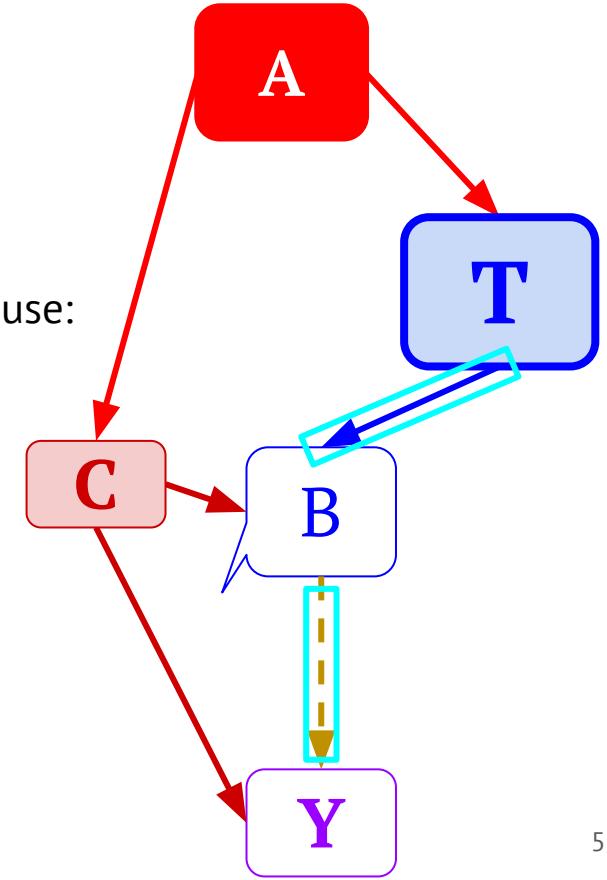


Challenge 1: **observed assignment**

Intuition: we can only observe outcomes in conversations that an agent is *assigned* to

We could observe that two **agents** get different outcomes because:

their **tendencies** are different



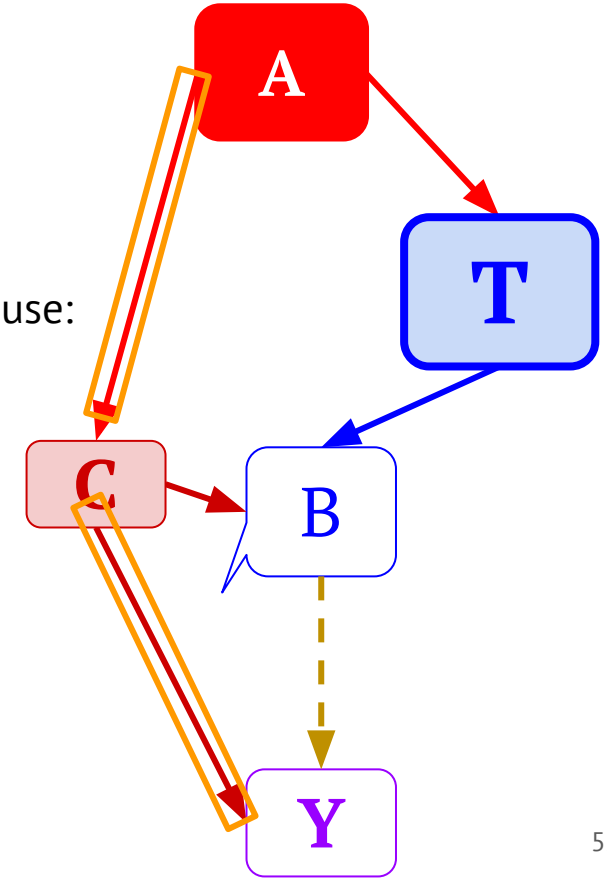
Challenge 1: **observed assignment**

Intuition: we can only observe outcomes in conversations that an agent is *assigned* to

We could observe that two **agents** get different outcomes because:

their **tendencies** are different,

or the **assigned circumstances** were different



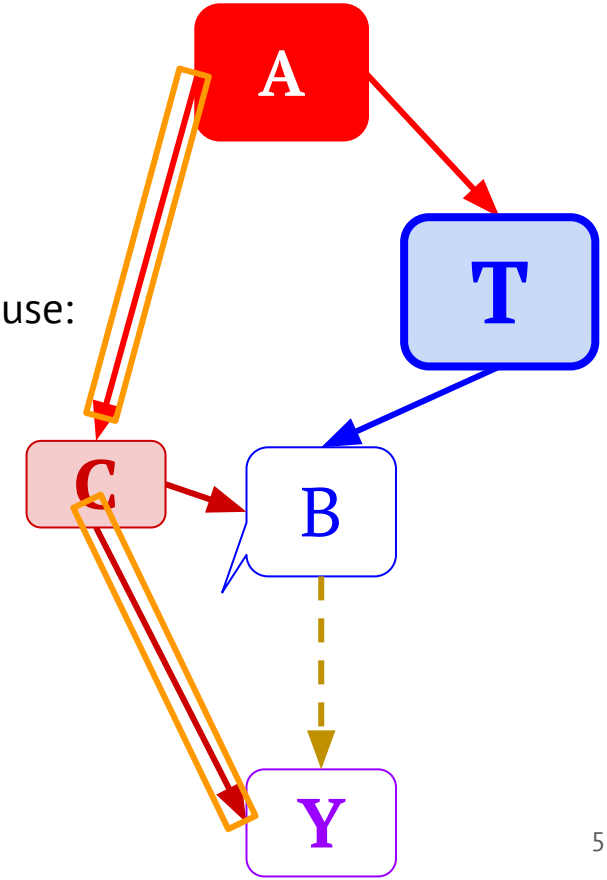
Challenge 1: **observed assignment**

Intuition: we can only observe outcomes in conversations that an agent is *assigned* to

We could observe that two **agents** get different outcomes because:

their **tendencies** are different,

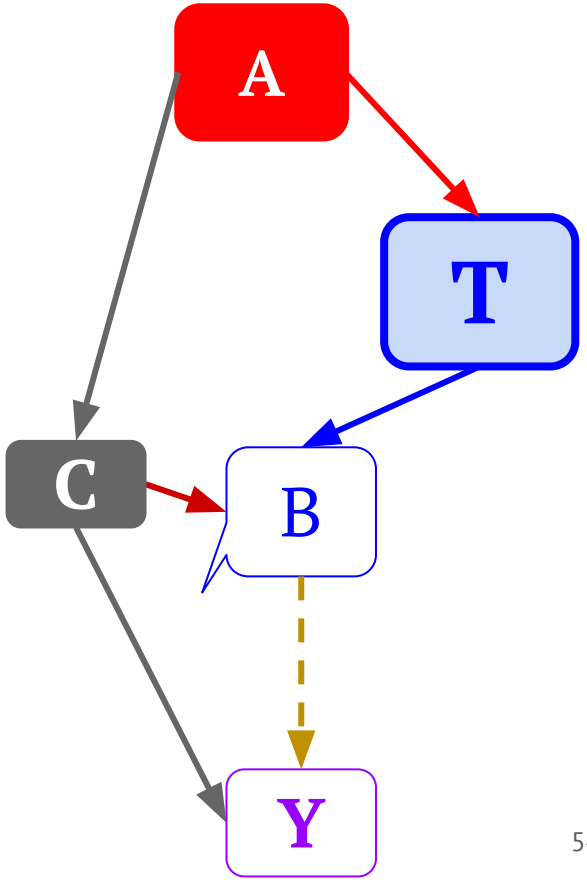
or the **assigned circumstances** were different



Goal: "break" this causal path

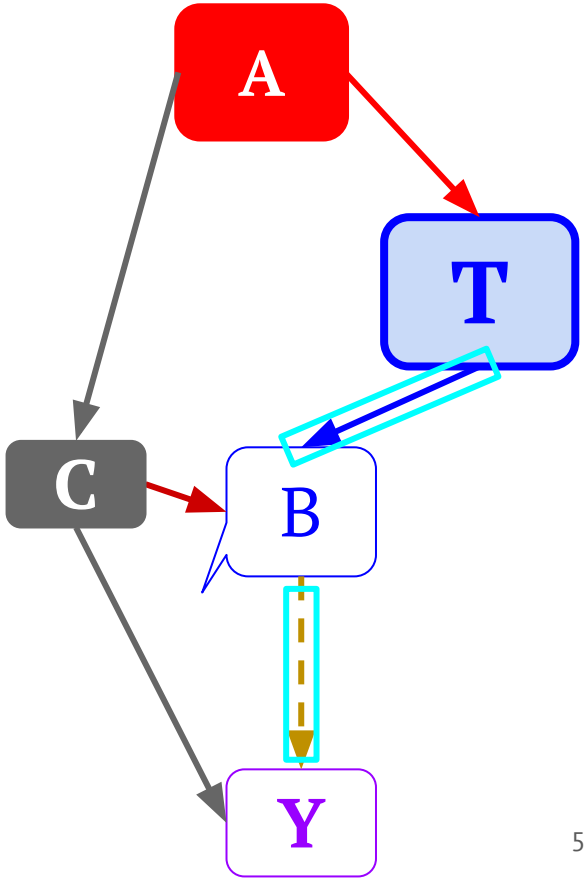
Challenge 1: **observed assignment**

Standard approach: **Control** for attributes of the circumstance
i.e., only compare agents that take conversations
in the same exact circumstances



Challenge 1: **observed assignment**

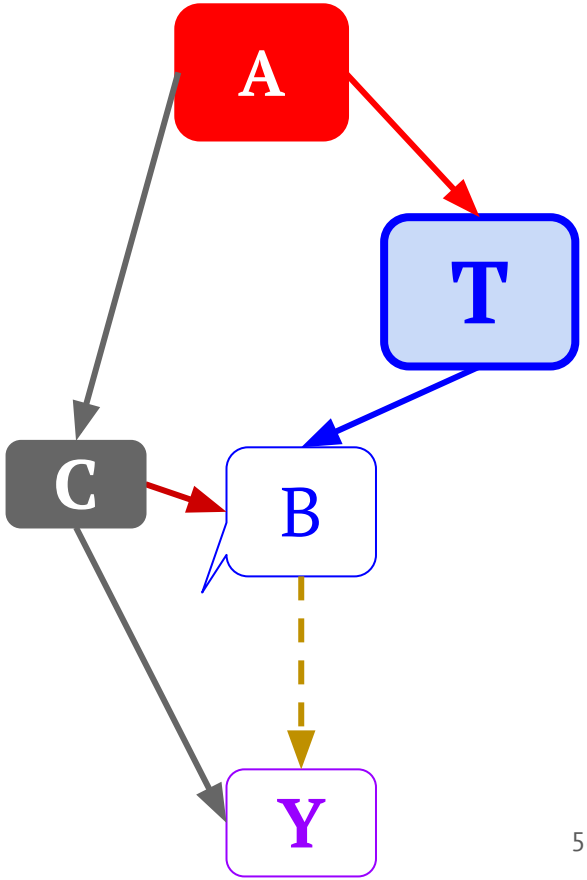
Standard approach: **Control** for attributes of the circumstance
i.e., only compare agents that take conversations
in the same exact circumstances



Challenge 1: **observed assignment**

Standard approach: **Control** for attributes of the circumstance
i.e., only compare agents that take conversations
in the same exact circumstances

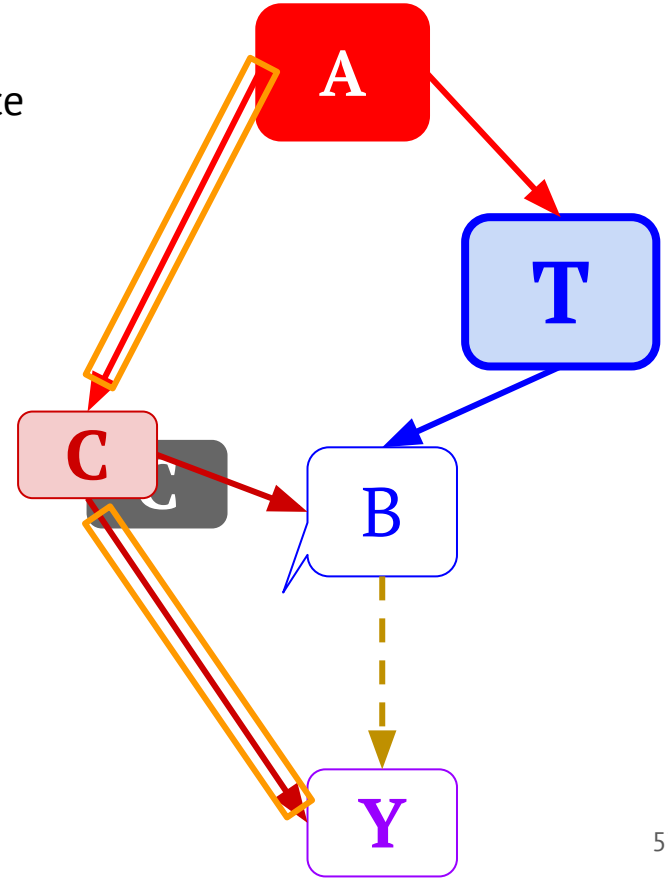
But: we can't control for **unobservable attributes**.



