Mental models and counterfactual thoughts about what might have been

Ruth M.J. Byrne

Counterfactual thoughts about what might have been ('if only…') are pervasive in everyday life. They are related to causal thoughts, they help people learn from experience and they influence diverse cognitive activities, from creativity to probability judgements. They give rise to emotions and social ascriptions such as guilt, regret and blame. People show remarkable regularities in the aspects of the past they mentally 'undo' in their counterfactual thoughts. These regularities provide clues about their mental representations and cognitive processes, such as keeping in mind true possibilities, and situations that are false but temporarily supposed to be true.

In a speech in 1968, Martin Luther King described how near to death he had come when he was stabbed, and he considered how things might have turned out differently:

> 'The tip of the blade was on the edge of my aorta…it came out in the *New York Times* the next morning that if I had sneezed I would have died…and I want to say tonight, I want to say that I'm happy I didn't sneeze…Because if I had sneezed, I wouldn't have been around here in 1960 when students from all over the South started sitting-in at lunch counters…If I had sneezed, I wouldn't have been here in 1963 when the black people of Birmingham, Alabama aroused the conscience of this nation and brought into being the civil rights bill…If I had sneezed I wouldn't have had the chance later that year in August to try to tell America about a dream that I had had… I'm so happy that I didn't sneeze.'

Counterfactual thoughts about what might have been are irresistible, especially after something bad happens [1–3], and they have attracted attention in each of the disciplines contributing to cognitive science.

Philosophical analyses of counterfactuals led to one of the most important advances in logic – 'possible world' semantics [4]. All counterfactuals have false antecedents (MLK didn't sneeze) and false consequents (he didn't die at that time) (see also Box 1), and distinguishing true from false ones requires an examination of properties that the world might have had or even 'parallel universes' [4]. Linguists have identified that the subjunctive mood is not necessary to convey counterfactuality [5], which can be expressed even in languages that do not contain specific linguistic markers, such as Chinese [6]. The ability to generate counterfactual thoughts can be lost however, following impairments to the frontal cortex [7]. Artificial intelligence systems have shown the usefulness of counterfactuals in planning sub-goals and robotic learning from experiences [8]. In addition, psychologists have performed hundreds and hundreds of experiments to determine the cognitive, social, motivational and emotional underpinnings of counterfactual thoughts in the 20 years since Daniel Kahneman and Amos Tversky's pioneering ideas [1]*.

A key question of interest to cognitive scientists is, how do people mentally 'undo' reality? People show remarkable regularities in the counterfactuals they generate, and a selective snapshot of even a few of the clues garnered in studies in just the past two or three years in this fast-moving field is illuminating.

Counterfactual thinking is pervasive

The day-to-day exercise of imagination skills starts early: children as young as two years of age can understand what 'nearly' or 'almost' happened and the development of counterfactual thought underlies their capacity for pretend play and to ascribe false beliefs to others [9–10]. The significance of great drama or moments of epiphany in literature often rests on an appreciation of a counterfactual alternative. For example, the characters in Samuel Beckett's *Waiting for Godot* instantiate a view of the human condition through their anticipation of something that never happens, and popular action films often exploit the suspense engendered by 'near misses'. Thoughts of how events might have turned out differently play an important, if controversial, role in historical analysis, say, in working out how the rise of the West occurred [11–13], and people rely on counterfactual alternatives to grasp the significance of current politics, for example, '*if LePen had won the French election*…' [11]. Most of all, counterfactuals figure prominently in daily musings, whether after dramatic life events such as accidents, or autobiographical reminiscence of lucky chances or opportunities lost [14].

Counterfactual thoughts are implicated in diverse cognitive activities, from daydreams and fantasy to deduction and probability calculation. They provide the building blocks for generating imaginary possibilities in creative and insightful cognition [15–17], such as in imagining novel categories and combinations [18–19]. Generating counterfactuals helps people to search for counterexamples to their deductions [20–21]. They may also help people avoid the 'certainty of hindsight' bias (i.e. the judgement, after an outcome is known,

*Neal Roose maintains an excellent counterfactual news and bibliography site at: http://www.sfu.ca/counterfactual/

Ruth M.J. Byrne
Psychology Dept,
University of Dublin,
Trinity College, Dublin 2,
Ireland.
e-mail: rmbyrne@tcd.ie

 1364-6613/02/$ – see front matter PII: S1364-6613(02)01974-5

Box 1. Counterfactual conditionals

People keep two possibilities in mind when they understand a counterfactual conditional (e.g. *'if the car had been out of petrol then it would have stalled'*) [a]. They believe that someone asserting the counterfactual means to imply that the facts are that the car was not out of petrol and it did not stall [b]. To verify the counterfactual they generate *'the car did not run out of petrol'* and *'it did not stall'* [a]. When they are given the counterfactual and then a surprise recognition test, they believe they were given *'the car did not run out of petrol'* and *'it did not stall'* [c]. To falsify the counterfactual they generate *'the car ran out of petrol'* and *'it stalled'* [a]. They judge that it is inconsistent with the situation in which the car actually was out of petrol and did stall [b].

