Simulation-based mentalizing generates a ‘proxy’ self-reference effect in memory

Allan, K., Morson, S., Dixon, S., Martin, D., and Cunningham, S. J.

2016

Simulation-based mentalizing generates a ‘proxy’ self-reference effect in memory

Kevin Allan[1]*, Suzannah Morson[1], Susan Dixon[1], Douglas Martin[1] & Sheila J. Cunningham[2]

[1] School of Psychology

College of Life Sciences and Medicine

University of Aberdeen

Aberdeen, UK

[2] Psychology Division,

School of Social and Health Sciences,

Abertay University,

Dundee, UK

**Corresponding author*

Keywords: Self-reference, Episodic Memory, Ownership, Binding, Simulation.

Abstract

The self-reference effect (SRE) in memory is a cognitive bias thought to depend on functionally specialised mechanisms that enhance memory for self-relevant information. These mechanisms may, however, by engaged by ‘proxy’ when we use our own mental states to simulate those of other people, but clear evidence of memory enhancements linked to such proxy self-reference is lacking. Here, young, healthy adult participants interacted with two virtual partners, one similar and one dissimilar to each participant in terms of their opinions and beliefs. Participants then viewed pairs of objects, and were instructed to pick one either for themselves, for their similar partner or for their dissimilar partner. A surprise memory test followed that required participants to view the object-pairs again and identify which object was chosen, and for whom. Participants were then shown their partners’ object pairs again, and asked to pick the objects that they preferred. Four key findings were observed. Overlap between participants’ own choice and those made for their partner’s was significantly higher for the similar vs. dissimilar partner – revealing participants use of their own preferences to simulate the similar partners. Recollection of chosen objects was significantly higher for self vs. both partners and, critically, significantly higher for the similar vs. dissimilar partner. Finally, we replicated prior findings of enhanced source confusion (here, over object-ownership) between self and the similar partner. These findings suggest that self-reference by proxy enhances memory for non-self relevant material, and we consider the theoretical implications for functional interpretation of the SRE.

Introduction

The self-reference effect (SRE) is a well-established bias favouring recollection of information linked to one's self versus information processed with reference to other people. The SRE in memory is elicited when information is deliberately evaluated with reference to self or when incidental self-stimulus associations are formed, e.g. by ownership (Cunningham, MacDonald, Turk, & Macrae, 1998). There is considerable debate about the SRE in memory and what it may or may not signify. An early and still influential view presupposes that the SRE in memory illuminates a special functional relationship that exists between the self and long-term memory (e.g. Rogers, Kuiper and Kirker, 1977). An alternative view is that the SRE in memory is more simply a side effect of deep, elaborate encoding (e.g., Klein & Kihlstrom, 1986; Klein & Loftus, 1988; Symons & Johnson, 1997). That is, access to a rich body of semantic knowledge about one's self allows numerous associations to be formed with stimuli that facilitates their later retrieval. According to this view, the SRE results from an interaction between semantic retrieval and episodic encoding processes that is not unique to self, and therefore sheds no light whatsoever on the mnemonic properties, requirements or functional capabilities of the self (see Gillihan & Farah, 2005).

Evidence from neuroimaging studies, however, has undermined the elaborate encoding account of the SRE. But in so doing, these studies have also generated new and intriguing questions about the functional mechanisms that generate the SRE. In particular, some of the neural substrates of the SRE appear to be engaged when we use our own mental states to simulate those of other people (e.g. Mitchell, Macrae and Banaji., 2006). A key finding from early studies was that specific neural circuitry in medial prefrontal cortex (mPFC) is engaged during self-referential encoding of character trait adjectives, over and above activation in left prefrontal and temporal cortex associated with elaborate semantic encoding per se (e.g. Kelley, Macrae, Wyland, Caglar, Inati, & Heatherton, 2002; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004). Based on these findings, Heatherton, Macrae and Kelley (2004) concluded that elaborative semantic encoding does not account for the mnemonic consequences of self-referential processing, although they pointed out that this assumes mPFC activation during encoding is related to how well the material is remembered.