Challenge 1: **observed assignment**

Standard approach: **Control** for attributes of the circumstance
i.e., only compare agents that take conversations
in the same exact circumstances

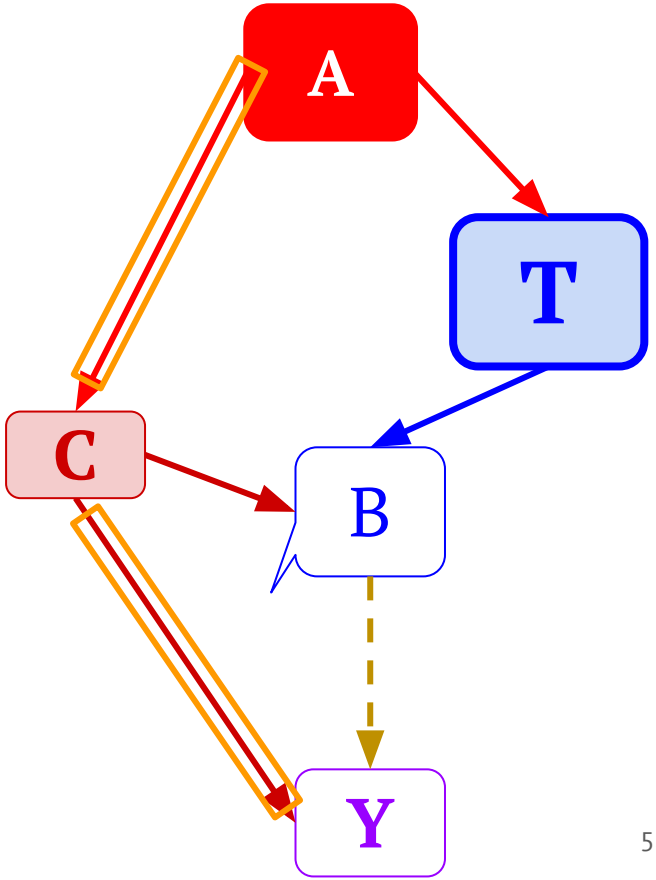
But: we can't control for **unobservable attributes**.



Challenge 1: **observed assignment**

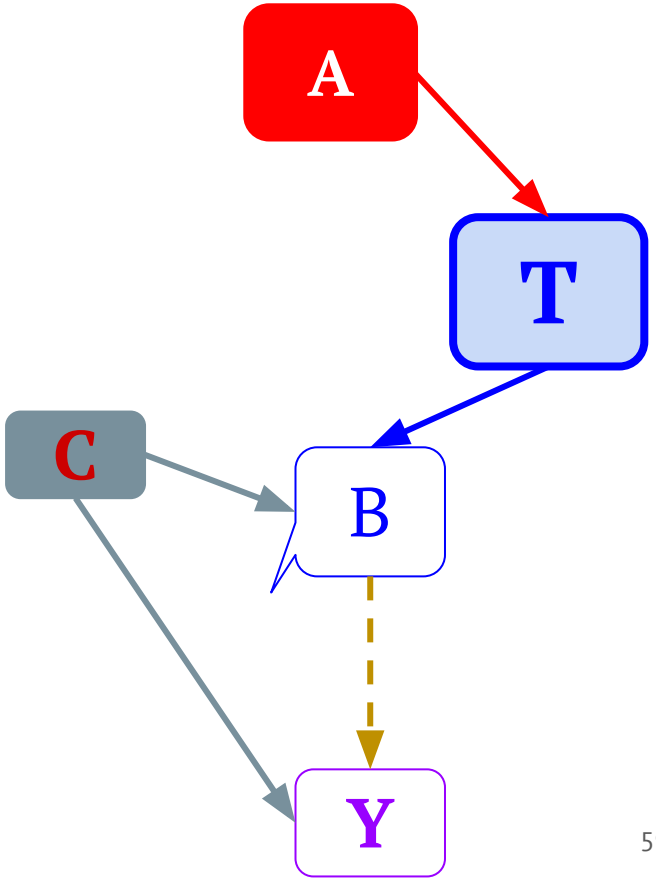
Wishful thinking: If only we had **random assignment**

,



Challenge 1: **observed assignment**

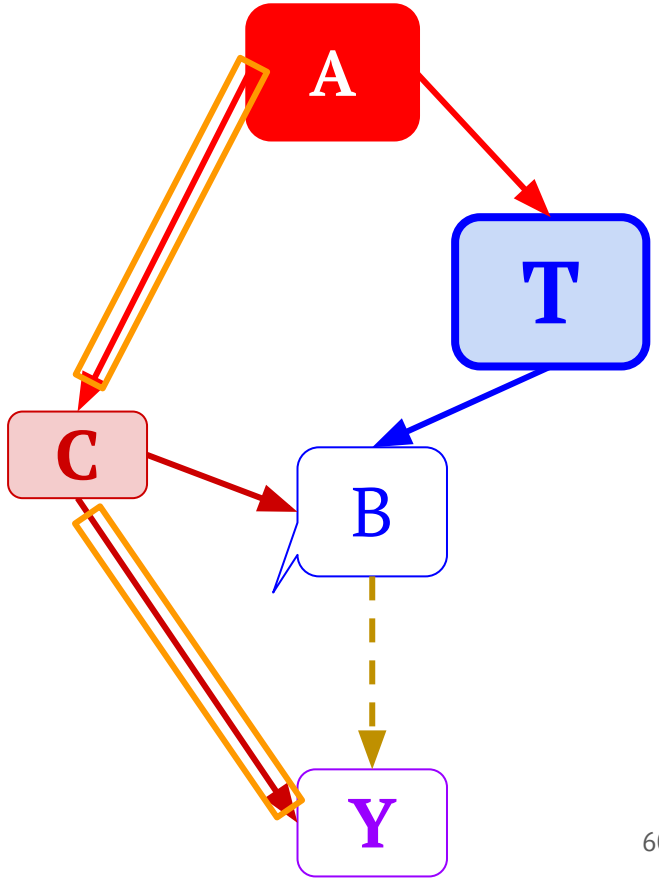
Wishful thinking: If only we had **random assignment**



Challenge 1: **observed assignment**

Wishful thinking: If only we had **random assignment**

But: we are in an observational setting…

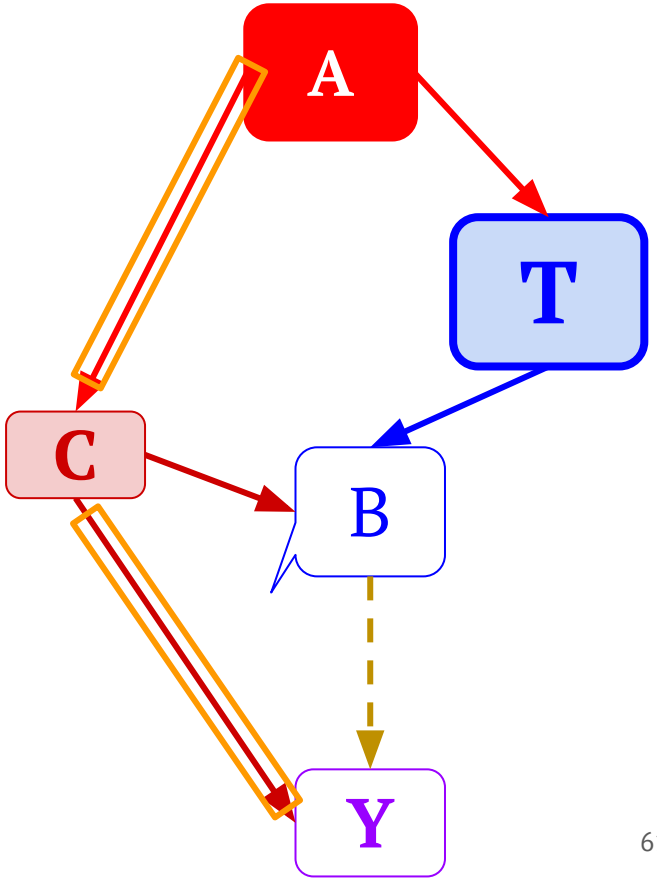


Challenge 1: **observed assignment**

Wishful thinking: If only we had **random assignment**

But: we are in an observational setting…

We could still exploit the assignment mechanism



Challenge 1: **observed assignment**

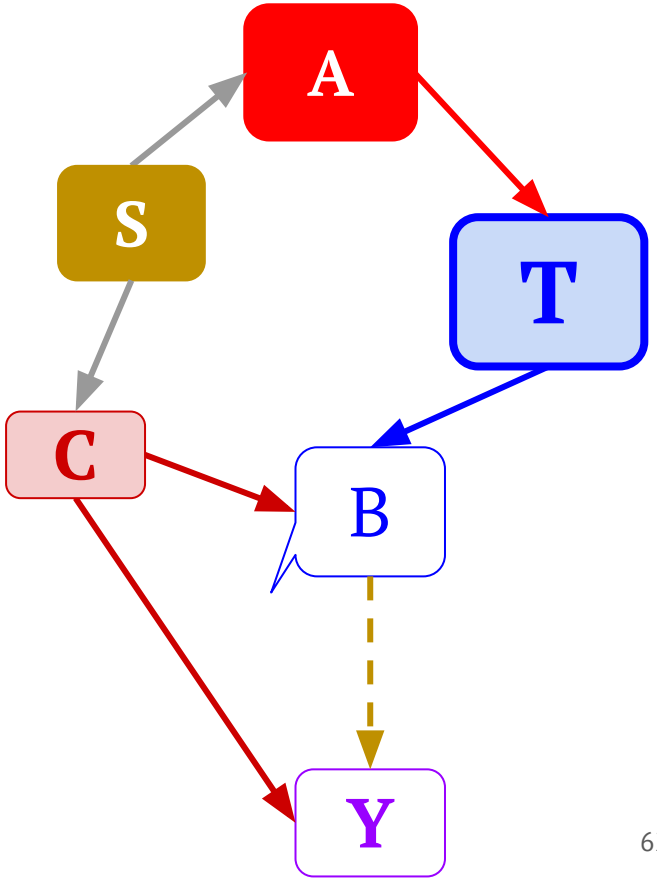
Wishful thinking: If only we had **random assignment**

But: we are in an observational setting…

We could still exploit the assignment mechanism

When assignment is governed by an accessible **selector variable S** (e.g., shift, location, department)…

e.g., crisis counselors are **randomly assigned within each shift**



Challenge 1: **observed assignment**

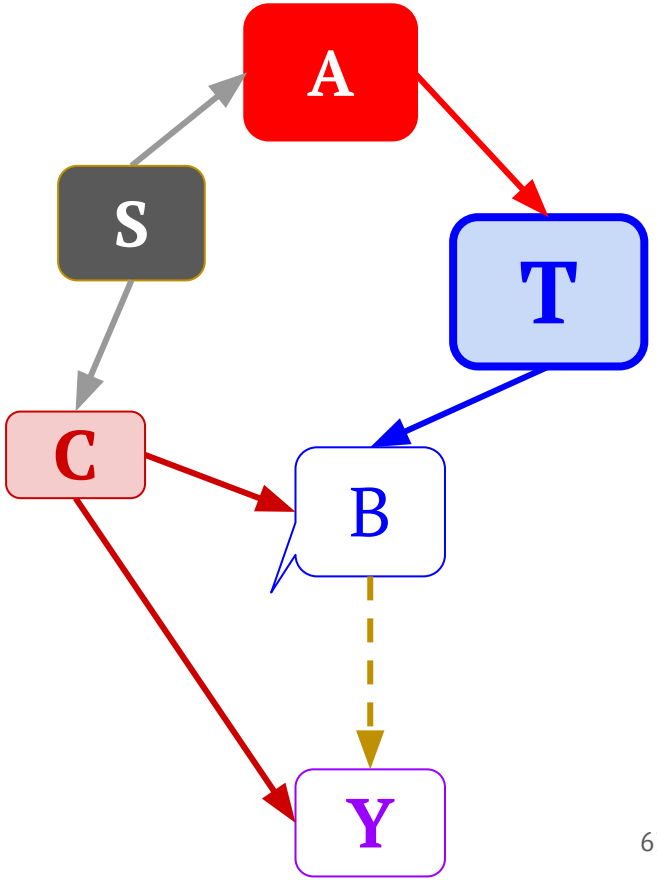
Wishful thinking: If only we had **random assignment**