People can readily make otherwise difficult inferences from a counterfactual conditional. When they are given the information *'the car did not stall'*, they readily infer, *'the car did not run out of petrol'* [a]. They do not make this inference readily given the regular conditional, *'if the car was out of petrol then it stalled'*. They make the easy inference from, *'the car ran out of petrol'* to *'it stalled'* just as readily from the counterfactual and the regular conditional [a]. However, they take longer to make inferences from counterfactuals [d].

Different counterfactuals focus people's attention on the facts or on the counterfactual possibility [e–f]. Priming people to think counterfactually (e.g. before Wason's selection task) affects the inferences they make [g]. People think counterfactually to revise their beliefs when inferences are contradicted [h–i], and they are less inclined to reject a belief expressed counterfactually [h]. Causal conditionals are interpreted as counterfactual more often than definitional ones, such as, *'if the animal had been a robin then it would have been a bird'* [b]. However, deontic conditionals about obligations, such as *'if the nurse had cleaned up the blood then she must have had to wear rubber gloves'*, are not interpreted as counterfactual [j].

References

a Byrne, R.M.J. and Tasso, A. (1999) Deductive reasoning with factual, possible, and counterfactual conditionals. *Mem. Cogn.* 27, 726–740

b Thompson, V.A. and Byrne, R.M.J. Reasoning counterfactually: making inferences about things that didn't happen. *J. Exp. Psychol. Learn. Mem. Cogn.* (in press)

c Fillenbaum, S. (1974) Information amplified: memory for counterfactual conditionals. *J. Exp. Psychol.* 102, 44–49

d Quelhas, C. and Byrne, R.M.J. (2000) Latencies to understand and reasoning from counterfactual conditionals. In *Mental Models and Reasoning* (Garcia-Madruga, J.A., ed.), pp. 315–326, UNED, Madrid, Spain

e McMullen, M.N. and Markman, K.D. (2000) Downward counterfactuals and motivation: the wake-up call and the pangloss effect. *Person. Soc. Psychol. Bull.* 26, 575–584

f Johnson-Laird, P.N. (1986) Conditionals and mental models. In *On Conditionals* (Traugott, E.C. *et al.*, eds), Cambridge University Press

g Galinsky, A.D. and Moskowitz, G.B. (2000) Counterfactuals as behavioural primes: priming the simulation of heuristics and consideration of alternatives. *J. Exp. Soc. Psychol.* 36, 384–409

h Byrne, R.M.J. and Walsh, C.R. Contradictions and counterfactuals: belief revision in conditional inference. In *Proc. 24th Ann. Conf. Cogn. Sci. Soc.*, Erlbaum (in press)

i Revlin, R. *et al.* Reasoning counterfactually: combining and rending. *Mem. Cogn.* (in press)

j Quelhas, A.C. and Byrne, R.M.J. Reasoning with deontic and counterfactual conditionals. *Think. Reason.* (in press)

that it was inevitable, such as the likelihood that a side would have won a war). Such biases can compromise the objectivity of historical analyses, but thinking of counterfactual alternatives, such as ways the side could have lost the war, can 'debias' thinking [11]. However, debiasing can backfire if people are asked to think of too many counterfactual alternatives: the difficulty of accessing more and more counterfactuals leads to the judgement that the outcome was inevitable after all [22].

What's it good for?

Many studies of the antecedents of counterfactual thoughts and their consequences, including studies of the relation between counterfactual and causal thought (see Box 2), show that they are generated most often after bad outcomes, particularly goal failures [23]. 'Upward' counterfactuals about how a situation could have turned out better, for example, '*if I'd studied I would have got an A*', serve a preparatory function, helping people to learn from mistakes [24]. Aviation pilots in near accidents who generate upward counterfactuals that focus on themselves, formulate effective plans to prevent a recurrence [25]. 'Downward' counterfactuals about how a situation could have turned out worse, for example, '*if I hadn't crammed last night I would have got only a C*', serve an affective function, helping people to feel better. People generate more upward counterfactuals [23], although they generate downward ones when they are in a good mood [26–27]. Their counterfactuals may be filtered by their motivations, for example, to console a victim, or to assign blame [23,28].

