Constraints on Counterfactuals
Ruth M.J. Byrne
Trinity College Dublin
University of Dublin, Ireland

Counterfactuals
“What if” and “If only” thoughts about imagined alternatives to reality
How a decision in the past could have been different, or how one in the future could be different

Many uses of counterfactuals in AI
- Sub-goals; planning failures; fault diagnosis
  Ginsberg, 1986; Hftpym & Pearl, 2005
- Counterfactual risk/regret minimization
  Swenerton & Zanow, 2011; Moravčík et al., 2017
- Generative adversarial networks (GANs)
  Swaminathan & Joachims, 2015; Moravčík et al., 2017

eXplainable AI (XAI)
- Decisions of complex AI systems
  e.g., criminal sentencing, creditworthiness, automated vehicles
- Artificial neural networks (ANNs)
  trained on vast amounts of data; can produce unintelligible decisions
- AI systems to explain decisions to humans to increase trust by users, accuracy in training by developers

Counterfactuals in XAI
- Contrastive explanations: why one decision was made instead of another
  Hoffman et al., 2018; McGloth et al., 2018; Miller, 2018; Wecht & et al., 2018

Examples
- Categorisation
  What if feature A had had a higher weighting, would the selected category have been different?
- Creditworthiness
  If the applicant’s income were $50,000 higher, would their loan application be approved?
- Autonomous vehicles
  If only the car had braked instead of swerving, the crash would have been avoided.
Which counterfactuals?

Number is potentially limitless
What to change?
Not all are equally helpful for humans
"Minimal" change is slippery concept
Evidence from human reasoners
Which counterfactuals do people create, how do people reason about counterfactuals?

1. Human counterfactual thought
2. Mental models of counterfactuals
3. How people create counterfactuals

Human counterfactual thought

Two enduring questions in cognitive science

Are people rational?

Does the mind have the competence to make inferences, constrained by performance?

How Are People Creative?
How does the human mind make discoveries, create new ideas, engage in literature and art?

Counterfactuals bridge reasoning and creativity

People can create rich alternatives to reality

We often imagine how things could have turned out differently ‘if only…’ usually after bad outcomes

Kahneman & Tversky 1982

Byrne 2016 Annual Review of Psychology
They help us to explain the past
We can use them to work out causes
McCrae, 2009; Spellman & Mandel, 1999;
Dixon & Byrne, 2011 Memory & Cognition

They help us to prepare for the future
We can learn from mistakes, form intentions, make decisions
Roese &Epstude, 2017; Tenente et al., 2013;
McElroy & Byrne, 2006; Walsh & Byrne, 2007 Thinking & Reasoning

Emotions
Regret, guilt, relief, hope
Roese & Epstude, 2017; Swenn & John, 2012;
Byrne & McElroy, 2000 JEP: Learning, Memory & Cognition

Moral judgments
Blame, fault, responsibility
Malle et al., 2014; Alicke et al., 2008;
Parkinson & Byrne, 2017 Judgment & Decision Making
Byrne & Timmons, 2018 Cognition

Counterfactual thoughts emerge early and develop through middle childhood
Harris, 2000; Bell et al., 2006;
Raspa, Quehhas & Byrne, 2015 Cognitive Development

Loss of imagination
Injuries to the prefrontal cortex;
Parkinson’s, Schizophrenia, Autism
Gomez-Beldarrain et al., 2005; McMamara et al., 2003
Raspa, Quehhas & Byrne, 2017 Journal of Autism and Developmental Disorders
From Van Hoeck, Watson, & Berney, 2015; see also De Brigard, Addis, Ford, Schacter, & Giovanello, 2013

**Experimental evidence**

**Methods**
- Remember episodes from own life
- Read stories about hypothetical events with fictional protagonists about accidents, illnesses, goal failures
- Engage in a task, e.g., solving a puzzle

**Measures**
- Types of counterfactuals people create
- Inferences from counterfactuals
- How quickly people read counterfactuals
- What is looked at for a counterfactual
- Brain imaging of activated areas

**Artificial Imaginative Intelligence (AII)**

Evidence from human reasoners can maximize effectiveness of XAI

AI systems that construct same counterfactuals humans naturally create

**Illustration: Counterfactuals and Causes**

Two sides of one coin?