But: we are in an observational setting…

We could still exploit the assignment mechanism

When assignment is governed by an accessible
selector variable S (e.g., shift, location, department)…
…controlling for S breaks the dependency on assignment

Just as good as random assignment
since **Y** and **A** are
conditionally independent
given **S** and **T**



Challenge 1: **observed assignment**

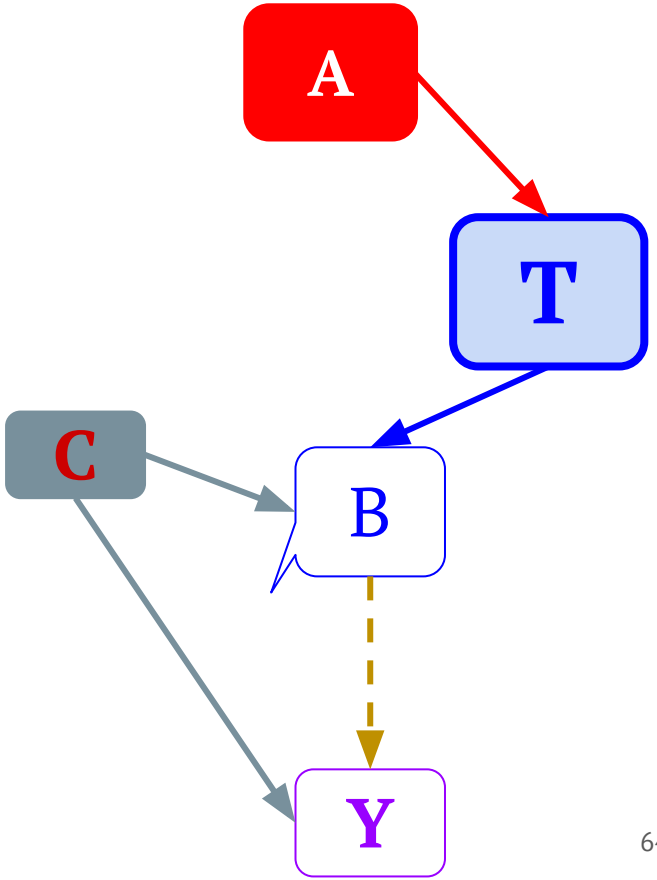
Wishful thinking: If only we had **random assignment**

But: we are in an observational setting…

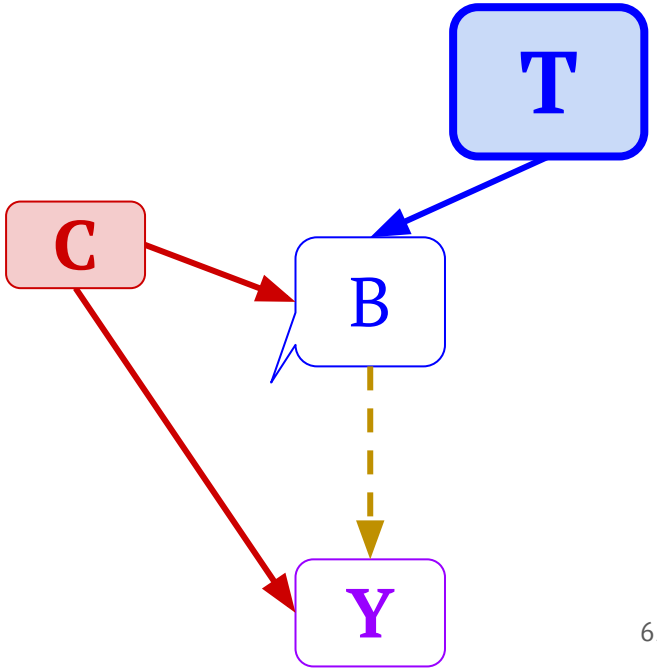
We could still exploit the assignment mechanism

When assignment is governed by an accessible
selector variable S (e.g., shift, location, department)…
…controlling for S breaks the dependency on assignment

Just as good as random assignment
since **Y** and **A** are
conditionally independent
given **S** and **T**

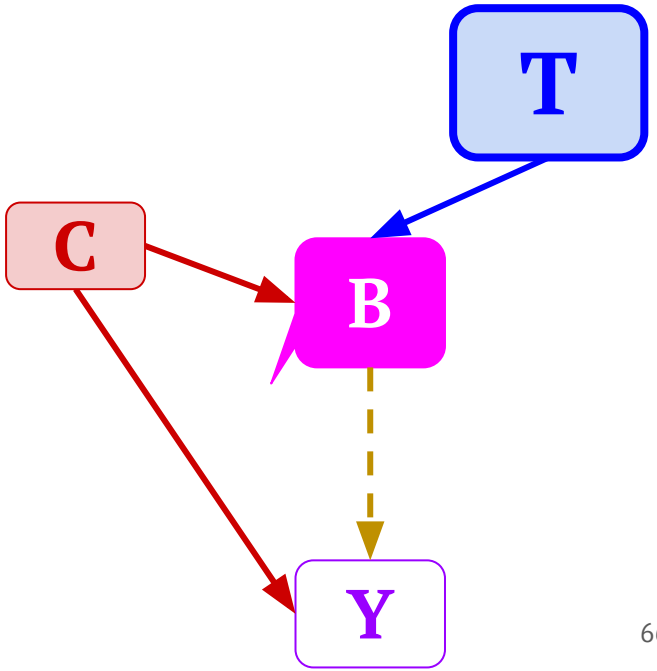


Challenge 2: **interactional effects**



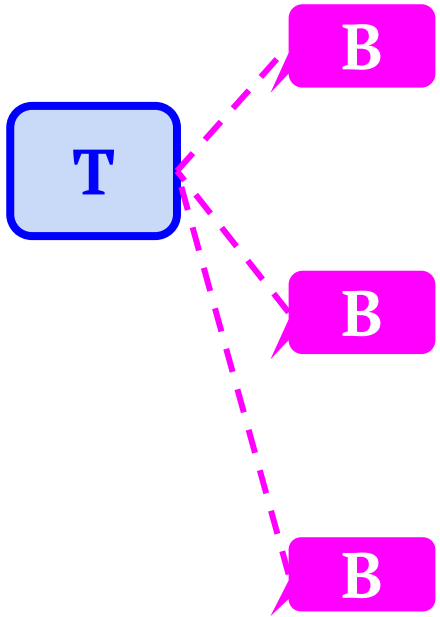
Challenge 2: **interactional effects**

Intuition: we infer an agent's tendencies using their **observed behaviours**.



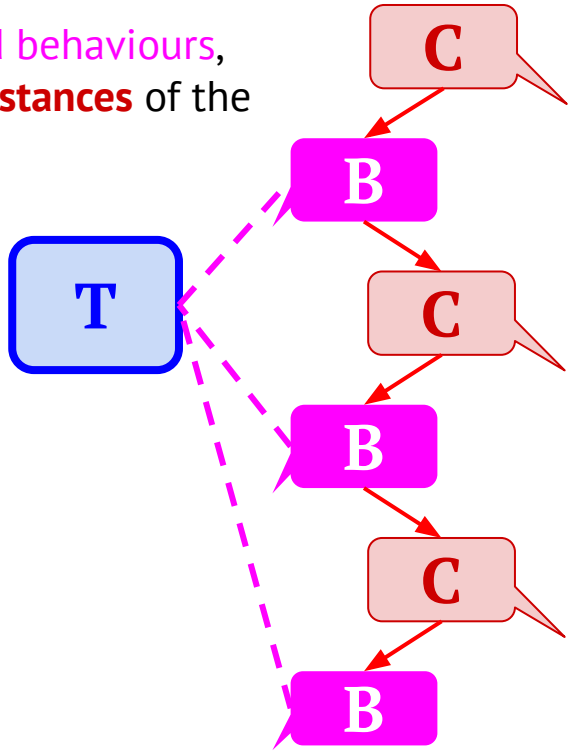
Challenge 2: **interactional effects**

Intuition: we infer an agent's tendencies using their **observed behaviours**.



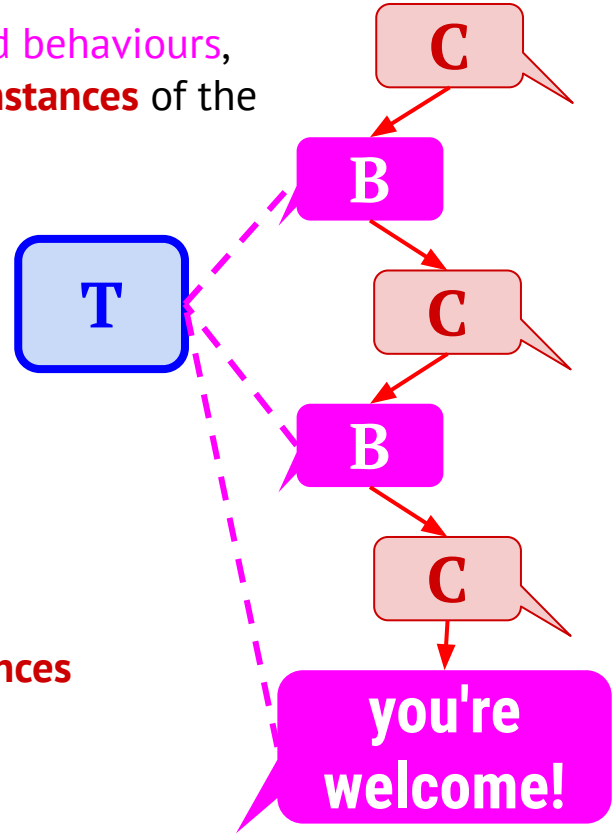
Challenge 2: **interactional effects**

Intuition: we infer an agent's tendencies using their observed behaviours, but these behaviours are entangled with the **circumstances** of the conversations in which they were observed



Challenge 2: **interactional effects**

Intuition: we infer an agent's tendencies using their observed behaviours, but these behaviours are entangled with the **circumstances** of the conversations in which they were observed



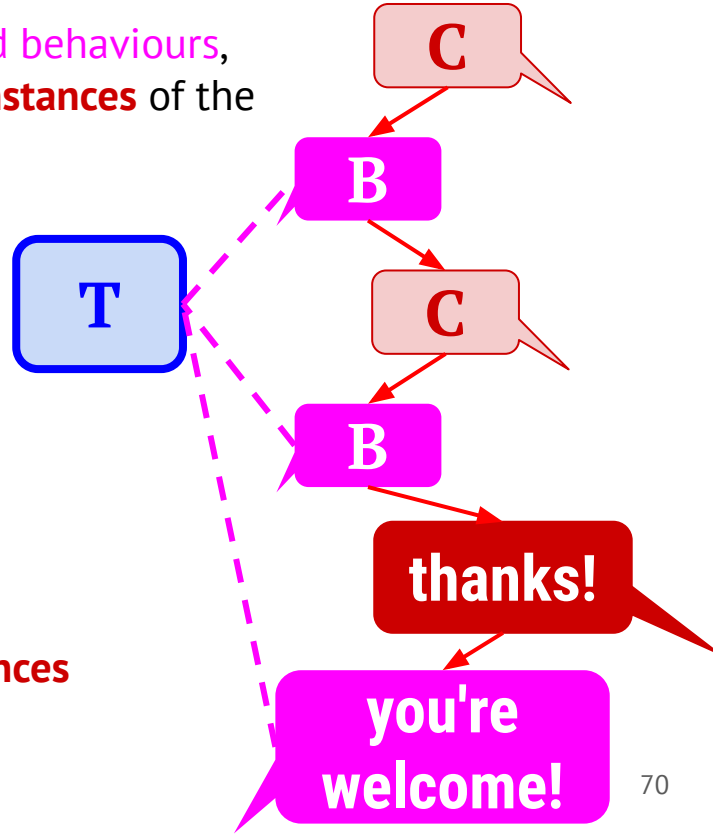
We could observe that **T** leads to better outcomes because:

that **tendency** is more effective,

or the **behaviours** reflect favorable **conversational circumstances**

Challenge 2: **interactional effects**

Intuition: we infer an agent's tendencies using their observed behaviours, but these behaviours are entangled with the **circumstances** of the conversations in which they were observed



We could observe that **T** leads to better outcomes because:

that **tendency** is more effective,

or the **behaviours** reflect favorable **conversational circumstances** (e.g., client already satisfied)

Challenge 2: **interactional effects**

Wishful thinking: Don’t infer the tendencies based on observed behaviors

Challenge 2: **interactional effects**

Wishful thinking: Don’t infer the tendencies based on observed behaviors

Where from then?