Considerable attention has been paid to one of the major consequences of counterfactual thinking: it can amplify emotions, such as guilt, shame and regret, or feelings of relief, satisfaction and luck, experienced through the comparison of an outcome with how it might have turned out [1–2,29]. These emotions can influence decision-making and choices about risk [30–32]. A great many social ascriptions are also consequences of counterfactual thoughts, such as judgements of accountability, fault, responsibility and blame; for example, counterfactual 'excuses' are most common when one must account publicly for unforeseeable outcomes [33–34]. These social, emotional and motivational antecedents and consequences sandwich the main cognitive meat of counterfactual thoughts: its content.

What is mutable?

People show extraordinary regularities in zooming in on the same things from the infinite set of possibilities [2]. In 'subjunctive instant replays' [16], what people focus on gives a clue about the 'fault lines' of the imagination [1–2]. Various constraints limit their counterfactuals: they tend not to alter natural laws, for example, '*if only the plane had fallen up…*' [28]; their counterfactuals are goal-driven [23,28,35]; and they are hugely influenced by the availability of alternatives [1]. They make minimal mutations, perhaps corresponding to core categories of thought: space, time, cause and intentionality [35].

'Close' counterfactuals, such as just missing an aeroplane by minutes [36], can distort our emotions: an objectively better outcome, such as coming second

Box 2. Causal and counterfactual thinking

The relation of counterfactual and causal thoughts has been hotly debated since philosophers first considered that to identify whether A caused B it is helpful to consider whether B would have happened if A hadn't [a]. Children by the age of four can refer to counterfactual alternatives when considering how an outcome could be prevented [b], and causal and counterfactual inferences are clearly related [c–d]. People judge an event to be more causally related to an outcome when an alternative is available. For example, a taxi driver's refusal to give a couple a lift is judged to be a cause of their deaths when their car was in an accident, but not when the taxi driver's car was also in the accident [e]. When people think that an outcome would still have happened *'even if'* a candidate cause had not occurred, they judge the candidate to be less causal [f].

People focus on first causes more than subsequent ones [g]: when a person gets to a store too late for a sale, they focus on the first in a series of delays, such as being stopped in a traffic jam, rather than subsequent causes, such as being stopped by a policeman, and they distinguish between necessary and sufficient causes [h]. Counterfactual and causal thoughts are activated by different triggers, negative outcomes and unexpected outcomes, and they have different consequences for emotions and social ascriptions (i–k).

The debate has moved beyond an initial 'which-comes-first' question (i.e. whether counterfactuals help people work out causes, or whether counterfactuals are generated from prior causal knowledge) [a]. Instead, counterfactual thoughts seem sensitive to particular sorts of cause, prevention rather than co-variation. For example, when a woman is injured in a car crash on a route she rarely takes, by a car driven by a drunk man swerving into her lane, people identify the drunk man as the cause of the accident, but they still generate counterfactuals about the woman's controllable decisions, such as *'if only I'd driven home by my usual route'* [l]. The distress of accident victims and bereaved individuals may be increased by counterfactual thoughts that focus on their own controllable behavior, despite awareness that it did not cause the bad outcome [m].

References

a Spellman, B.A. and Mandel, D.R. (1999) When possibility informs reality: counterfactual thinking as a cue to causality. *Curr. Dir. Psychol. Sci.* 8, 120–123

b Harris, P.L. *et al.* (1996) Children's use of counterfactual thinking in causal reasoning. *Cognition* 61, 233–259

c Goldvarg, Y. and Johnson-Laird, P.N. (2001) Naive causality: a mental model theory of causal meaning and reasoning. *Cogn. Sci.* 25, 565–610

d Yarlett, D. and Ramscar, M. Uncertainty in causal and counterfactual inference. *Proc. 24th Ann. Conf. Cogn. Sci. Soc.* (in press)

e Wells, G.L. *et al.* (1987) The undoing of scenarios. *J. Pers. Soc. Psychol.* 53, 421–430

f McCloy, R. and Byrne, R.M.J. (2002) Semifactual 'even if' thinking. *Think. Reason.* 8, 41–67

g Wells, G.L. and Gavanski, I. (1989) Mental simulation of causality. *J. Pers. Soc. Psychol.* 56, 161–169