A caused B

If A hadn’t happened, B wouldn’t have happened


**Example**

Autonomous vehicle

Decision to swerve to avoid hitting cyclist

Outcome: hit wall, passenger minor injuries

**Counterfactuals amplify causal judgments**

**Counterfactual**
- different action - different outcome

If the car had braked instead of swerving, the passenger wouldn’t have been injured

Amplifies judgments of causal relation

**Semi-factual**
- different action - same outcome

Even if the car had braked instead of swerving, the passenger would still have been injured

Reduces judgments of causal relation

McCloy & Byrne, 2002
(1) But they focus on different sorts of causes

A drunk driver swerved into a protagonist driving home on an unusual route

What was the cause of the accident? The drunk driver

How could things have turned out differently? If the protagonist had driven home by his usual route

Mandel & Lehman, 1996

(1) But they focus on different sorts of causes

A drunk driver swerved into a protagonist driving home on an unusual route

What was the cause of the accident? The drunk driver

How could things have turned out differently? If the protagonist had driven home by his usual route

Strong causes that convey with outcome; necessary and sufficient

Background enabling conditions that could prevent outcome; necessary and not sufficient

(2) People spontaneously create twice as many causal thoughts as counterfactual thoughts

Example

Spontaneous undirected thoughts

Causal explanations about the facts as they happened

Counterfactual thoughts about imagined alternatives

I didn’t make friends because I didn’t go to the party

If I had gone to the party I would have made friends

McEleney & Byrne, 2006 Thinking & Reasoning

(3) Cause-effect sequences

CAUSE Effect

Bad weather conditions caused the car accident

If only the weather hadn’t been bad...

Focus on cause

Clare Walsh
Plymouth University

Wells et al., 1989

CAUSE Effect

Bad weather conditions caused the car accident

If only the weather hadn’t been bad...

Focus on cause

Reason ACTION

Wanting to get milk was the reason the person drove to the shop

If only he hadn’t driven to the shop...

Focus on action

Clare Walsh
Plymouth University
If AI agent provides counterfactual, human will readily infer a causal relation. But given the differences between causal and counterfactual thoughts, explicit causal explanations as well as counterfactual alternatives may be needed.

Implications for XAI

1. Human counterfactual thought
2. Mental models of counterfactuals
3. How people create counterfactuals

Counterfactuals in Cognitive Science

Truth functional semantics
Unable to account for meaning of counterfactuals
Adams, 1975

Possible worlds semantics
Truth in the closest possible world
Lewis, 1973; Stalnaker, 1968

Closest possible world is ill-defined concept
Kratzer, 2012; Williamson, 2007

Counterfactuals and Mental Models

People construct a mental model
Represent aspects of a situation
Create counterfactual
Select aspect represented in model
Modify to create an alternative model
Byrne, 2002 Trends in Cognitive Sciences

People construct mental models to simulate possibilities

If the car turned, it hit the cyclist

<table>
<thead>
<tr>
<th>turned</th>
<th>not-turned</th>
<th>hit</th>
<th>not-hit</th>
</tr>
</thead>
</table>
| not-turned | hit | Possibilities

Phil Johnson-Laird, 1983, 2006;
Johnson-Laird & Byrne, 2009 Trends in Cognitive Sciences
If the car turned, it hit the cyclist

- turned: hit
- not-turned: not-hit
- not-turned: hit

Principle of truth

If the car turned, it hit the cyclist

- turned: hit
- not-turned: ...

Principle of parsimony

If the car turned, it hit the cyclist

- turned: hit
- not-turned: not-hit
- not-turned: hit

Working memory constraints

If the car had turned, it would have hit the cyclist

- turned: hit
- not-turned: not-hit
- not-turned: hit

Counterfactual

Factual

1. What do counterfactuals imply?

If the car had turned, it would have hit the cyclist

What does someone uttering the assertion mean to imply?
- The car did not turn.
- It did not hit the cyclist.

