NeuroImage 178 (2018) 332–345

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

What if? Neural activity underlying semantic and episodic counterfactual thinking

Natasha Parikh [a,b,*], Luka Ruzic [a,b], Gregory W. Stewart [a,b], R. Nathan Spreng [c,d], Felipe De Brigard [a,b,e,f]

[a] Department of Psychology and Neuroscience, Duke University, USA
[b] Center for Cognitive Neuroscience, Duke University, USA
[c] Laboratory of Brain and Cognition, Montreal Neurological Institute, Department of Neurology and Neurosurgery, McGill University, Canada
[d] Human Neuroscience Institute, Department of Human Development, Cornell University, USA
[e] Department of Philosophy, Duke University, USA
[f] Duke Institute for Brain Sciences, Duke University, USA

ARTICLE INFO

Keywords:
Counterfactual thinking
Episodic memory
Semantic memory
Plausibility
Partial least squares

ABSTRACT

Counterfactual thinking (CFT) is the process of mentally simulating alternative versions of known facts. In the past decade, cognitive neuroscientists have begun to uncover the neural underpinnings of CFT, particularly episodic CFT (eCFT), which activates regions in the default network (DN) also activated by episodic memory (eM) recall. However, the engagement of DN regions is different for distinct kinds of eCFT. More plausible counterfactuals and counterfactuals about oneself show stronger activity in DN regions compared to implausible and other- or object-focused counterfactuals. The current study sought to identify a source for this difference in DN activity. Specifically, self-focused counterfactuals may also be more plausible, suggesting that DN core regions are sensitive to the plausibility of a simulation. On the other hand, plausible and self-focused counterfactuals may involve more episodic information than implausible and other-focused counterfactuals, which would imply DN sensitivity to episodic information. In the current study, we compared episodic and semantic counterfactuals generated to be plausible or implausible against episodic and semantic memory reactivation using fMRI. Taking multivariate and univariate approaches, we found that the DN is engaged more during episodic simulations, including eM and all eCFT, than during semantic simulations. Semantic simulations engaged more inferior temporal and lateral occipital regions. The only region that showed strong plausibility effects was the hippocampus, which was significantly engaged for implausible CFT but not for plausible CFT, suggestive of binding more disparate information. Consequences of these findings for the cognitive neuroscience of mental simulation are discussed.

Introduction

The term *counterfactual*, coined by philosopher Nelson Goodman (1947), was introduced in reference to conditional statements whose antecedent is false, such as "If kangaroos had no tails, they would topple over" (Lewis, 1973, p. 1). Since, in fact, kangaroos do have tails, the antecedent of this statement asserts something that is contrary-to-fact or counterfactual. In the last four decades, a number of psychological studies have shed light on the cognitive processes that give rise to the thoughts expressed by counterfactual statements. However, until a few years ago, the scientific study of *counterfactual thinking* (CFT), as it is now known, was almost entirely confined to behavioral economics (Kahneman and Tversky, 1982), social psychology (Roese, 1997), and, more recently, cognitive science (Byrne, 2005).

Only in the past few years has cognitive neuroscience started to uncover the neural underpinnings of counterfactual thinking. Inspired by the striking commonalities found between *episodic memory* (eM) and episodic future thinking—i.e., thoughts about possible events that may occur to oneself in the future (Schacter et al., 2012; Spreng and Grady, 2010; Szpunar, 2010)—researchers began to compare these two kinds of mental simulations to *episodic counterfactual thinking* (eCFT): thoughts about alternative ways one's past personal events could have occurred but did not (De Brigard et al., 2016; De Brigard, Rodriguez and Montañés, 2017; De Brigard, Spreng, Mitchell and Schacter, 2015; De Brigard and

* Corresponding author. LSRC Box #90999, Duke University, Durham, NC, 27708, USA.
E-mail address: natasha.parikh@duke.edu (N. Parikh).

https://doi.org/10.1016/j.neuroimage.2018.05.053
Received 25 January 2018; Received in revised form 4 May 2018; Accepted 22 May 2018
Available online 25 May 2018
1053-8119/Published by Elsevier Inc.

Giovanello, 2012; Özbek et al., 2016, 2017; Stanley et al., 2017). In a pioneer study (Addis et al., 2009), participants were asked to remember *actual past* events, to mentally recombine elements from their memories into *possible future* events, or to mentally recombine them into *possible past* events, while undergoing fMRI. When compared to baseline, the results revealed substantial overlap across these three kinds of mental simulations in core regions of the brain's default network (DN), a set of functionally connected neural structures including ventral medial prefrontal cortex (vmPFC), posterior cingulate cortex (PCC), inferior parietal lobule (IPL), lateral temporal cortex (LTC), dorsal medial prefrontal cortex (dmPFC), and the hippocampal formation in the medial temporal lobe (MTL; Andrews-Hanna et al., 2014; Benoit and Schacter, 2015; Buckner et al., 2008; Schacter et al., 2012).