This link between mPFC activation at encoding and subsequent memory has been confirmed. Ventral mPFC activity during self-referential encoding of trait adjective stimuli correlates with (i.e. predicts) subsequent memory for these stimuli (Macrae et al., 2004). The finding was replicated by Benoit, Gilbert, Volle, and Burgess (2010) , and recently by Bergstrom. Vogelsang, Benoit, and Simons (2015), who showed that regions within ventral mPFC are also differentially active during retrieval of stimuli related to self vs. other controls. Similar findings have been reported that link dorsal mPFC activity during self-referential encoding to subsequent source memory for visual object stimuli (Leshikar & Duarte, 2012; 2014), and for objects associated with self vs. other by ownership (Turk, van Bussel, Brebner, Toma, Krigolson, & Handy, 2011).

The degree to which ventral mPFC is differentially active for self vs. other, however, appears to depend upon the perceived similarity of the other to one’s self (Benoit et al., 2010; Bergstrom et al., 2015; and see Mitchell, Macrae and Banaji., 2006). By comparing responses to character traits applied to one’s self vs. friend, Benoit et al. (2010) were able to calculate how similar self and friend were, and then look for brain regions whose activity covaried significantly with this self-other perceived similarity index. They observed that high similarity was associated with reduced differential self vs. other ventral mPFC activation while judging whether character trait adjectives were descriptive of self vs. a friend control condition. Bergstrom et al. (2015) also observed differential activation of ventral mPFC during retrieval of self vs. other encoded traits. In this case, the ventral mPFC response to self-reference was specific to source memory for the person (self vs. other) to whom the character trait adjectives were applied. As was the case in the earlier study by Benoit et al. (2010), Bergstrom et al. calculated a perceived similarity index between Self and other control (in this case, US president Obama) and found that high similarity was associated with increased source confusion (i.e., a reduced ability to correctly judge the person to whom trait adjectives were applied at encoding).

To account for their finding of overlapping ventral mPFC activation for self and similar others, Benoit et al. (2010) and Bergstrom et al. (2015) used Mitchell et al.’s

(2006) proposal that the region is involved when one simulates other people's mental states based on one's own (i.e., when conceptual information about one's self is accessed in order to represent or understand other individuals). To account for the source confusion that appears to result from higher similarity between self and other, and the greater overlap in ventral mPFC activation for self and similar other, Benoit et al. and Bergstrom et al. cited a general principle of reinstatement thought to govern hippocampally-driven retrieval processing within the neocortex (e.g. McLelland, McNaughton & O'Reilly, 1995; Thakral, Wang and Rugg, 2015). According to this principle, the goal of episodic retrieval processes is to reinstate neural activity patterns from encoding episodes within the neocortical sites that were active during encoding. Increased overlap in these activity patterns during retrieval (i.e. for items encoded in relation to one's self and for a similar friend/other), might therefore induce source confusion (Johnson, Hashtroudi and Lindsay, 1993) by reducing participants' ability to discriminate between sources.

What is not clear from these neuroimaging and associated behavioural findings, is whether simulating others leads to enhanced memory, or if the SRE in memory is specific to information truly associated with one's self. In other words, it is not clear whether the SRE in memory truly reflects an encoding mechanism specialised to retain information specifically linked to self, or whether that mechanism can also promote memory for information associated with (similar) others. Serbun, Shih and Gutchess (2010) speculated that simulating others might enhance recollection of associated visual details, in their study examining recollection of source specifying information for self vs. significant other (Mother) vs. a famous other control (ex-US president Clinton). Serbun et al. (2010, Experiment 1) found that enhanced recollection of specific source specifying information for visual details of encoded objects was present to the same degree both for Self and for Mother vs. Clinton, in a modified ownership/shopping task based on the procedure introduced by Cunningham et al. (2008). Serbun et al. tentatively interpreted their finding of enhanced recollection for mother vs. famous control as being due to simulation of mother.

Although enhanced recollection of material associated with significant others can result in a reduced SRE (i.e. a reduced difference between memory for material associated with one's self vs. such others), the reasons why are far from clear. Close other-referents (i.e., mother or best friend) may trigger affective, attentional and knowledge-based processing biases that might enhance encoding and retrieval processing independently of a specific association with self (see Bower & Gilligan, 1979; Ray, Shelton, Hollon, Michel, Frankel, Gross, & Gabrieli, 2009; Sui, He, & Humphreys, 2012; Symons and Johnson, 1997). To properly test whether simulation (i.e. accessing conceptual self on behalf of others) produces a proxy SRE in memory, it would be necessary to try to remove these other potential confounds. The aim of the experiment reported here was therefore to test whether or not simulating a stranger can lead to enhanced recollection, and thereby gain insight into the nature of the mechanism producing the SRE. Specifically, this will allow us to examine whether the SRE in memory occurs by proxy when conceptual self access occurs during encoding on behalf of other individuals.