h Ngbala, A. and Branscombe, N.R. (1995) Mental simulation and causal attribution: when simulating an event does not affect fault assignment. *J. Exp. Soc. Psychol.* 31, 139–162

i Sanna, L.J. and Turley, K.J. (1996) Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence. *Pers. Soc. Psychol. Bull.* 22, 909–919

j McEleney, A. and Byrne, R.M.J. (1999) Consequences of counterfactual reasoning and causal reasoning. In *Eur. Conf. Cogn. Sci.* '99 (Bagnara, S., ed.), pp. 199–205, University of Siena

k McEleney, A. and Byrne, R.M.J. (2000) Counterfactual thinking and causal explanation. In *Mental Models in Reasoning* (García-Madruga, J.A. *et al.*, eds), pp. 301–314, UNED, Madrid, Spain

l Mandel, D.R. and Lehman, D.R. (1996) Counterfactual thinking and ascriptions of cause and preventability. *J. Pers. Soc. Psychol.* 70, 450–463

m Davis, C.J. *et al.* (1995) The undoing of traumatic life events. *Pers. Soc. Psychol. Bull.* 21, 109–124

in a race, makes people feel worse than a poorer outcome, such as coming third [37]. 'Closeness' is related to plausibility [11,37]. People judge whether a counterfactual is plausible (e.g. '*if Kennedy had listened to his Hawk advisers he would have engaged in a nuclear strike during the Cuban missile crisis*') by how well it fits with their causal generalizations, for example, beliefs that nuclear deterence works effectively to prevent either side striking [13].

There is a huge array of factors that people focus on in their counterfactual thoughts. They focus on exceptional events more than routine ones [1]: when a person is killed in a car crash, the route they chose is focused on more when it is not their normal one. They focus on events within their control [38–39]: when a man arrives home too late to save his dying wife, people focus on his intentional decision (e.g. stopping at a bar for a beer), rather than events outside his control (a traffic jam). This 'controllability' effect is sensitive to social obligations: people focus on socially unacceptable events, such as stopping at a bar, more than obligations, such as stopping to see elderly parents [40]. People mentally undo more recent events in an independent sequence rather than earlier events (see Box 3), and they focus on their actions rather than inactions, at least in some circumstances (see Box 4). This set of factors is by no means exhaustive but it sheds some light on the key cognitive question concerning why some aspects of reality are more mutable than others.

How do people mentally undo reality?

One view of counterfactual generation is that people compute norms: a normal event evokes representations that resemble it, whereas an abnormal event has highly available alternatives [2]. The mutability or 'slippability' [16] of attributes is controlled by heuristics [2,41], and the availability of alternatives is distinct from expectations. For example, two travellers delayed by 30 minutes miss their respective flights. One flight departed on time, the other 25 minutes late (and so was missed by 5 minutes). People judge the individual who missed their flight by 5 minutes to feel worse; it is easier to recruit ways to make up the 5-minute difference and so there is an available alternative, even though both travellers expected to miss their flights. Another view is that norms account for the content of counterfactual generation, but its activation depends on goals [3,23]. Counterfactuals, which can be constructed either automatically or deliberately, are more common following goal failure regardless of norm violation, and they provide a roadmap for the future based on avoiding bad things in the past [23]. A third view is that counterfactual thinking depends on judgements of probability [42]. The baseline probability of an outcome (e.g. a woman being raped) before the antecedent occurred (the perpetrator gave her a lift home) may be judged to be low when people can think of many alternatives (she got the bus home, she got a lift from a friend),

Box 3. The temporal order effect

People generate counterfactual alternatives that undo the most recent event in a sequence of independent events [a]. Consider a game in which John and Michael each toss a coin. If the two coins come up the same (both heads or both tails), they each win $1000; if they do not come up the same, neither wins anything. John goes first and tosses a head; Michael goes next and tosses a tail, and so neither wins anything. Most people think they could have won if only Michael tossed a head (rather than if only John tossed a tail). They also judge that Michael will experience more guilt, and will be blamed more by John. The effect occurs for sequences of more than two events [b], and in everyday situations; for example, people place more weight on recent games in a basketball league [c], and they judge an individual to be lucky when a good outcome, such as a well-rated jump in a ski competition, is described after a bad outcome, such as a poorly-rated jump, rather than vice versa [d]. Counterfactuals that undo historical events, such as the rise of the West, often focus on the 'last chance' juncture when things could have taken a different turn [e].