Although related, the possible past condition in Addis et al.'s study was importantly different from the mental operation we normally associate with eCFT. By randomly recombining fragments from past events, participants may have been asked to imagine not only events that did not happen, but also events that could not have happened at all. Nevertheless, the hypothesis that core regions of the DN were preferentially engaged during eCFT was further supported by two studies (De Brigard, Addis, Ford, Schacter and Giovanello, 2013; Van Hoeck et al., 2013) directly asking participants to imagine alternative ways in which past personal events could have occurred while undergoing fMRI. Further findings have refined the nature of this initial observation. De Brigard et al. (2013) showed that the engagement of the DN was modulated by the perceived plausibility of eCFT, such that episodic counterfactual events that were perceived as more plausible-to-occur recruited core regions of the DN to a greater extent than those that were perceived as less plausible. In a subsequent study (De Brigard et al., 2015), participants were asked to engage in eCFT about themselves, other people, or objects while undergoing fMRI. The results showed little involvement of DN regions during counterfactual thinking involving objects but substantial involvement of DN regions during eCFT involving people—both self and others. However, the degree to which DN was recruited during people-based eCFT was modulated by how closely related participants were to the protagonist of the mental simulation. Specifically, person-based eCFT involving oneself as well as familiar people were more likely to engage core regions of the DN relative to person-based eCFT that involved unfamiliar people.

What could be the reason for the differential engagement of DN regions during eCFT? One possibility is that person-based counterfactual thoughts that involve oneself or familiar people are perceived as more plausible than those that involve either unfamiliar people or objects; this would be consistent with the aforementioned result suggesting greater involvement of core regions of the DN during plausible relative to implausible eCFT (De Brigard et al., 2013). Another possibility is that the mental simulations generated during counterfactual thoughts involving objects or unfamiliar people as well as implausible events recruit information and processes from semantic memory (sM) to a greater degree than they do so from eM (Addis and Schacter, 2012; Irish and Piolino, 2016). Partial support for this hypothesis comes from a recent neuroimaging study whereby counterfactual thoughts about factual or semantic statements—i.e., *semantic counterfactual thinking* (sCFT)—correlate with activation in the cuneus and caudate (Kulakova et al., 2013), rather than core regions of the DN, such as the vmPFC, the PCC, or the MTL. However, this study did not manipulate plausibility, and it only compared sCFT against non-counterfactual hypothetical statements, not against eCFT.

The purpose of the current study is to uncover the precise relationship between the neural regions engaged during plausible and implausible eCFT and sCFT. Specifically, set up as a 2×3 within-subject design, the current study allows us to compare neural activity associated with eM and sM when they are reactivated during memory recollection, plausible counterfactual generation, or implausible counterfactual generation, to answer three questions. First, we seek to determine commonalities and differences in brain activity associated with both plausible and implausible eCFT and sCFT as compared to eM and sM. Second, we directly compare both plausible and implausible eCFT and sCFT to test whether there is differential engagement of DN activity as a function of plausibility, episodicity, or both. Finally, given previous results suggesting changes in hippocampal activity as a function of perceived plausibility and amount of episodic detail during episodic future thinking (Addis and Schacter, 2008; Spreng et al., 2015; Weiler et al., 2010), we seek to investigate changes in hippocampal activity as a function of both plausibility and episodicity during eCFT and sCFT.

To these ends, three strategies to analyze the fMRI data were planned. First, a data-driven mean-centered spatiotemporal partial least squares analysis (PLS; McIntosh et al., 2004) was employed to explore reliable commonalities and differences in brain activity during the mental simulation of the six conditions of interest. We predicted that this analysis would yield three distinct patterns of brain activity: 1) segregating memory (both episodic and semantic) from counterfactual thinking (both eCFT and sCFT); 2) segregating semantic simulations from episodic simulations; and 3) separating plausible from implausible simulations. Next, to directly explore differences in brain activity for plausible and implausible eCFT and sCFT, we conducted a mean-centered PLS analysis restricted to the four counterfactual thinking conditions. We hypothesized that, if perceived plausibility is the main factor driving the engagement of DN activation during counterfactual simulations, we should see greater activity in core regions of DN during plausible relative to implausible counterfactual thinking, whether episodic or semantic. However, if it is episodicity rather than perceived plausibility that is driving the DN activity, we should expect to see greater activation of core regions of the DN during plausible and implausible eCFT relative to both plausible and implausible sCFT. Finally, a region-of-interest (ROI) analysis was planned to explore differences in hippocampal activation as a function of perceived plausibility and episodicity during both eCFT and sCFT. This ROI analysis would allow us to explore whether increased activity in the hippocampus responds to the episodic nature of the CFT simulation, its perceived plausibility, or a combination of the two.

Materials and methods

Participants

Twenty-seven individuals completed both sessions of our study (Age $M = 22.15 \pm 2.98$, 14 females). Participants were right-handed, native English speakers, and had no history of psychiatric or neurological disorders. Participants were recruited using flyers posted on campus and on a Duke recruitment website, and they gave written consent according to the requirements of the Duke University Health System Institutional Review Board. Participants received monetary compensation for their time.

Pre-scan stimulus preparation

Prior to the study, we generated a list of 180 possible life events to cue participants' autobiographical memories. The list of these possible autobiographical episodes included common events that were specific to a spatiotemporal location and had clear outcomes that would have followed directly from the event or action (e.g. "You were in a snowball fight." or "You were caught in the rain without an umbrella."). These 180 events were also classified by two independent coders (NP and GWS) on whether they could be easily manipulated in an implausible manner (see Supplementary Information). For example, "You broke something that did not belong to you" was rated as easily manipulated, while "You accidentally pushed on a pull door" was classified as difficult to make into an implausible eCFT that was still credible. Additionally, 80 true and 80 false semantic statements were created (e.g., "Combining red paint and blue paint will create purple paint." and "Rhode Island is one of the largest states in the U.S."). To minimize the chances of including semantic statements that would cue episodic autobiographical