We employ a procedure recently used by Wheeler, Allan, Tsivilis, Martin, and Gabbert (2013) to gain experimental control over simulation-based mentalizing (see Methods for details). Wheeler et al. (2013) asked participants to interact via computer with two virtual partners. First, the participants answered a series of questions about their beliefs and opinions on various issues (e.g., "should nations be held responsible for acts of terror perpetrated by their citizens?"). After providing their own answer, the participants were shown computer generated feedback purporting to represent the partners' views, however these were experimentally manipulated so that one partner tended to agree with the participants' own views (similar other) while the other partner did not (dissimilar other). Participants were then asked to predict their partners' views on a new set of issues, before viewing images of household scenes for which a subsequent collaborative memory test was given. Following the memory test the participants were re-presented with the new set of issues and asked to give their own views. This allowed Wheeler et al. to quantify the extent to which participants had used their own views when predicting their partners.

Via this procedure, Wheeler et al. (2013) showed that increased mirroring of the similar partner was associated with increased trust in (i.e. conformity to) that partner's memory, rather than the dissimilar partner's memory, during the collaborative memory test. This conformity to their partner's memory lead participants to accept and report not only accurate but also inaccurate details about what they had encoded, producing a misinformation effect – i.e. a socially induced memory distortion. In the context of a social interaction with a similar other, Wheeler et al did not interpret this memory distortion as a 'malfunction' per se. Instead, conformity was interpreted as an adaptive social behaviour, regulated by explicit mentalizing mechanisms that allow simulation of others, to promote social learning and cooperation between like-minded individuals. By the same token, it is not clear whether a proxy SRE generated by simulation of similar others – if observed – should be interpreted as a malfunction within memory mechanisms 'designed' to enhance memory for self-relevant material. Alternatively, a proxy SRE generated by simulation of similar others might be functional in the context of social interaction. We return to this issue in the discussion.

In the experiment reported here, we embed Wheeler et al.'s similar/dissimilar virtual target manipulation within an object ownership task similar to that used previously by Sebrun, Shih and Gutchess (2010). Following exposure to their virtual partners opinions (i.e., Bear's, Tiger's) in the first phase of the experiment, participants were then asked to choose objects to own from a series of object pairs, and on each trial they either picked an object for themselves, for Bear or for Tiger. A surprise memory test was then given, which required participants to view the object pairs again and to recollect which particular object was previously picked, and for whom. Finally, a mirror score indicating similarity to self is calculated by asking participants to pick objects they would prefer for themselves from the object-pairs previously associated with Bear and Tiger, and then an opinion mirror score is calculated by the same task performed on the opinion statements.

By definition, the mirror scores for each partner quantify the extent to which participants used their own preference to make judgements about their partners' preferences when viewing the object stimuli during encoding. Following the previous study by Wheeler et al. (2012) we predict significantly higher mirror scores for the similar vs. dissimilar partner. Second, we predict an enhanced ability to recollect which of the two objects was picked for one's self vs. that for both virtual partners. Third, to the extent that simulating others is sufficient to enhance recollection of associated information, we predict enhanced ability to recollect objects associated with the similar vs. dissimilar partner. Finally, we predict that recollection of the objects' owners will show evidence of confusion between self and similar other (Benoit et al., 2010; Bergstrom et al., 2015).

Methods

Participants

Participants were 46 young healthy adults (35 female, mean age 19.8yrs, SD = 1.5) recruited from the University of Aberdeen undergraduate population.