Challenge 2: **interactional effects**

Wishful thinking: Don't infer the tendencies based on observed behaviors

Where from then? From "other" conversations of the same agent

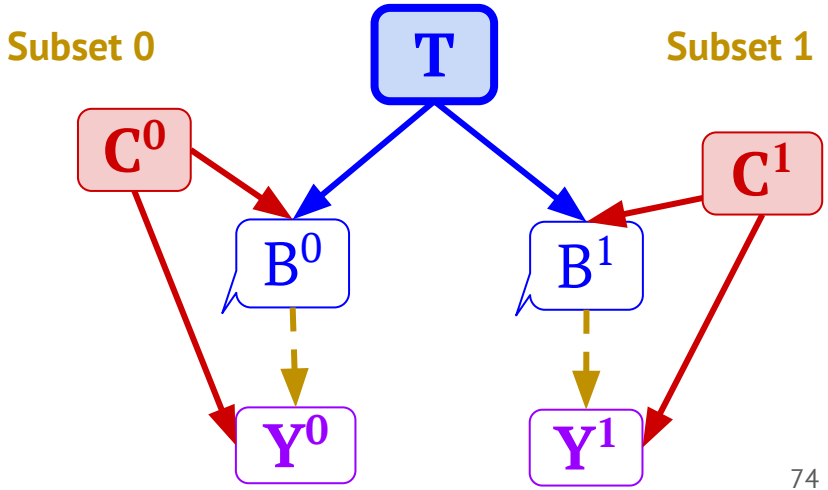
Challenge 2: **interactional effects**

Wishful thinking: Don't infer the tendencies based on observed behaviors

Where from then? From "other" conversations of the same agent

Suppose the agent take ***many*** conversations
with different clients in different situations.

We split their conversations into **two subsets**



Challenge 2: **interactional effects**

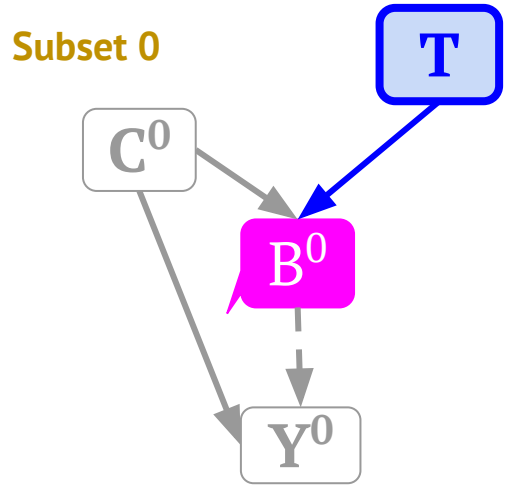
Wishful thinking: Don't infer the tendencies based on observed behaviors

Where from then? From "other" conversations of the same agent

Suppose the agent take ***many*** conversations
with different clients in different situations.

We split their conversations into **two subsets**

We measure their **tendencies** on **one subset**



Challenge 2: **interactional effects**

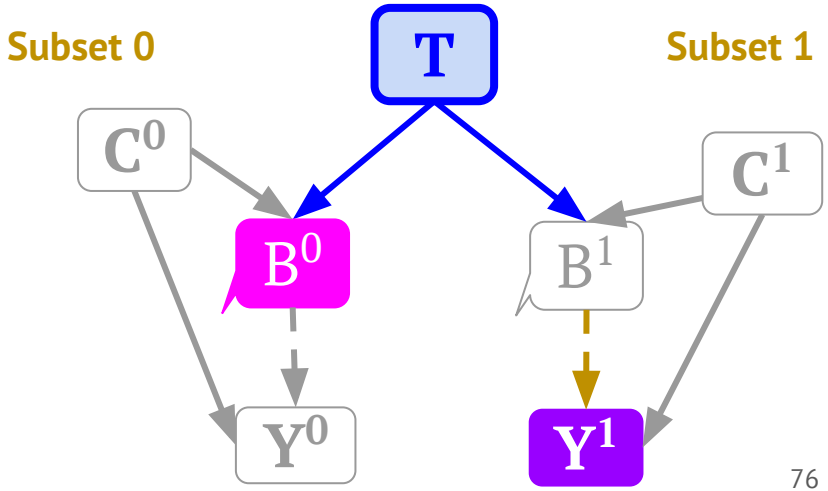
Wishful thinking: Don't infer the tendencies based on observed behaviors

Where from then? From "other" conversations of the same agent

Suppose the agent take ***many*** conversations
with different clients in different situations.

We split their conversations into **two subsets**

We measure their **tendencies** on **one subset**
and **outcomes** on the other subset



Challenge 2: **interactional effects**

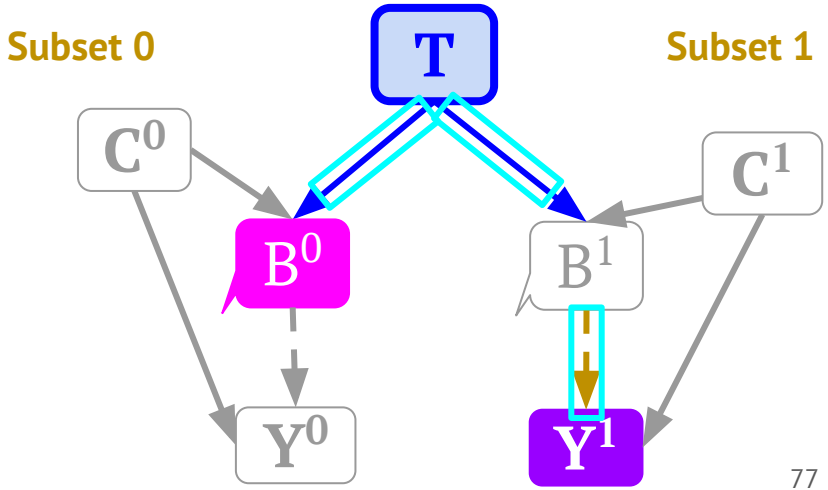
Wishful thinking: Don't infer the tendencies based on observed behaviors

Where from then? From "other" conversations of the same agent

Suppose the agent take ***many*** conversations
with different clients in different situations.

We split their conversations into **two subsets**

We measure their **tendencies** on **one subset**
and **outcomes** on the other subset



Challenge 2: **interactional effects**

Wishful thinking: Don't infer the tendencies based on observed behaviors

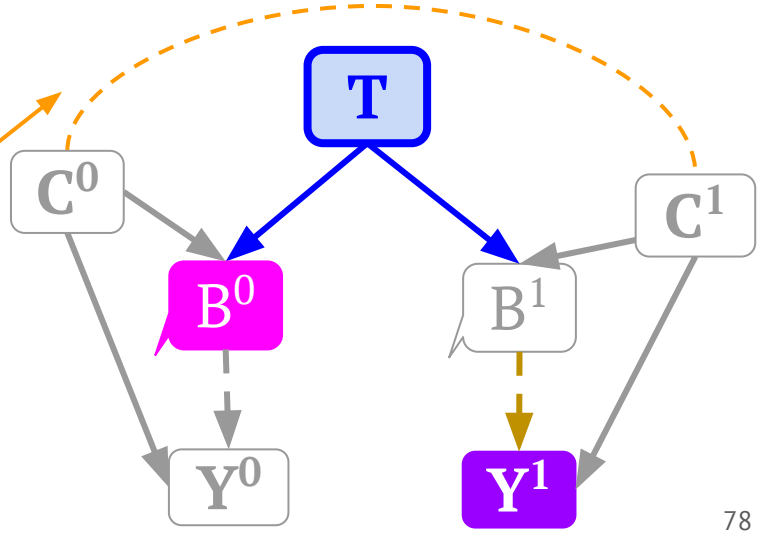
Where from then? From "other" conversations of the same agent

Suppose the agent take ***many*** conversations
with different clients in different situations.

We split their conversations into **two subsets**

We measure their **tendencies** on **one subset**
and **outcomes** on the other subset

Just as good as having pre-established tendencies
provided circumstances of conversations
aren't related to each other



What do we gain from our concrete description?

What do we gain from our concrete description?

More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

What do we gain from our concrete description?

More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

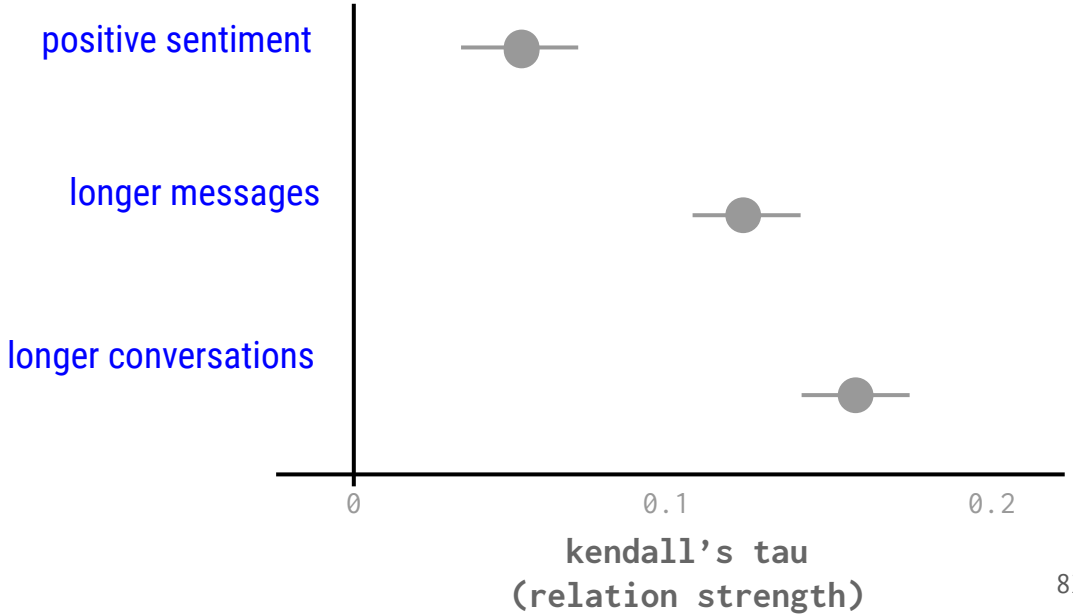
positive sentiment

longer messages

longer conversations

What do we gain from our concrete description?

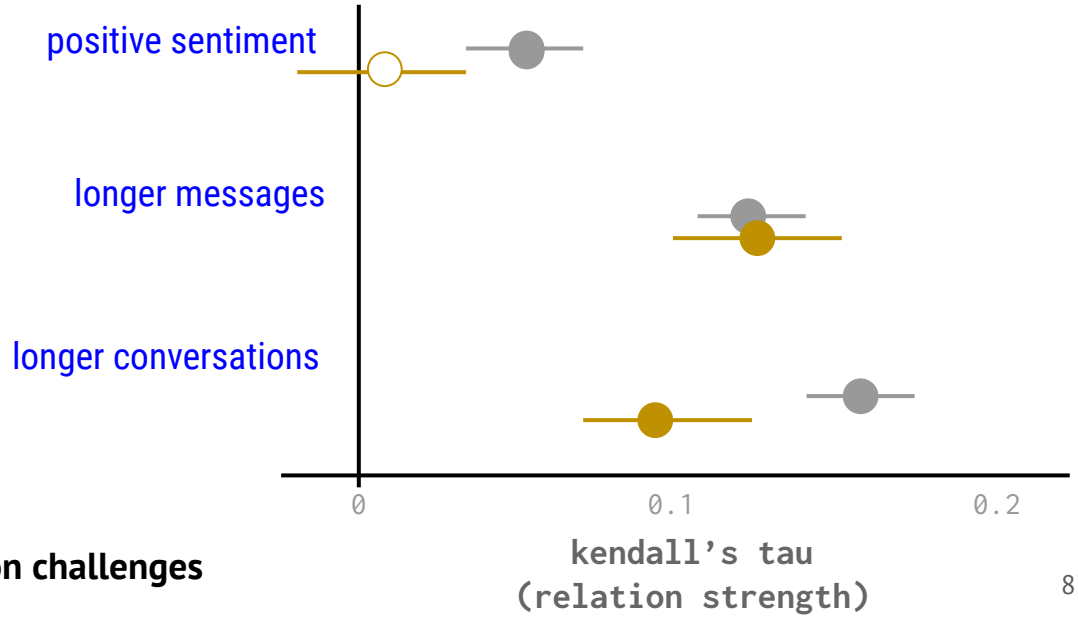
More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:



Simple counselor-level correlations

What do we gain from our concrete description?