People may calculate the probability of a good outcome after each player's contribution, which changes only after the second player's selection [f]. But the effect can be eliminated without changing the probabilities, by providing an explicit alternative to the first player's throw [g]. For example, John tosses tails, but there is a technical hitch and John must throw again and this time he tosses heads; Michael tosses tails. People undo John's throw as often as Michael's. A description that focuses attention on the first event leads people to undo the first event more than the second [h]. The results suggest that people keep in mind just some counterfactual possibilities, and several computational models simulate these effects [g–h].

References

a Miller, D.T. and Gunasegaram, S. (1990) Temporal order and the perceived mutability of events: implications for blame assignment. *J. Pers. Soc. Psychol.* 59, 1111–1118

b Segura, S. *et al.* Temporal and causal order effects in thinking about what might have been. *Q. J. Exp. Psychol.* (in press)

c Sherman, S.J. and McConnell, A.R. (1996) The role of counterfactual thinking in reasoning. *Appl. Cogn. Psychol.* 10, 113–124

d Tiegen, K.H. *et al.* (1999) Good luck and bad luck: how to tell the difference. *Eur. J. Soc. Psychol.* 29, 981–1010

e Tetlock, P.E. The logic and psycho-logic of counterfactual thought experiments in the rise of the West debate. In *Unmaking the West: Exploring Alternative Histories of Counterfactual Worlds* (Tetlock, P.E. *et al.*, eds), Cambridge University Press (in press)

f Spellman, B.A. (1997) Crediting causality. *J. Exp. Psychol. Gen.* 126, 323–348

g Byrne, R.M.J. *et al.* (2000) The temporality effect in counterfactual thinking about what might have been. *Mem. Cogn.* 28, 264–281

h Walsh, C. and Byrne, R.M.J. (2001) A computational model of the temporal order effect. In *Proc. 23rd Ann. Conf. Cogn. Sci. Soc.*, pp. 1078–1083, Erlbaum

Box 4. The agency effect

People are more concerned with harm caused by actions, such as death from a vaccine, than with harm caused by omissions, such as death from a disease that could have been vaccinated against, especially when the actions involve important values such as human life and civil rights [a–b]. They generate counterfactual alternatives that undo actions rather than inactions [c]. Consider Dave and Jim, both unhappy at the same university and both considering transferring to another college. Dave opts to stay where he is and Jim decides to transfer, but their decisions turn out badly. Most people judge that Jim, the one who acted, feels more regret [c], when a comparison between the actor and non-actor is required [d]. People prefer to do nothing, even when doing nothing itself leads to change [a–b], and the things they do not do far outnumber the things they do [e]. This 'omission bias' or 'agency effect' occurs in everyday life (e.g. people apologize on a national television apology show more often for things they did than for things they failed to do) [f], and people focus more on the actions of victims and perpetrators when they defend their roles [g].

However, the focus on actions seems to reverse from a long-term perspective. When people look back over their past lives, they regret failing to spend time with family and friends and failing to avail of educational opportunities [h]. They focus on inactions when the situation called for action, such as a soccer manager's inaction in fielding the same players despite a series of lost matches [i–j]. Although autobiographical regrets in retrospective recall are often inactions, the actions are judged to be more intense [k]. Regret for both actions and inactions is associated equally with hot emotions such as anger, but regret for inactions is associated more with wistful emotions such as nostalgia, and also with despair emotions, such as emptiness [l].

But the 'inaction effect' in the long-term may be limited to special sorts of inactions. People focus on actions even in the long-term for bad outcomes, when they have equal information about the actor and non-actor; for example, the actor lost $1000 because of his investment action and would have gained it if he had not acted, and equally, the non-actor lost $1000 because of his inaction and would have gained it if he had acted [m]. They focus on the actions of matched actors and non-actors even when the counterfactual possibilities are not explicity described [n]. The inactions people focus on might be those whose counterfactual outcomes are unknowable [m].