Stimuli

The similar-dissimilar manipulation was created by controlling the level of agreement between self and two virtual targets personal opinions, which were elicited using the set of 190 different opinion statements employed by Wheeler et al. (2013; see Mitchell et al., 2006). Briefly, these 190 statements had been previously rated by a group of 24 young healthy adults who were asked to judge the extent to which each opinion gave insight into a person's character using a 5-point Likert scale that ranged from 'not very informative' up to 'very informative'. All 190 statements were then ranked according to their mean rating. The 30 top ranked statements (i.e. those consistently rated as providing high insight into character, for example: "I believe that nations should be held responsible for acts of terror perpetrated by their citizens") were split into three sets of ten statements, ensuring that the mean ranking of each set was equivalent. These three sets were then used in a counterbalanced design within the opinion and mentalizing phases described below. The bottom 10 ranking (i.e. the least insightful statements, for example: "I prefer to drink coffee rather than tea") were used as filler items.

A total of 96 images were selected for the object-preference and prediction phases of the study. Images were taken from Allan et al. (2012), where a more detailed description of their selection can be found. In brief, the images were grouped into 48 pairs of similar but not identical common everyday objects. Image pairs were randomly allocated to create 3 sets of 16 image pairs that were used in a counterbalanced design according to the procedure described below. The 48 image pairs shown in the prediction phase were used again as retrieval cues during the memory task.

Procedure

Participants were run in groups of 10 - 20 individuals in a large computing lab. Each participant was asked to sit facing the screen of a desktop PC (15inch monitor) on which the experimental software (EPRIME, v2) could be accessed. Participants were told that the aim of the experiment was to investigate how people come to understand one another's character, and that during the experiment they would be interacting anonymously via PC with two other individuals in the room. To ensure anonymity, their two virtual partners would be labelled throughout as 'Bear' and 'Tiger'.

To manipulate similar/dissimilar status of the two virtual targets, participants were first asked to complete an opinion rating task. In this task a series of opinion statements would be shown on the computer monitor, and under each statement they would see two boxes, one labelled 'AGREE' and the other labelled 'DISAGREE'. Their task was to click within one of the boxes using the computer mouse to indicate whether they agreed or disagreed with each statement. Participants were further instructed that, after giving each response, Bear and Tiger's names would be shown underneath one of the response boxes to indicate whether they agreed or disagreed with each statement. During this phase of the experiment each participant viewed 20 different opinion statements, 10 rated high for insight and 10 filler statements rated low for insight.

Levels of agreement/disagreement with each statement by Bear and Tiger were manipulated so that one always *shared* the view of the participant on the 10 high insight statements, while the other partner always took the *opposite* view on the 10 high insight statements. On the low insight statements both partners shared the view of the participants on 5 (of 10) trials. Hence, over all 20 opinion trials, one partner shared the participants view on 75% (15/20) trials, while the other partner took the opposite view to the participant on 75% (15/20) trials.

Following the initial opinion phase, participants were asked to utilise their acquired knowledge of Bear and Tiger in order to judge their partner's views on some new opinion statements. They were then shown a further set of 20 statements, one at a time onscreen, below which 'agree' and a 'disagree' response box were shown. All of these 20 statements were high insight, and consisted of the two unseen sets of 10 high insight statements that were not used during the prior opinion phase. Above each statement the name of one partner (i.e., Bear or Tiger) was shown, and the participants' task was to decide whether that partner would agree or disagree with the statement. 10 of the statements were given with respect to Bear's opinions, and 10 were given with respect to Tiger's opinions. Partner order (i.e. Bear vs Tiger) was randomised from one trial to the next. This phase was self-paced.

Once the opinion prediction task was complete, the object preference and prediction phase began. Participants were told that they would now see pairs of object images depicting everyday items, such as footballs, items of clothing, vehicles, foodstuffs, etc. Their task was to decide which particular item in each pair either themselves or one of their partners might prefer to own. On each trial they were shown two objects side by side, below text which stated: "Which object would [referent] prefer?". On 16 of the trials the referent was SELF, on 16 it was BEAR and on the remaining 16 it was TIGER. Across these 48 trials the order of referent was randomised for each participant. Object selection was carried out using the computer mouse to click on the object image that was preferred.

After the object-preference trials were complete, the participants were given a brief five-minute rest. Following this a surprise memory test was administered. The 48 object pairs were represented, and participants were asked to first identify which of the two objects they had previously selected in the object-preference phase. After identifying the object, participants were then asked to identify for which referent the object had been chosen ('SELF', 'TIGER' or 'BEAR').