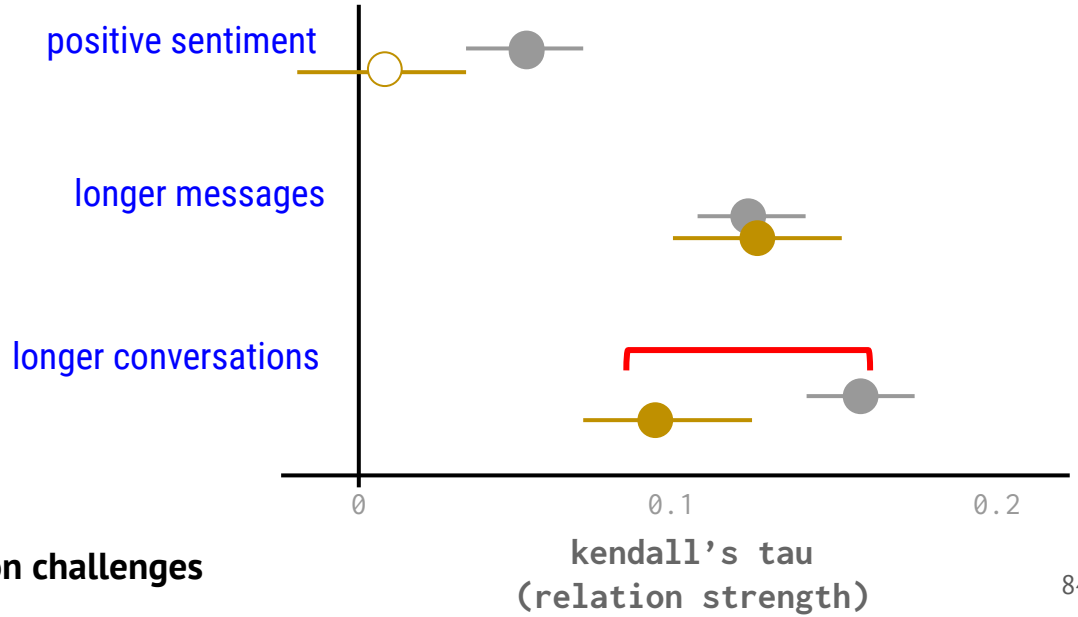
More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:



- Simple counselor-level correlations
- **Addressing assignment and interaction challenges**

What do we gain from our concrete description?

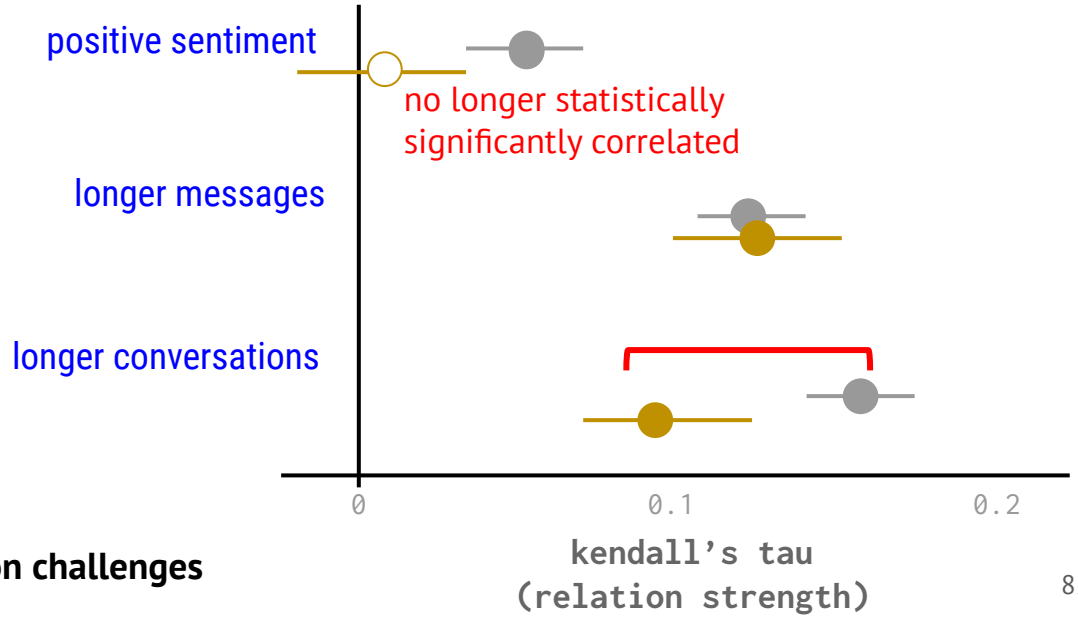
More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:



- Simple counselor-level correlations
- **Addressing assignment and interaction challenges**

What do we gain from our concrete description?

More **precise** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:



- Simple counselor-level correlations
- **Addressing assignment and interaction challenges**

What do we gain from our concrete description?

More **actionable** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

What do we gain from our concrete description?

More **actionable** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

Back-of-the-envelope estimate of the potential impact of a tendency-based assignment policy:

What do we gain from our concrete description?

More **actionable** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

Back-of-the-envelope estimate of the potential impact of a tendency-based assignment policy:

current % of positively rated conversations: **88%**

What do we gain from our concrete description?

More **actionable** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

Back-of-the-envelope estimate of the potential impact of a tendency-based assignment policy:

current % of positively rated conversations: **88%**

estimated % under tendency-based assignment policy: **92%****

****Wilcoxon p<0.01; improved (counterfactual) outcome in 74% of shifts**

What do we gain from our concrete description?

More **actionable** inferences about the **relation** between **tendencies** and outcome, i.e., **helpfulness rating**:

Back-of-the-envelope estimate of the potential impact of a tendency-based assignment policy:

current % of positively rated conversations: **88%**

estimated % under tendency-based assignment policy: **92%****

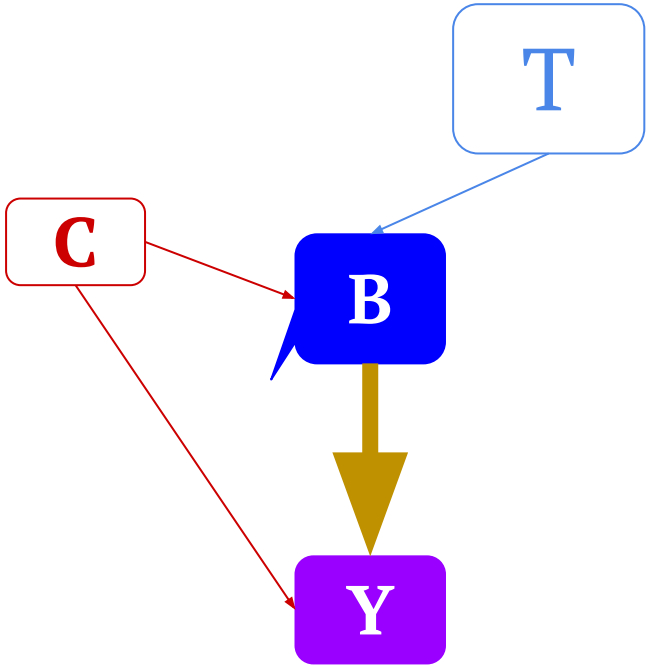
estimated % under historical efficiency assignment policy: **90%**

****Wilcoxon p<0.01; improved (counterfactual) outcome in 74% of shifts**

What we did **not** gain from our concrete description?

Causal effects of individual **actions in a conversation**

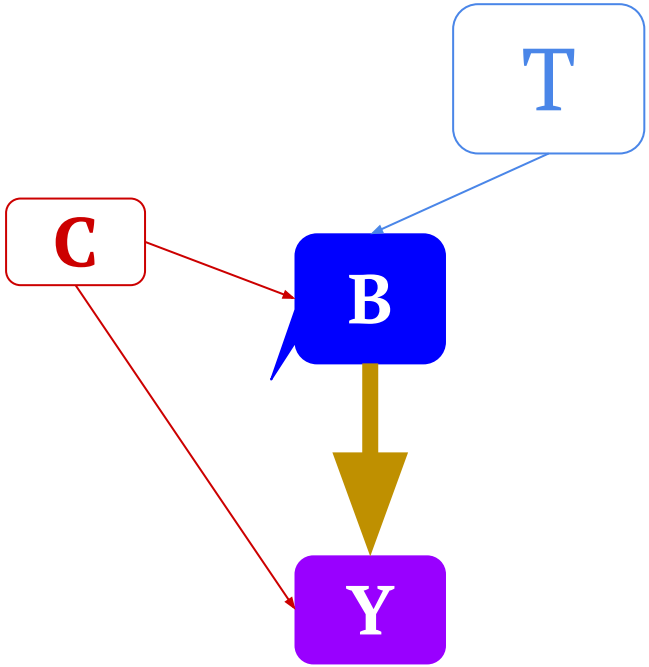
needed for training policies or assistance

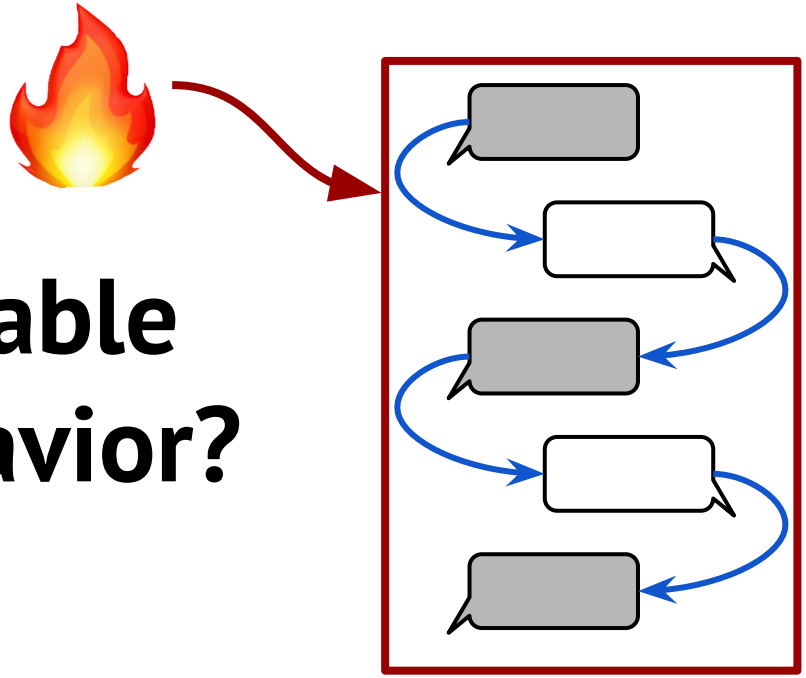


What we did **not** gain from our concrete description?

Causal effects of individual **actions in a conversation**

needed for training policies or assistance



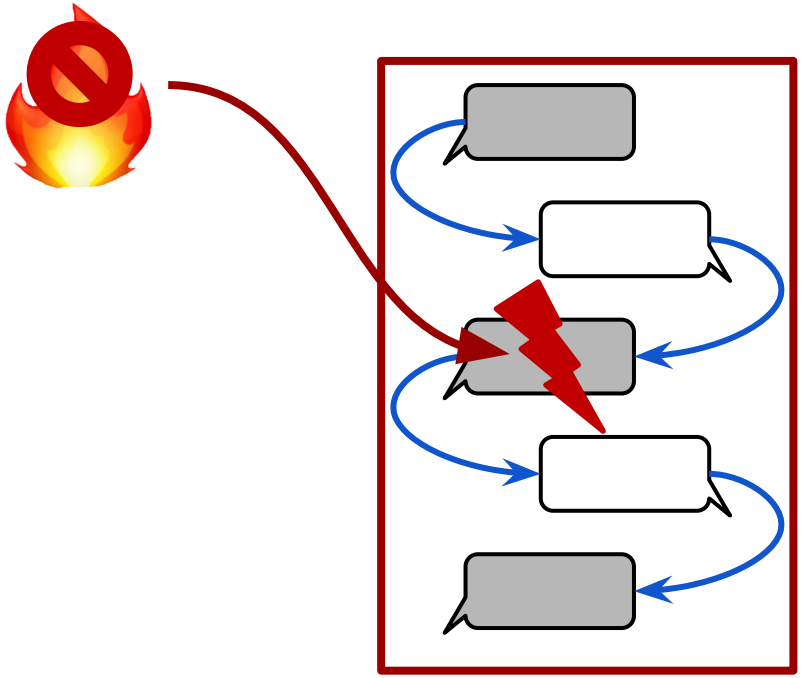


What leads to desirable conversational behavior?