Dual possibilities for counterfactuals

<table>
<thead>
<tr>
<th>Counterfactual</th>
<th>Factual</th>
</tr>
</thead>
<tbody>
<tr>
<td>if the car had turned, it would have hit the cyclist</td>
<td>if the car turned, it hit the cyclist</td>
</tr>
<tr>
<td>turned: hit</td>
<td>turned: hit</td>
</tr>
<tr>
<td>not-turned: not-hit</td>
<td>C’Fact:</td>
</tr>
<tr>
<td></td>
<td>Fact:</td>
</tr>
</tbody>
</table>

Valerie Thompson
University of Saskatchewan

Thompson & Byrne, 2002
Journal of Experimental Psychology: Cognition, Memory & Emotion
(2) Comprehension

Michael goes to a fruit market with his sister who tells him:
If there had been oranges there would have been pears.
The seller tells them:
There are no oranges and there are no pears.

Santamaria, Espino & Byrne, 2005
Journal of Experimental Psychology: Learning Memory & Cognition

Orlando Espino
La Laguna, Tenerife

(2) Comprehension

Michael goes to a fruit market with his sister who tells him:
If there are oranges there are pears.
The seller tells them:
There are no oranges and there are no pears.

Santamaria, Espino & Byrne, 2005
Journal of Experimental Psychology: Learning Memory & Cognition

Orlando Espino
La Laguna, Tenerife

People read both possibilities rapidly from counterfactuals

Adapted from Santamaria et al., 2005, Exp 1

(3) Eye-tracking

Adults listened to stories on headphones; looked at displays on screen; cameras recorded their eye movements
Where people look indicates what they are thinking about

Isabel Orenes
UNED Madrid
If there are oranges, there are pears

If there had been oranges, there would have been pears

Individual differences

Dual possibilities

People understand counterfactual conditionals by simulating two possibilities
If the car turned, it hit the cyclist

The car turned. Did it hit the cyclist?

100% modus ponens

If the car turned, it hit the cyclist

The car did not hit the cyclist. Did it turn?

50-60% modus tollens

If the car turned, it hit the cyclist

The car did not hit the cyclist. Did it turn?

40-50% nothing follows

If the car turned, it hit the cyclist

The car did not hit the cyclist. Did it turn?

Limitations in working memory

Johnson-Lehr & Byrne, 2002 Psychological Review
Espino & Byrne, 2012 Cognitive Psychology
If the car had turned, it would have hit the cyclist.

The car did not hit the cyclist.

Did it turn?

80% modus tollens

If the car had turned, it would have hit the cyclist.

The car turned.

Did it hit the cyclist?

90% modus ponens

(4) People make more inferences, e.g., modus tollens, from counterfactuals compared to factual conditionals

Byrne & Tasso, 1999; Thompson & Byrne, 2002; Frosch & Byrne, 2012; Egan & Byrne, 2012; Quelhas & Byrne, 2003

Causal Bayes nets

Counterfactuals overcome working memory limitations

Byrne, 2017 Current Directions in Psychological Science

People make more inferences from counterfactuals

Byrne & Tasso, 1999; Thompson & Byrne, 2002; Frisch & Byrne, 2012; Egan & Byrne, 2012; Quelhas & Byrne, 2003

Counterfactual Inference Effect

Byrne, 2017 Current Directions in Psychological Science
Confirms potential usefulness of counterfactuals in XAI
People construct models constrained by working memory
Counterfactuals enable representation of dual possibilities
More inferences from counterfactuals

Individual differences
AI agent - probe human user has mentally simulated dual possibilities

1. Human counterfactual thought
2. Mental models of counterfactuals
3. How people create counterfactuals

How people create counterfactuals
People simulate events
They are influenced by the availability of alternatives
The availability of alternatives depends on norms
Kahneman & Tversky, 1982; Kahneman & Miller, 1986

People zoom in on 'fault-lines' junctures or joints in the representation of reality

“Fault-lines”

Exceptions
If only he had driven home by his usual route...

Actions
If only the car hadn't swerved...

Recent events
If only the cyclist had stayed in the cycle lane...

Controllable
If only he hadn’t stopped at a bar for a beer...