Finally, the participants completed two tasks designed to assess simulation ('mirror scores'). On the first task they were shown the 20 opinion statements previously presented for prediction of Bear and Tiger's opinions. In this instance, participants were asked to indicate whether they themselves agreed or disagreed with each statement. Likewise for the second task, participants were shown the 32 (i.e. 16 per target) object pairs for which they had previously tried to predict their partner's preferences. For each of these pairs, they were asked to pick the one object that they would prefer to own themselves. The entire procedure took on average 55minutes.

Results

Mirror Scores

Following Wheeler et al. (2013), we use the term mirror score to refer to the overlap between a participant's own responses and what they predicted for their partners'. This provided two different mirror scores: one for opinions and one for object preferences. Replicating Wheeler et al. (2013), the opinion mirror score for the similar partner (81.7%, SD = 12.3) was significantly higher than the 50% null value given the binary agree/disagree response option (*t*(45) = 17.4, p < 0.001). The opinion mirror score for the dissimilar partner was 44.3% (SD = 19.6), lower than the 50% null value and approaching significance (*t*(45) = 1.95, *p* = 0.057). The opinion mirror scores for each partner were significantly different from one another (*t*(45) = 9.81, *p* < 0.001). Participants therefore made systematic use of their own opinions when explicitly mentalizing about the opinions of the similar vs. dissimilar targets.

Analysis of the object mirror scores revealed the same effect of partner similarity as shown in the opinion mirror scores. The object mirror score for the similar partner was 66.4% (SD = 18.9), significantly higher than the 50% chance value (*t*(45) = 5.90, *p* < 0.001). The object mirror score for the dissimilar partner was lower than 50%, at 45.2% (SD = 22.6) although this failed to reach significance (*t*(45) = 1.43, *p* = 0.16). The object mirror scores for each partner were also significantly different from one another (*t*(45) = 4.13, *p* < 0.001). Together, the object and opinion mirror scores suggest that participants used the similar-dissimilar distinction as a rationale for simulation. We take this as essential confirmation that our manipulation of partner similarity during the initial exposure phase of the experiment worked as intended. In particular, by systematically altering participants' use of their own preferences when predicting their partners' preferences during encoding in the memory phase of the experiment. The question now is whether differential simulation of the similar and dissimilar partners is accompanied by the predicted pattern of memory effects.

Chosen-Object Recollection

One-way ANOVA, using the factor of target (self vs. similar partner vs. dissimilar partner), upon the percentage correct chosen-object recollection rates shown in Figure 1 gave a main effect of target (*F*(2,90) = 12.67, *p* < 0.001). Follow-up t-tests showed that chosen-object recollection was significantly higher for self-owned objects (93.2%, SD = 8.52) versus objects picked either for the similar (89.5%, SD = 13.4) or the dissimilar (85.1%, SD = 13.6) targets (*t*(45) = 2.33, *p* = 0.024 and *t*(45) = 4.70, *p* < 0.00005, respectively). As predicted, chosen-object recollection was also significantly enhanced for similar relative to dissimilar targets (*t*(45) = 2.89, *p* = 0.006).

Ownership Judgement

The chosen-object accurate recollection rates shown in Figure 1 for each referent may be decomposed further according to the accuracy of the subsequent ownership judgement. The accurate owner rates are shown in Figure 2a. The rates of incorrect owner judgements for correct chosen-objects belonging to the two targets is shown in Figure 2b, and the incorrect owner judgements for correct chosen-objects belonging to Self are shown in Figure 2c.

As is clear from Figure 2a, correct ownership rates are selectively enhanced for objects belonging to self vs. both the similar and dissimilar targets. This pattern was confirmed by one-way ANOVA which gave a significant effect of target (*F*(2,90) = 7.75, *p* = 0.001). Follow-up t-tests showed that correct Owner judgements were significantly enhanced for objects associated with self (mean 81.6%, SD = 14.9) vs. both partners (similar partner mean 71.6, SD = 21.2, *t*(45) = 3.58, *p* = 0.001; dissimilar partner mean 74.9%, SD = 20.3, *t*(45) = 2.35, *p* = 0.023). However, no effect of partner similarity was observed (*t*(45) = 1.60, *p* = 0.12).