References

a Ritov, I. and Baron, J. (1990) Reluctance to vaccinate: omission bias and ambiguity. *J. Behav. Decision Making* 3, 263–277

b Ritov, I. and Baron, J. (1999) Protected values and omission bias. *Organ. Behav. Hum. Decision Process.* 79, 79–94

c Kahneman, D. and Tversky, A. (1982) The simulation heuristic. In *Judgment Under Uncertainty: Heuristics and Biases* (Kahneman, D., ed.), pp. 201–208, Cambridge University Press

d Ngbala, A. and Branscombe, N.R. (1997) When does action elicit more regret than inaction and is counterfactual mutation the mediator of this effect. *J. Exp. Soc. Psychol.* 33, 324–343

e Tykocinski, O.E. and Pittman, T.S. (1998) The consequences of doing nothing. Inaction inertia as avoidance of anticipated counterfactual regret. *J. Pers. Soc. Psychol.* 75, 607–616

f Zeelenberg, M. *et al.* (1998) Undoing regret on Dutch television. *Pers. Soc. Psychol. Bull.* 24, 1113–1119

g Catellani, P. and Milesi, P. (2001) Counterfactuals and roles: mock victims' and perpetrators' accounts of judicial cases. *Eur. J. Soc. Psychol.* 31, 247–264

h Gilovich, T. and Medvec, V.H. (1995) The experience of regret: what, when, and why. *Psychol. Rev.* 102, 379–395

i Zeelenberg, M. *et al.* The inaction effect in the psychology of regret. *J. Pers. Soc. Psychol.* (in press)

j Connolly, T. *et al.* (1997) Regret and responsibility in the evaluation of decision outcomes. *Organ. Behav. Hum. Decision Process.* 70, 73–85

k Feldman, J. *et al.* (1999) Are actions regretted more than inactions? *Organ. Behav. Hum. Decision Processes* 78, 232–255

l Gilovich, T. *et al.* (1998) Varieties of regret. *Psychol. Rev.* 102, 379–395

m Byrne, R.M.J. and McEleney, A. (2000) Counterfactual thinking about actions and failures to act. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1318–1331

n Babad, D.A. Mental undoings of actions and inactions. *Br. J. Psychol.* (in press)

increasing the judgement of the probability that the antecedent contributed causally to the outcome.

Our view is that some facts are more alterable because of the mental representations people construct of the facts and of the alternative possibilities [35]. Their mental models are governed by a small set of principles [20]: they keep alternative possibilities in mind; for example, to understand '*John tossed a head*

Table 1. The facts and three counterfactual alternative possibilities to the coin-toss game (see Box 3)

Possibilities	First player	Second player
Factual	John tossed heads	Michael tossed tails
Counterfactual	John tossed heads	Michael tossed heads
	John tossed tails	Michael tossed heads
	John tossed tails	Michael tossed tails

or he tossed a tail' they keep in mind two possibilities; in one John tosses heads, in the other he tosses tails. They represent as little as possible because of the limitations of working memory (43), and so they represent only what is true [20]. For example, they do not keep in mind the two possibilities in which John tossed neither a head nor a tail, or he tossed both a head and a tail. They represent some information explicitly and some implicitly [43]. In the model in which John tossed a head, for example, they do not represent explicitly that he did *not* toss a tail. Implicit information is not inaccessible, it may be 'fleshed out' to be explicit if need be, and it is akin to an unfinished thought, a means to construct a representation rather than a representation in itself. People can represent what is false, but temporarily supposed to be true [35]; for example, '*John would have tossed a head*' calls for two possibilities, one corresponding to the presupposed facts, John did not toss a head, and the other corresponding to what was once possible but is so no longer, John tossed a head.

Mental models have been corroborated in many domains of deductive inference [43–44]†, and this small set of principles can account for effects such as the 'temporal order effect' and the 'agency effect' (see also Boxes 3 and 4). Consider a game in which John and Michael each toss a coin: if the two coins come up the same (both heads or both tails), they each win $1000; if they do not come up the same, neither wins anything. John goes first and tosses a head; Michael goes next and tosses a tail, and so neither wins anything. Most people think they could have won if only Michael had tossed a head (rather than if only John tossed a tail). The temporal order effect arises because people keep in mind the facts (John tossed a head and Michael tossed a tail), and just one of the three situations in which the players could have won (see Table 1), in which John and Michael both tossed heads, because the first event, John tossed heads, is presupposed [35]. An alternative shakes loose this presupposition, as corroborated by the technical hitch scenario in which John tosses tails first and then the game is re-started.

†The mental models website is at: http://www.tcd.ie/Psychology/Ruth_Byrne/mental_models/

Questions for future research

- How are epistemic thoughts about what might have been and deontic thoughts about what should have been related?
- What factors influence judgements of the plausibility and closeness of counterfactuals?
- How do counterfactuals focus attention on the factual or counterfactual possibility?
- What commonalities exist between counterfactual thoughts and creative cognition?