Content removal as a moderation strategy
Kumar Bhargav Srinivasan,Cristian Danescu-Niculescu-Mizil,
Lillian Lee, Chenhao Tan
Proceedings of CSCW, 2019

Empirical setting:
Moderation of online discussions

r/ChangeMyView moderators
manually removed ***23k***
comments over 3 years

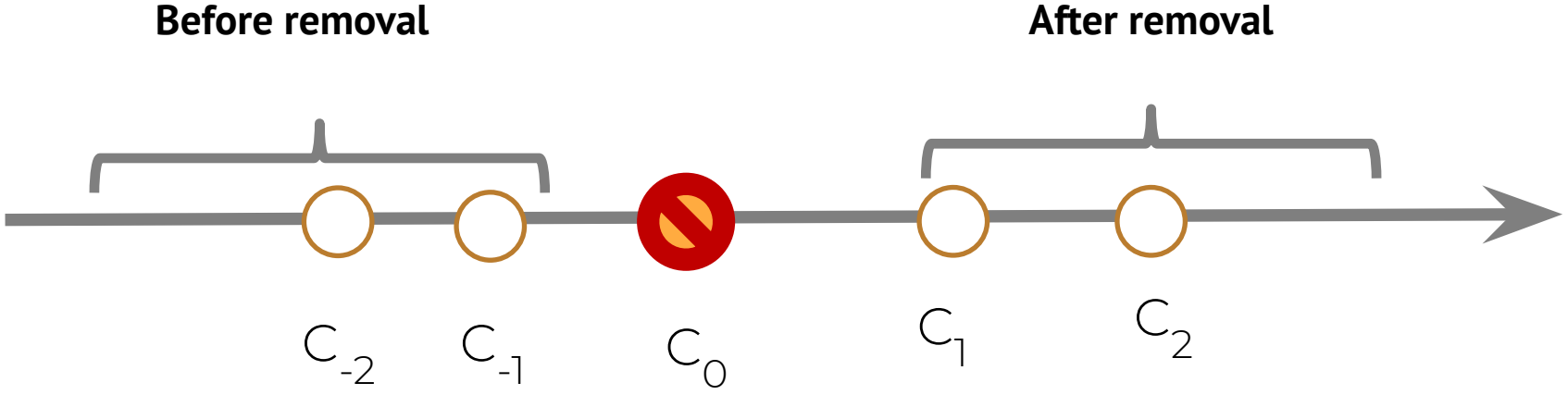


Does the removal of a problematic comment (by a moderator)
lead to a compliant* behavior (by the author of the removed comment)?

*: compliant ≠ better (moderation can be both good and bad)

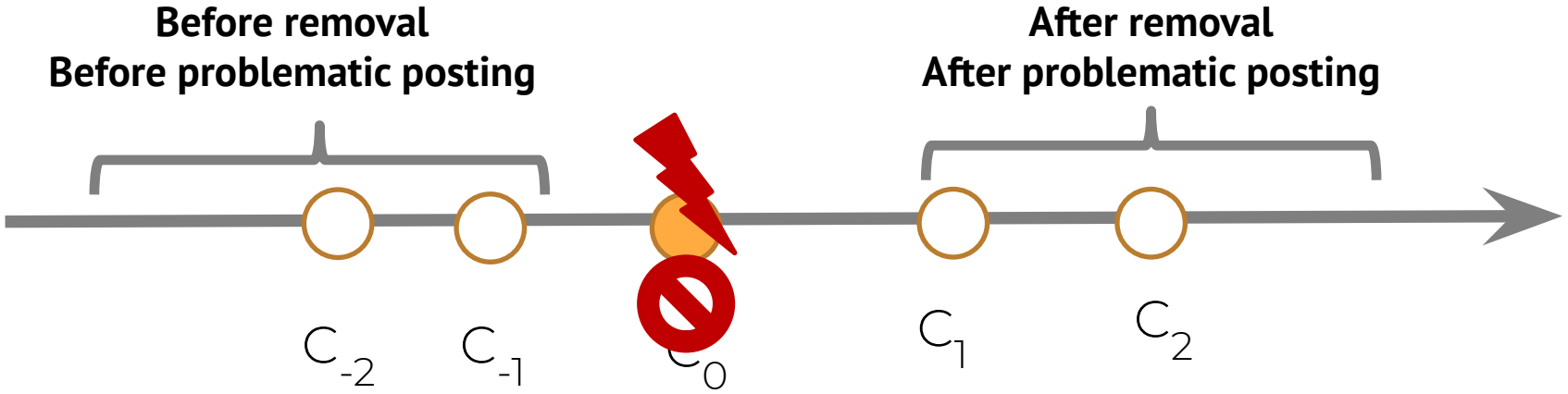
Does removal of a noncompliant message (by a moderator) lead to compliant behavior (by the author of the removed comment)?

Standard attempt: compare behavior before removal with behavior after removal, i.e., interrupted time-series approach



Does removal of a noncompliant message (by a moderator) lead to compliant behavior (by the author of the removed comment)?

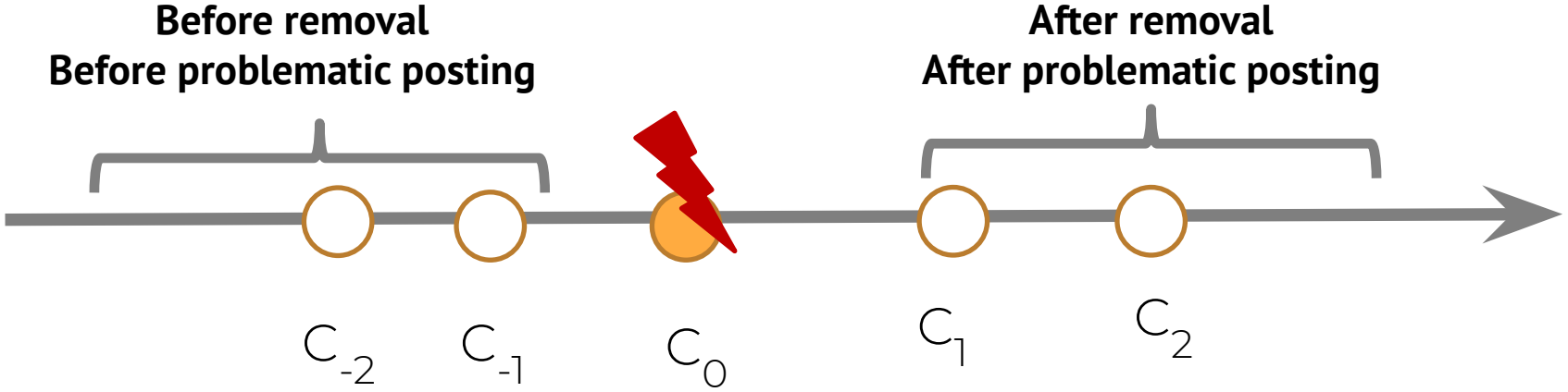
Standard attempt: compare behavior before removal with behavior after removal, i.e., interrupted time-series approach



But: the removal and the circumstances comment are confounded

Does removal of a noncompliant message (by a moderator) lead to compliant behavior (by the author of the removed comment)?

Standard attempt: compare behavior before removal with behavior after removal, i.e., interrupted time-series approach



But: the removal and the circumstances comment are confounded
Missing counterfactual: problematic comments that are **not** removed

Challenge 3: missing counterfactual

Wishful thinking: find situations in which the problematic comment is **not** removed

Challenge 3: missing counterfactual

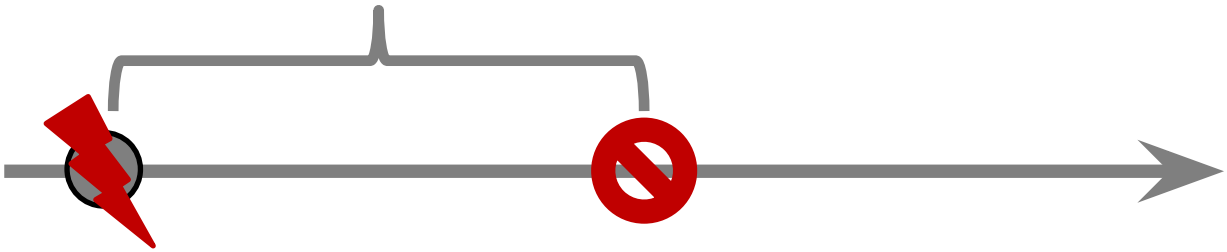
Wishful thinking: find situations in which the **same** problematic comment is **not** removed

Challenge 3: missing counterfactual

Wishful thinking: find situations in which the **same** problematic comment is **not** removed

Observation: removals don't happen immediately

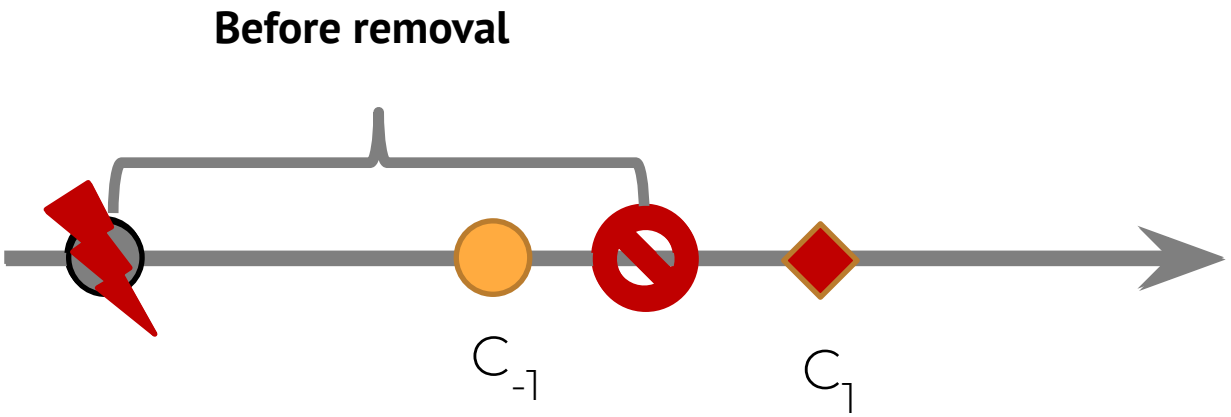
Delay between posting and removal
>2 hours in 40% of cases



Challenge 3: missing counterfactual

Wishful thinking: find situations in which the **same** problematic comment is **not** removed

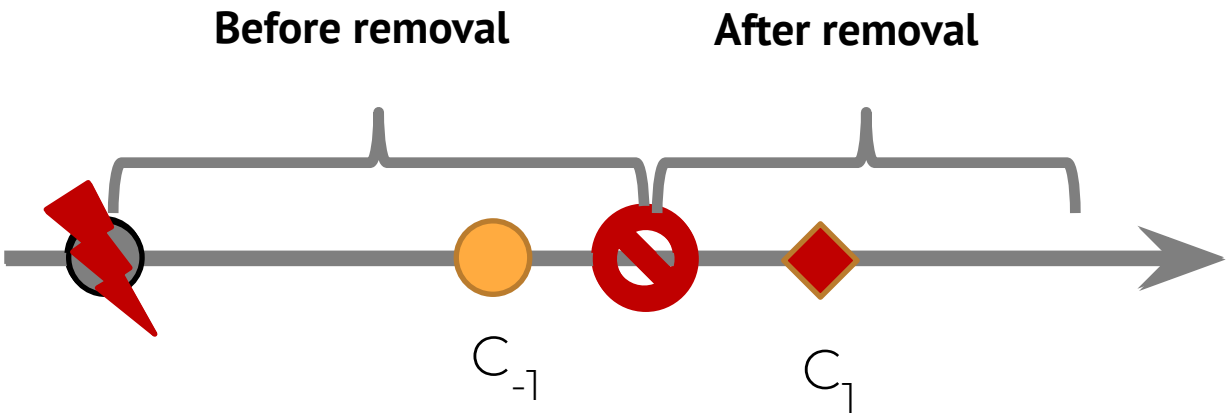
Observation: removals don't happen immediately



Challenge 3: missing counterfactual

Wishful thinking: find situations in which the **same** problematic comment is **not** removed

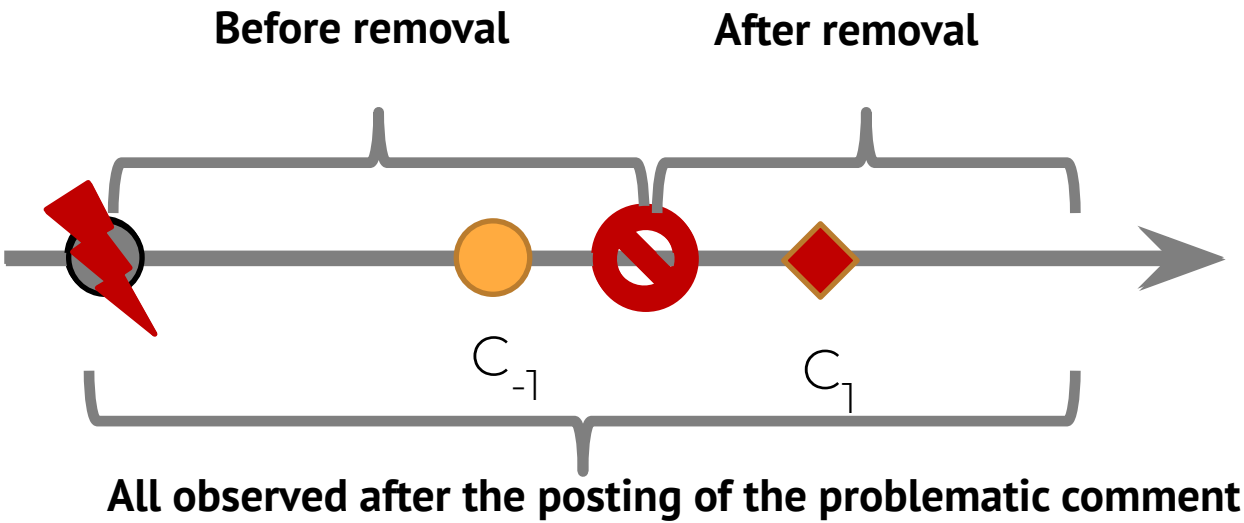
Observation: removals don't happen immediately