Although the accurate other owner judgement rates shown in Figure 2a did not differ statistically according to target similarity, differences do emerge between the similar vs. dissimilar target conditions when we examine the pattern of incorrect judgments shown in Figures 2b and 2c. In particular, Figure 2b shows that the pattern of errors, and the overall error rate, differ for the similar vs. dissimilar others. Using a 2x2 ANOVA (with factors of target (Similar / Dissimilar) and incorrect owner (Self vs. Other) we observed significant main effects of target (F(1,45) = 17.2, p < 0.001), incorrect owner (F(1,45) = 4.14, p < 0.05) and an interaction between the two factors (F(1,45) = 4.0, p = 0.05). The incorrect owner effect reflects a reduced tendency to incorrectly ascribe ownership of these objects to Self vs. other targets per se. The main effect of target (similar vs. dissimilar) reflects a higher overall error rate for the similar vs. dissimilar target. (17.9% (SD = 16.3) vs. 10.2% (SD = 13.4), *t*(45) = 4.15, p < 0.0005). This inflated error rate is due specifically to differences in the incorrect self owner judgments, which is reflected by the significant two-way ANOVA interaction.

The two-way interaction occurred because objects encoded in reference to the similar vs. dissimilar target are more likely to receive an incorrect self owner judgment (8.4% (SD = 11.11) vs. 2.3% (SD = 4.8), *t*(45) = 4.61, p < 0.0005), while the rate of other incorrect judgments (i.e. to similar vs. dissimilar) do not differ (t(45) = 1.0, p = .31). That is, participants showed a significantly elevated tendency to claim, as their own, objects belonging to the similar vs. dissimilar target. In Figure 2c, we show the pattern of incorrect owner judgments for objects belonging to Self, which reveals a significantly elevated rate of similar vs. dissimilar owner judgments (9.5% (SD = 10.0) vs. 2.0% (SD = 3.7), *t*(45) = 4.94, *p* < 0.0005). That is, errors in ownership judgment for objects that actually belonged to self were biased towards the similar target.

A final check upon the data was carried out to establish whether there was any effect of target (self vs. similar vs. dissimilar) on how long images were viewed while participants chose objects during the encoding phase. The mean reaction times (RTs), respectively, for each condition were 3.8s (SD = 2.8), 4.2s (SD = 3.8) and 3.8s (SD = 2.2). A one-way ANOVA revealed no effect of condition (*F*(2,90) = 0.30, *p* = 0.74). Hence, differences in image encoding duration do not account for the pattern of target effects on object and owner recollection.

Discussion

We observed a pattern of results which confirms that simulating another individual's object choices based on one's own preference is associated with enhanced recollection of those objects. Specifically, we found that the ability to judge which of two previously seen (hence equally familiar) visually similar objects had previously been picked was enhanced for objects belonging to self vs. others, replicating and extending prior work using the ownership paradigm (Cunningham et al., 2008; Cunningham, van den Bos, & Turk, 2011; Serbun, Shih and Gutchess, 2010; Turk et al., 2011). Critically, we also found that recollection of picked objects was significantly higher for objects belonging to the similar vs. dissimilar other. To determine whether preferences of similar others were simulated during encoding via use of the participants own object preferences, we calculated a mirror score (Wheeler et al., 2013) reflecting the overlap between participants own preferences and the simulated preferences from the encoding phase of the experiment. These mirror scores confirmed that during encoding, participants used their own object preferences significantly more often for the similar vs. dissimilar partner. Moreover, this pattern was also present in mirror scores when participants predicted their partners' opinions and beliefs. Finally, participants' ability to recollect the owner of the objects showed both a selective enhancement for self-owned vs. other owned objects, but no evidence of enhanced memory for the similar vs. dissimilar partner. Rather, we observed errors in ownership judgement indicating increased source confusion between self and the similar other. Ownership judgements were biased, such that participants were significantly more likely to judge that they were the owner of the similar vs. dissimilar person's objects, and were also significantly more likely to falsely ascribe ownership of self-owned objects to the similar vs. dissimilar target.