Consider also the tendency to undo an action, such as Jim's decision to move to a new college, in contrast to an inaction, such as Dave's decision to stay at his original college. The agency effect arises because there are more possibilities to keep in mind when someone does something than when they do nothing: when people think about Dave, the non-actor, they think about his being in college A, but when they think about Jim, the actor, they think about his being not only in college B now, but also in college A at the outset. The post-action state can be replaced with the pre-action state, '*if only Jim had stayed in college A*….' People keep two possibilities in mind to understand a counterfactual conditional (see Box 1). Consider the counterfactual '*if MLK had sneezed he would have died*'. They keep in mind the facts (he did not sneeze and he did not die), as well as the suggested possibility (he sneezed and he died). For a regular conditional, '*if he sneezed he died*', they keep in mind a single possibility (he sneezed and he died). Because the counterfactual is represented more explicitly, people make more of the inferences that require access to the facts, although it takes them longer to do so. Our account implies that just as logical thought has turned out to be far more imaginative than previously supposed [43], so too imaginative thought might be far more logical than previously supposed [45].

Conclusions

People think about what might have been to try to prevent bad outcomes and to feel better. They mentally undo events within their control that are intentional, exceptional, recent, or a first cause, among other factors. These events may be most readily undone in imaginary simulations because of the sorts of representations they construct of reality. They represent what is true, including what is false but temporarily supposed to be true. Their mental representations make only some information explicit. The cognitive constraints on the sorts of representations people construct about the real world in turn limit their thoughts about imaginary worlds.

Acknowledgements
Thanks to Phil Johnson-Laird and Mark Keane for their comments on an earlier draft of this manuscript.

References

1 Kahneman, D. and Tversky, A. (1982) The simulation heuristic. In *Judgment Under Uncertainty: Heuristics and Biases* (Kahneman, D., ed.), pp. 201–208, Cambridge University Press
2 Kahneman, D. and Miller, D.T. (1986) Norm theory: comparing reality to its alternatives. *Psychol. Rev.* 93, 136–153
3 Roese, N.J. (1997) Counterfactual thinking. *Psychol. Bull.* 121, 133–148
4 Stalnaker, R.C. (1999) *Context and Content*, Oxford University Press
5 Dudman, V.H. (1988) Indicative and subjunctive. *Analysis* 48, 113–122
6 Au, T.K. (1983) Chinese and English