Kahneman & Tversky, 1982; Dixon & Byrne, 2011; Byrne & McEvoy, 2000; Walsh & Byrne, 2004; De Brigard et al., 2013

What people don’t change
NOT miracle worlds
NOT - if only the passenger had been made of steel...

NOT Probability
NOT - if only the car and bike hadn’t been in the exact same place at the exact same time...

NOT continuous variables
NOT - if only the car had been going at 25 km instead of 30 km...

Kahneman & Tversky, 1982

A descriptive counterfactual is not necessarily an explanatory one
Controllable events

People change events within their own control
Protagonist delayed by traffic jam, weather conditions, ... and intentional decision to visit a bar for a beer
If only he hadn’t called into the bar for a beer
Girotto, Legrenzi & Rizzo, 2001

Readers

Anna can win a prize if she multiplies a sum in 30 secs. She has to select envelope A or B, one with an easy sum, one a hard sum. She chose A, it contains the hard sum, 68 x 87, and she fails to solve it in 30 seconds.
How could things have been better for Anna?
If only she had chosen the other envelope
Girotto, Ferrante, Pighin, & Gonzalez, 2007

Observers

How could things have been better for the player?
If only he’d had more time, a pen-and-paper...
Pighin, Byrne, Ferrante, Gonzalez, & Girotto, 2011 Thinking & Reasoning

Players

YOU can win a prize if you multiply a sum in 30 secs. You have to select envelope A or B, one with an easy sum, one a hard sum. You chose A, it contains the hard sum, 68 x 87, and you fail to solve it in 30 seconds.
How could things have been better for you?
If only I’d had more time, a pen-and-paper...
Pighin, Ferrante, Pighin, & Gonzalez, 2007

(3) Cooperative dilemmas

Reader – actor effect in cooperative dilemmas
Prisoner’s dilemma
Trust dilemma
Pighin, Byrne, & Tentori, 2010 submitted

(1) Constrained by moral norms
Protagonist delayed by traffic jam, weather conditions, ... and intentional decision to visit elderly parents
If only there hadn’t been a traffic jam, bad weather
McClay & Byrne, 2000 Memory & Cognition

(1) Constrained by moral norms
Protagonist delayed by traffic jam, weather conditions, ... and intentional decision to visit elderly parents
If only there hadn’t been a traffic jam, bad weather
Girotto, Legrenzi & Rizzo, 2001

(2) Constrained by moral norms
Protagonist delayed by traffic jam, weather conditions, ... and intentional decision to visit elderly parents
If only there hadn’t been a traffic jam, bad weather
McCloy & Byrne, 2000 Memory & Cognition

(3) Cooperative dilemmas
Reader – actor effect in cooperative dilemmas
Prisoner’s dilemma
Trust dilemma
Pighin, Byrne, & Tentori, 2010 submitted

(3) Cooperative dilemmas
Reader – actor effect in cooperative dilemmas
Prisoner’s dilemma
Trust dilemma
Pighin, Byrne, & Tentori, 2010 submitted
Readers change the protagonist’s own decision to cooperate or not - controllable

Actors change the other player’s decision – uncontrollable

The probabilities remain the same in all versions

Pighin, Byrne, & Tentori, 2019, submitted

**Implications for XAI**

People create counterfactuals by constructing models, modulated by knowledge
They zoom in on “fault-lines”, whether minimal or not Descriptive counterfactuals are not always explanatory
It matters whether people read about an AI system and its decisions, or observe it in action Actors and observers simulate events differently from readers

**Conclusions**

How are people creative?

People create counterfactual alternatives guided by ‘fault-lines’ in their mental simulation of reality

Are people rational?

People consider a conclusion to be valid if they cannot imagine a counterexample - rational in principle even if err in practice

Imagined alternatives to reality help people reason
Conclusions for XAI

- Counterfactuals aid explanation
  - but focus on different causes from causal thoughts

- Counterfactuals ensure people simulate dual possibilities
  - enable otherwise difficult inferences

- Explanatory counterfactuals zoom in on "fault-lines" in mental simulations
  - actors and observers simulate events differently from readers

---

Evidence from human reasoners can maximize effectiveness of XAI

- AI agent that constructs same counterfactuals human naturally creates

http://reasoningandimagination.com/ @ruthmjbyrne