Our findings suggest that memory enhancements linked to self-reference are not circumscribed or limited to material that (notionally) belong to one's actual self. Instead, mentalizing about similar others based on one's own preferences appears to be sufficient to enhance subsequent recollection – although not to an identical degree, given that memory for one's own objects was significantly enhanced for self vs. similar other. Hence, the effect of simulating others was to reduce the size of the

SRE for self vs. similar other, compared to the size of the SRE for self vs. dissimilar other. Findings such as this have been reported previously in the literature when the control condition in SRE experiments involves a significant other such as one's Mother (e.g. Bower and Gilligan, 1997; Serbun, et al., 2010; Symons and Johnson, 1997). But our present findings show, for the first time to our knowledge, that reductions in the size of the SRE can be elicited for strangers, via a proxy SRE in memory triggered by conceptual self access on behalf of similar others. These findings avoid potential confounds from pre-existing affective, attentional or knowledge-based processes linked to significant others that in prior work may have potentially enhanced memory and reduced the size of the SRE.

In addition to enhanced memory for objects belonging to the similar other, we also observed biased errors in ownership judgment indicating greater confusion between self/similar target than self/dissimilar target - as was reported by Benoit et al. (2010) and Bergstrom et al. (2015) using source memory (for target, self vs. other) for character trait adjectives. So our present findings confirm, using a different SRE paradigm, that simulating similar others can lead to source confusion. We see no reason not to adopt the interpretation of such source confusion given by Benoit et al. (2010) and Bergstrom et al. (2015). Increased overlap in encoding operations and associated neural activity patterns for self and similar other during encoding may have led to increased overlap in retrieval processing, inducing source confusion over object-ownership. This pre-supposes a general principle of reinstatement that accounts for our ability to re-experience a past event via re-activation of neural activity patterns specific to encoding episodes (e.g. McLelland, McNaughton & O'Reilly, 1995; Thakral, Wang and Rugg, 2015). Source memory – inferences based on retrieved information specific to a past episode – may thus be confounded if there is increased overlap in retrieved information for self and similar other. If correct, our present findings along with those previously reported by Benoit et al. (2010) and Bergstrom et al. (2015) suggest that the encoding mechanism generating the SRE does not selectively operate upon information truly, specifically, associated with one's self. Put another way, it appears that the special relationship between one's conceptual self and memory (Heatherton et al., 2004) can extend to similar others, via simulation.

Before we consider functional interpretation of the SRE in light of these findings, it is worth explicitly acknowledging that the behavioural data described here cannot of course provide a direct link to specific neural mechanisms, whether these be within the mPFC (or more generally in terms of overlap in neural activity during encoding and retrieval). But to the extent that our findings reflect conceptual self access during the object preference decision at encoding, we predict that future related fMRI work should reveal differential ventral mPFC activation under similar conditions. That is, introducing the element of choice into the ownership paradigm as we have done here (and see Serbun et al., 2010) should lead to ventral mPFC activation sensitive to self-other similarity in addition to, or instead of, dorsal mPFC activation associated with self vs. other ownership in the basic ownership task (Turk et al., 2011).

Turning finally to functional interpretation of the SRE in memory, it seems natural to ask whether the proxy SRE in memory should be considered as a socially induced memory distortion, a memory malfunction perhaps, that uselessly elevates recollection of items that do not belong to self and produces confusion over ownership. Consistent with this socially induced memory distortion account, our finding of errors in ownership judgment indicating confusion between self/similar other, and prior findings of source confusion over trait adjectives associated with similar individuals (Benoit et al., 2010; Bergstrom et al., 2015) suggest that the mnemonic consequences of simulation may be better thought of as a cost, rather than a (functional) benefit. Similar patterns have also been observed previously in work from Perfect and colleagues (e.g. Stark and Perfect, 2007; Perfect, Field and Jones, 2009) in their work on plagiarism effects. In these studies, individuals mistakenly come to believe that concepts and ideas originated with themselves, rather than with another person. This source confusion – i.e. wrongly attributing an idea or a concept to ones self rather than another – resembles the object ownership biases we observed in the present experiment[1].

Memory, however, need not always function to maintain an accurate representation of the past, and some kinds of distortion within episodic memory may in fact be

[1] We would like to thank a reviewer for pointing out the work of Perfect and colleagues to us.

adaptive, as has been argued in relation to future simulation (Schacter, Guerin & St Jacques, 2011). To make this point more concrete, Conway (2005) has proposed a self-memory-system (SMS) in which conceptual self-knowledge is integral to a ‘working self’ that regulates the transfer of goal-relevant information to and from long-term memory. In these terms, the ownership task may produce a SRE in memory by activating a goal to keep track of one’s resources – i.e. items that belong to one’s self. If the function of this goal-based mechanism is to enhance memory for information truly linked to one’s self, then our current findings indicate that engaging in simulation apparently undermines this function, in two senses. Firstly, items that belong to similar others are confused with items that belong to one’s self and, secondly, one’s recollection of episodic detail is enhanced for objects that are not truly related to one’s self - only by proxy.