counterfactuals: the Sapir–Whorf hypothesis revisited. *Cognition* 15, 155–187
7 Knight, R.T. and Grabowecky, M. (1995) Escape from linear time: prefrontal cortext and conscious experience. In *The Cognitive Neurosciences* (Gazzaniga, M.S., ed), pp. 1357–1371, MIT Press
8 Costello, T. and McCarthy, J. (1999) Useful counterfactuals. *Linkoping Electronic Articles in Computer and Information Science* 3, 1–28
9 Riggs, K.J. and Peterson, D.M. (2000) Counterfactual thinking in preschool children: mental state and causal inferences. In *Children's Reasoning and the Mind* (Mitchell, P. and Riggs K.J., eds), Psychology Press
10 Riggs, K.J. *et al.* (1998) Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cogn. Dev.* 13, 73–90
11 Tetlock, P.E. The logic and psycho-logic of counterfactual thought experiments in the rise of the West debate. In *Unmaking the West: Exploring Alternative Histories of Counterfactual Worlds* (Tetlock, P.E. *et al.*, eds), Cambridge University Press (in press)
12 Lebow, R.N. (2000) What's so different about a counterfactual? *World Politics* 52, 550–585
13 Tetlock, P.E. and Lebow, R.N. (2001) Poking counterfactual holes in covering laws, cognitive styles and historical reasoning. *Am. Polit. Sci. Rev.* 95, 829–843
14 Gilovich, T. and Medvec, V.H. (1994) The temporal pattern to the experience of regret. *J. Pers. Soc. Psychol.* 67, 357–365
15 Sternberg, R.J. and Gastel, J. (1989) If dancers ate their shoes: inductive reasoning with factual and counterfactual premises. *Mem. Cogn.* 17, 1–10
16 Hofstadter, D.R. (1985) *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Penguin
17 Thomas, N.J.T. (1999) Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cogn. Sci.* 23, 207–245
18 Costello, F.J. and Keane, M.T. (2000) Efficient creativity: constraint guided conceptual combination. *Cogn. Sci.* 24, 299–349
19 Finke, R.A. *et al.* (1992) *Creative Cognition: Theory, Research and Applications*, MIT Press
20 Johnson-Laird, P.N. and Byrne, R.M.J. Conditionals: a theory of meaning, pragmatics, and inference. *Psychol. Rev.* (in press)
21 Byrne, R.M.J. *et al.* (1999) Counterexamples and the suppression of inferences. *J. Mem. Lang.* 40, 347–373
22 Sanna, L.J. *et al.* When debiasing backfires: accessible content and accessibility experiences in debiasing hindsight. *J. Exp. Psychol. Learn. Mem. Cogn.* (in press)
23 Roese, N.J. *et al.* The mechanics of imagination: automaticity and control in counterfactual thinking. In *The New Unconscious* (Bargh, J.A. *et al.*, eds), Cambridge University Press (in press)
24 Roese, N.J. (1994) The functional basis of counterfactual thinking. *J. Pers. Soc. Psychol.* 66, 805–818
25 Morris, M.N. and Moore, P.C. (2000) The lessons we (don't) learn: counterfactual thinking and organizational accountability after a close call. *Admin. Sci. Q.* 45, 737–765
26 Sanna, L.J. *et al.* (1999) Mood, self-esteem, and simulated alternatives: thought provoking affective influences on counterfactual direction. *J. Pers. Soc. Psychol.* 76, 543–558
27 Sanna, L.J. and Meier, S. (2000) Looking for clouds in a silver lining: self-esteem, mental simulations and temporal confidence changes. *J. Res. Pers.* 34, 236–251
28 Seeleu, E.P. *et al.* (1995) Counterfactual constraints. In *What Might Have Been: The Social Psychology of Counterfactual Thinking* (Roese, N.J. and Olson, J.M., eds), Erlbaum
29 Mandel, D.R. Counterfactuals, emotions, and context. *Cogn. Emot.* (in press)
30 Mellers, B.A. *et al.* (1999) Emotion based choice. *J. Exp. Psychol. Gen.* 128, 332–345
31 Slovic, P. *et al.* The affect heuristic. In *Heuristics of Intuitive Judgment: Extensions and Applications* (Gilovich, T. *et al.* eds), Cambridge University Press (in press)
32 Crawford, M.T. *et al.* (2002) Reactance, compliance, and anticipated regret. *J. Exp. Soc. Psychol.* 38, 56–63
33 Markman, K.D. and Tetlock, P.E. (2000) I couldn't have known: accountability, foreseeability, and counterfactual denials of responsibility. *Br. J. Soc. Psychol.* 39, 313–325
34 Hegarty, P. and Pratto (2000) The effects of social category norms and stereotypes on explanations for ntergroup differences. *J. Pers. Soc. Psychol.* 80, 723–735
35 Byrne, R.M.J. (1997) Cognitive processes in counterfactual thinking about what might have been. In *The Psychology of Learning and Motivation, Advances in Research and Theory*, Vol. 37 (Medin, D., ed.), pp. 105–154, Academic Press
36 Markman, K.D. and Tetlock, P.E. (2000) Accountability and close call counterfactuals: the loser who nearly won and the winner who nearly lost. *Pers. Soc. Psychol. Bull.* 26, 1213–1224
37 McMullen, M.N. and Markman, K.D. (2002) Affective impact of close counterfactuals: implications of possible futures for possible pasts. *J. Exp. Soc. Psychol.* 38, 64–70
38 Girotto, V. *et al.* (1991) Event controllability in counterfactual thinking. *Acta Psychol.* 78, 111–133
39 Klauer, K.C. *et al.* (1995) Counterfactual processing: test of an hierarchical correspondence model. *Eur. J. Soc. Psychol.* 25, 577–595
40 McCloy, R. and Byrne, R.M.J. (2000) Counterfactual thinking about controllable actions. *Mem. Cogn.* 28, 1071–1078
41 Kahneman, D. and Frederick, S. Representativeness revisited: attribute substitution in intuitive judgment. In *Heuristics of Intuitive Judgment: Extensions and Applications* (Gilovich, T. *et al.*, eds), Cambridge University Press (in press)
42 Spellman, B.A. (1997) Crediting causality. *J. Exp. Psychol. Gen.* 126, 323–348
43 Johnson-Laird, P.N. and Byrne, R.M.J. (1991) *Deduction*, Erlbaum
44 Legrenzi, P. *et al.* (1993) Focusing in reasoning and decision-making. *Cognition* 49, 37–66
45 Byrne, R.M.J. (1996) Towards a model theory of imaginary thinking. In *Mental Models in Cognitive Science: Essays in Honour of Phil Johnson-Laird* (Oakhill, J. and Garnham, A., eds), pp. 155–174, Erlbaum