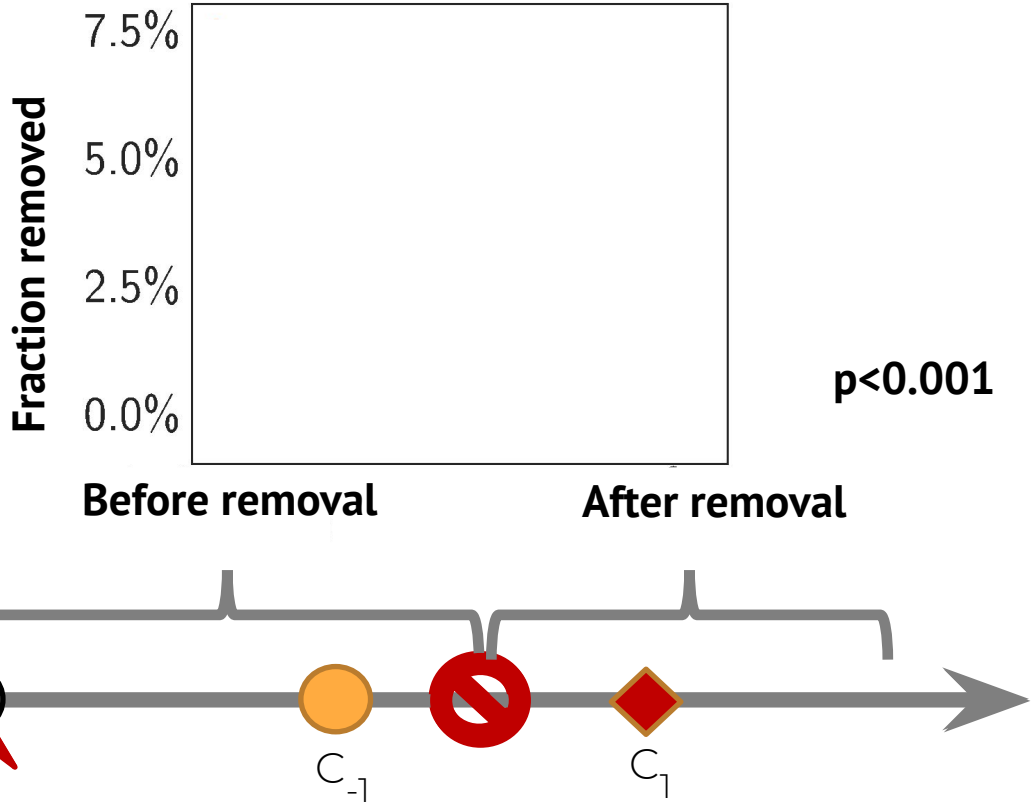


Challenge 3: missing counterfactual

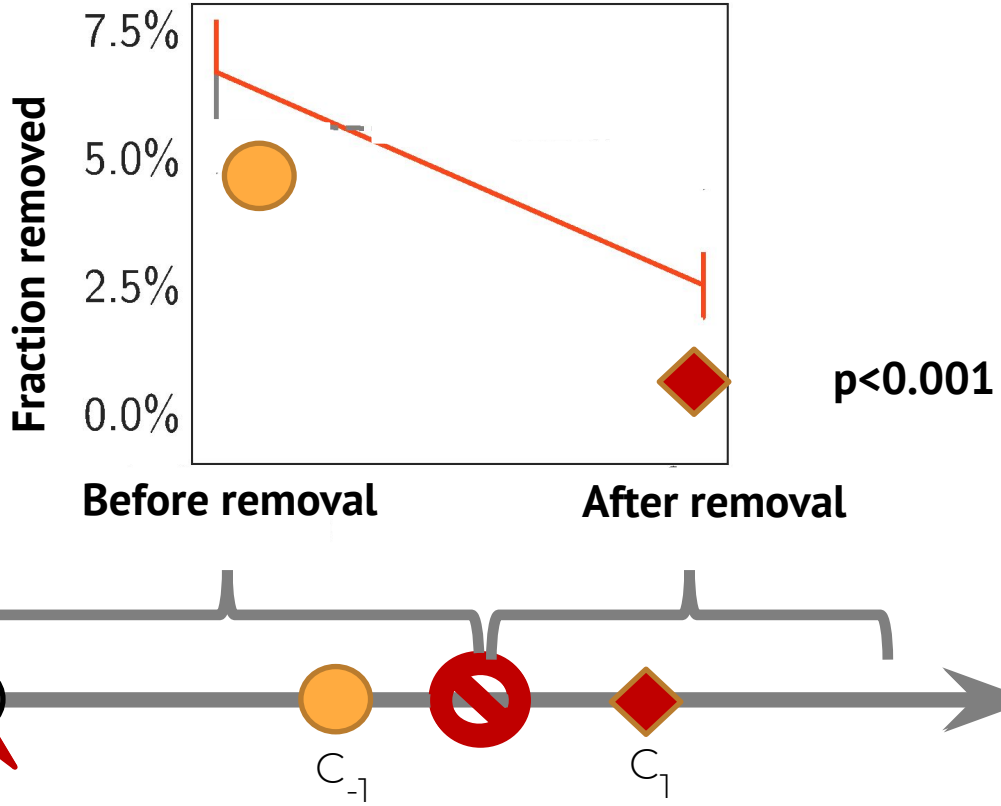
Wishful thinking: find situations in which the **same** problematic comment is **not** removed

Observation: removals don't happen immediately



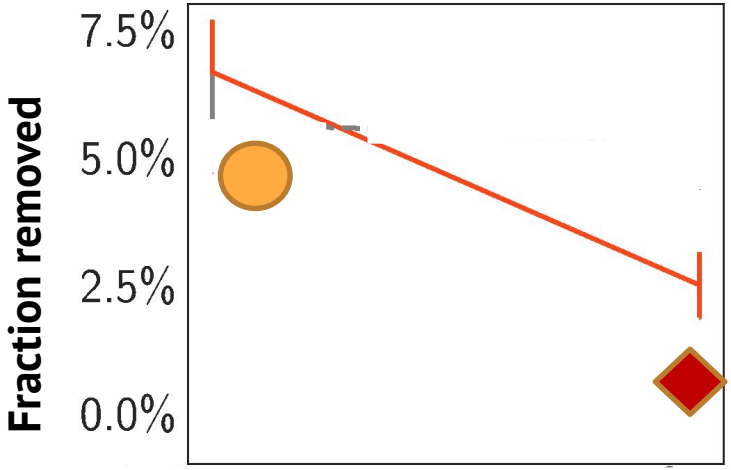


Does removal of a noncompliant message lead to compliant behavior?

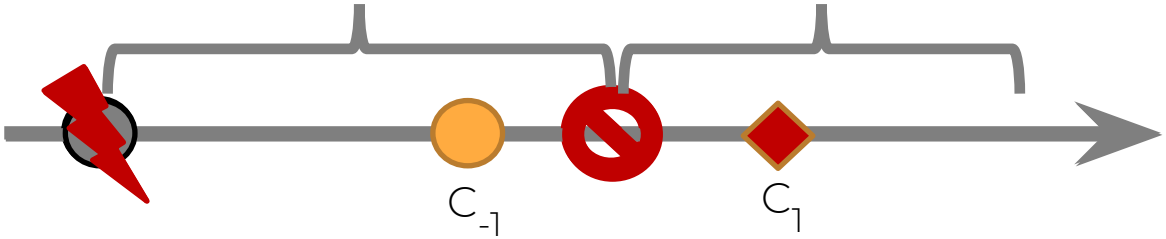


Does removal of a noncompliant message lead to compliant behavior?

Looks like it.

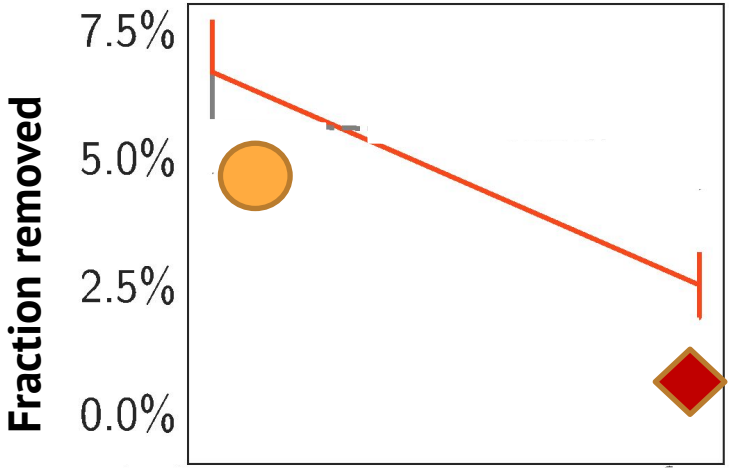


p<0.001

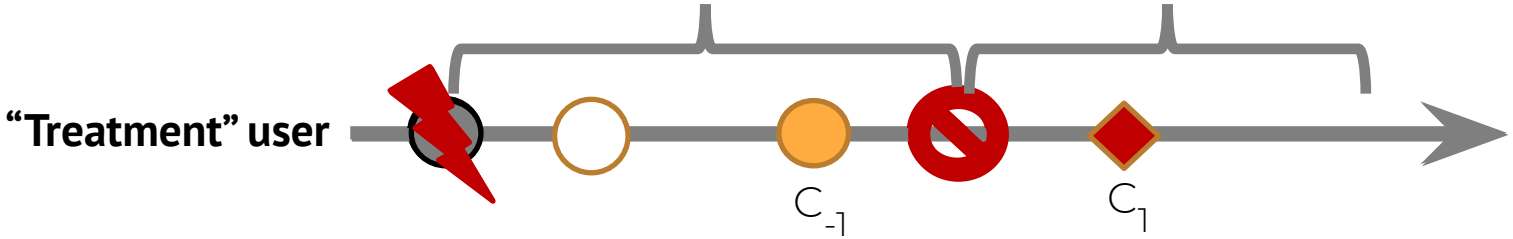


Does removal of a noncompliant message lead to compliant behavior?

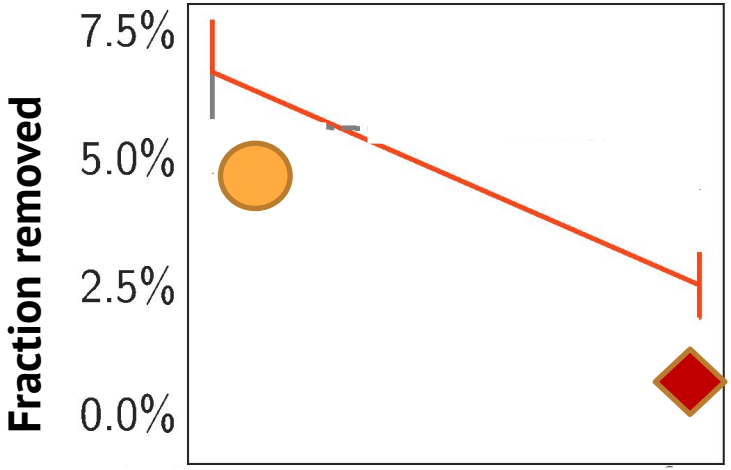
Looks like it. Or are these just temporal effects?


Fraction removed
7.5%
5.0%
2.5%
0.0%

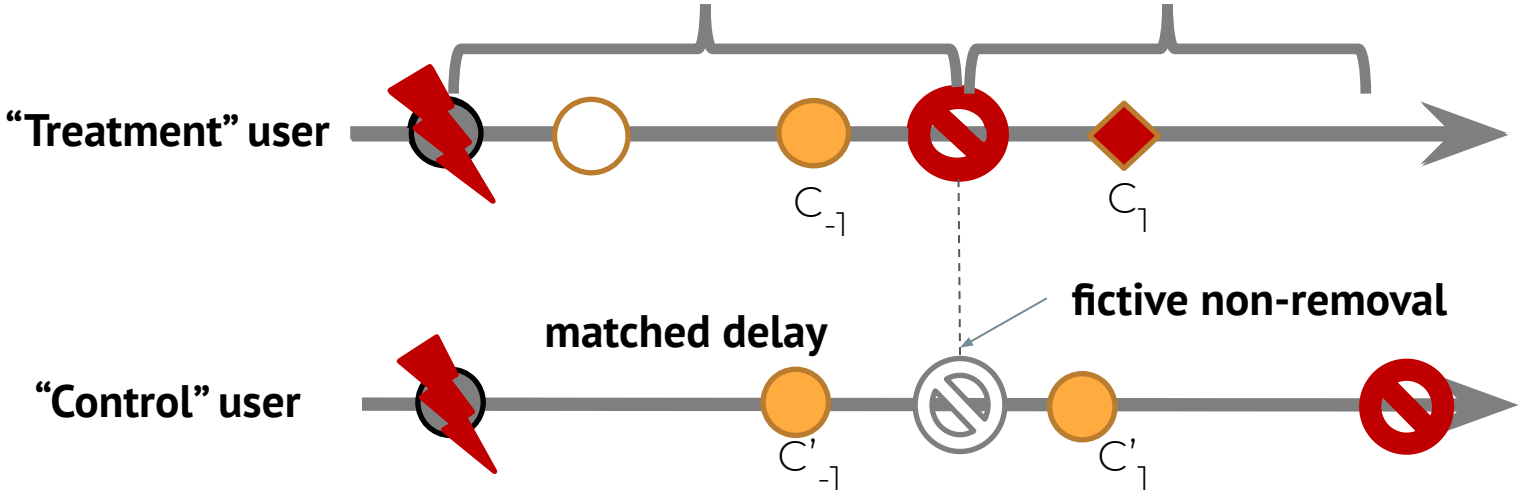
p<0.001

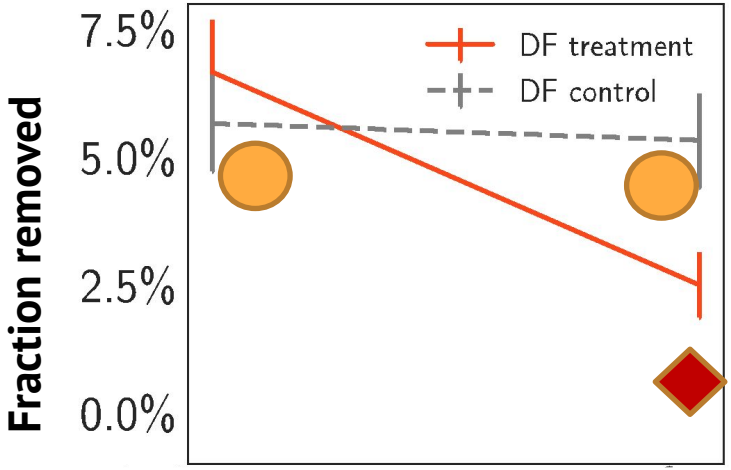

"Treatment" user
C-1
C1

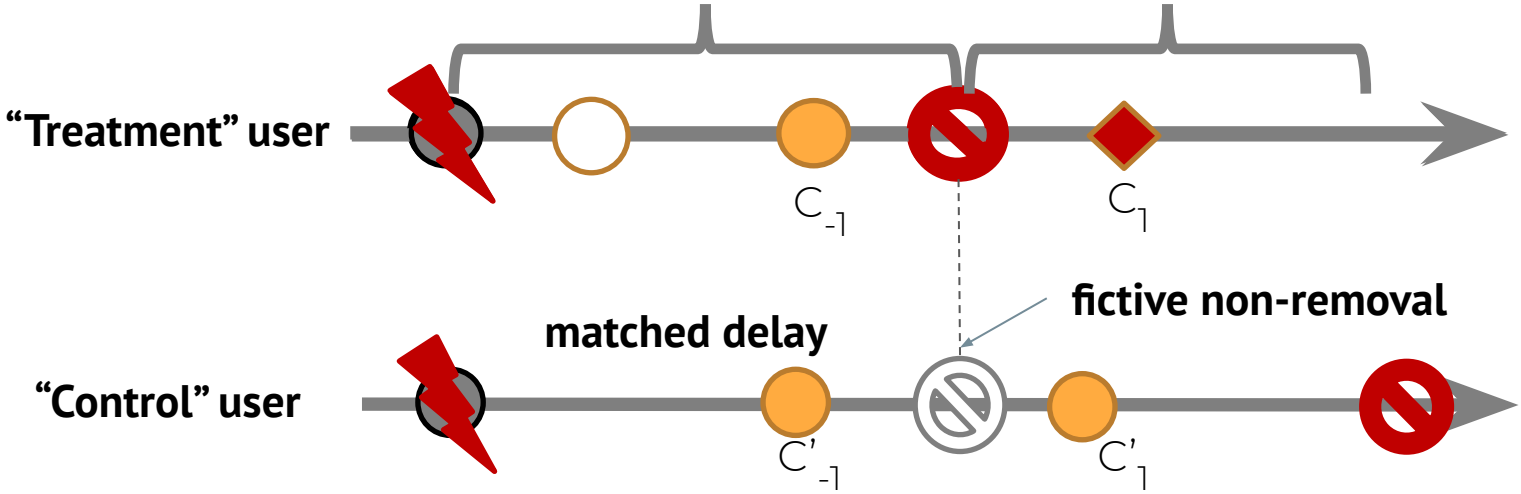

"Control" user

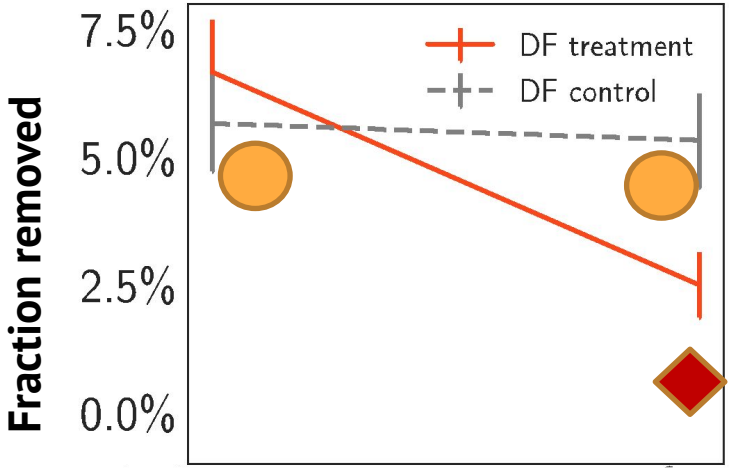

Fraction removed
7.5%
5.0%
2.5%
0.0%

p<0.001

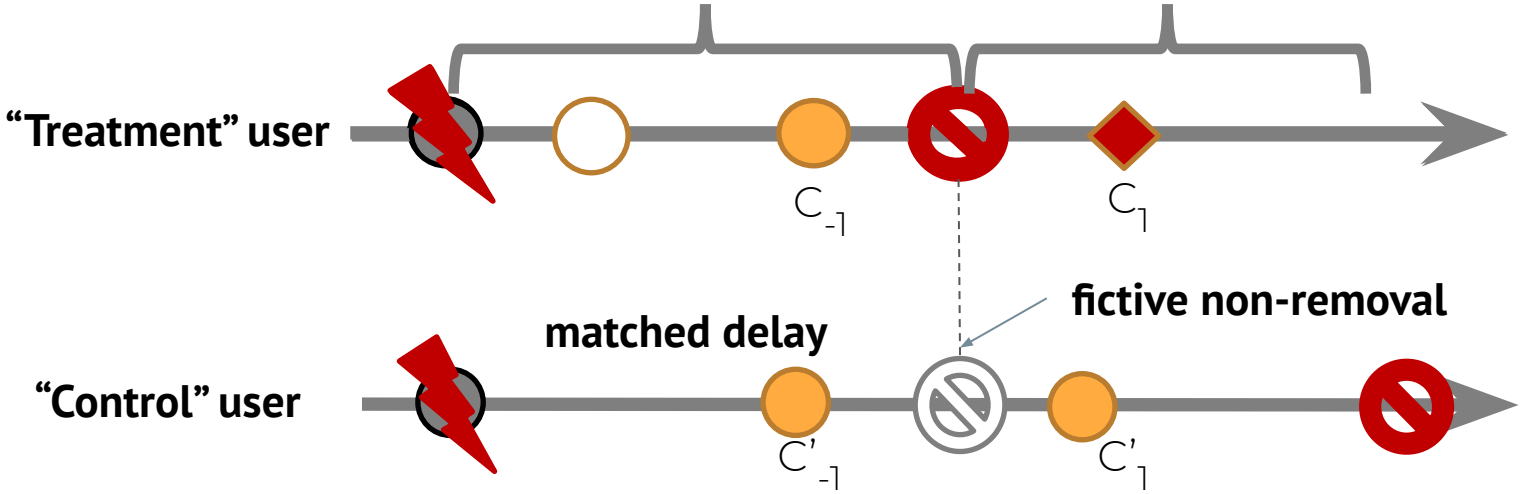

"Treatment" user
C -1
C 1
fictive non-removal
matched delay
"Control" user
C' -1
C' 1


Fraction removed
7.5%
5.0%
2.5%
0.0%
DF treatment
DF control


"Treatment" user
C -1
C 1
fictive non-removal
matched delay
"Control" user
C' -1
C' 1



No temporal difference without removal



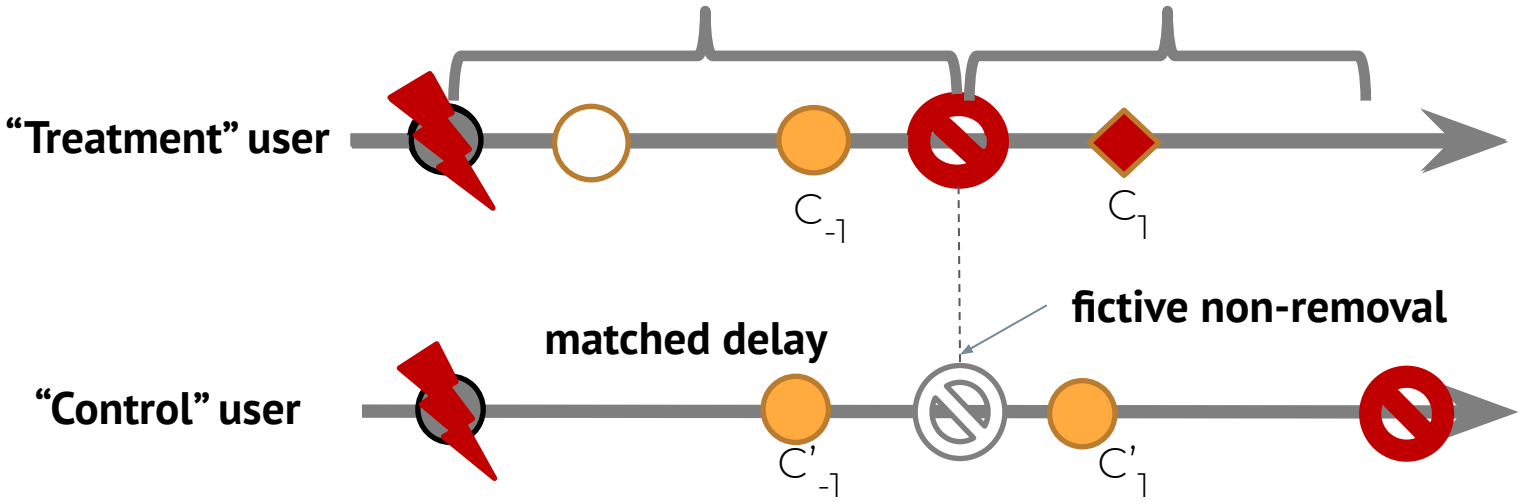
“Delayed feedback” paradigm

Provides a way to find the missing counterfactual in observational data

Limitation: we can only claim causality about active users,
who comment twice between posting and removal
(other caveats in the paper)

Good enough if we only care about active users

Again, we narrowed the scope of our claim to render it tractable



Takeaways

Making causal claims about conversations is hard, but don't give up

Formally describing why that is hard pays off

Wishful thinking as a conceptual mechanism

- Imagine ideal setting needed for establishing "truly causal" relations
  - e.g., random assignment, access to counterfactual situations
- Formally describe what is missing in our non-ideal observational setting
- Find an alternative that is just as good (or good enough)
  - Potentially by narrowing the scope of the claim to make it tractable

Thank you!