However, although this ‘malfunction’ account of the proxy SRE is viable, it is not clear that we need to interpret the proxy SRE as a malfunction at all. It seems just as reasonable instead that the proxy SRE could give us insight into the social functioning of memory. As noted in the introduction, in a prior study from our lab Wheeler et al (2013) found other evidence that simulation influences memory. In that study, we observed that simulating similar others influenced conformity to their memory for a shared experience. Subjective trust in another similar person’s memory (i.e. conformity to their memory judgments) was enhanced, even though there were no objective grounds (e.g. increased accuracy) to trust that person’s memory. We used this finding, as well as our own and others prior work on memory conformity, to argue for a reorientation in how we theorize about social influences upon memory (Wheeler et al., 2013). Socially induced distortions of memory, such as acceptance of misinformation from another person (e.g. Allan et al., 2012), or indeed the source confusions demonstrated here between one self and similar others (and see Benoit et al., 2010; Bergstrom et al., 2015) can be and usually are viewed as a malfunction.

But instead of viewing another person as a source of undesirable influence over memory, one may just as readily view them as a potential source of better information about the past than that provided by our own memory. In other words, even from a perspective emphasizing accuracy, it is reasonable to be open to another person’s knowledge about the past if that can be used to improve our own mental model of reality (Allan et al., 2012; Frith and Frith, 2012). Similarly, a willingness to accept another person’s version of the past may also have indirect adaptive benefits if it fosters trust and cooperation between individuals. Reviewing the literature, we came to the conclusion (Wheeler et al., 2013) that conformity to other people’s memory appears to be closely regulated by explicit mentalizing mechanisms sensitive to the accuracy of our own memory and that of other individuals, and sensitive to the social status of other individuals with whom we interact.

It’s possible that the proxy SRE demonstrated here may turn out to have a similar *functional* basis in social cognition, although what this may be remains to be elucidated and empirically tested. There are, however, various promising ways to test the hypothesis. For example, the enhanced recollection of objects owned by self and similar, trusted, others may play a functional role in competition and cooperation over resources. Accordingly, we would predict that the SRE, and the proxy SRE, should systematically co-vary with factors, such as one’s physical dominance, known to be important for inter-individual resource competition (e.g. Watkins, Jones, & DeBruine, 2010). We are currently testing predictions based upon this hypothesis.

References

Allan, K., Jones, BC., DeBruine, L. & Smith, DS. (2012). Evidence of adaptation for mate choice within human females' long-term memory. *Evolution and Human Behavior*, 33, 193-199.

Allan K., Midjord J.P., Martin D., & Gabbert, F. (2012) Memory conformity and the perceived accuracy of self versus other. Memory & Cognition, 40: 280-286.

Benoit, R. G., Gilbert, S. J., Volle, E., & Burgess, P. W. (2010) When I think about me and simulate you: Medial rostral prefrontal cortex and self-referential processes. *NeuroImage, 50,* 1340-1349.

Bergstrom, Z. M., Vogelsang, D. A., Benoit, R. G., & Simons, J. S. (2015). Reflections of oneself: Neurocognitive evidence for dissociable forms of self-referential recollection. *Neuroscience, 25,* 2648-2657. doi: 10.1093/cercor/bhu063

Bower, G. H., & Gilligan, S. G. (1979). Remembering information related to one's self. *Journal of Research in Personality 13,* 420-432. doi:10.1016/0092-6566(79)90005-9

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language, 53,* 594–628.

Cunningham, S. J. (in press). The function of the self-attention network. *Cognitive Neuroscience.* doi: 10.1080/17588928.2015.1075485.

Cunningham, S. J., Brady-Van den Bos, M., Gill, L., & Turk, D. J. (2013). Survival of the selfish: Contrasting self-referential encoding and survival processing. *Consciousness and Cognition, 22,* 237-244. doi: 10.1016/j.concog.2012.12.005.

Cunningham, S. J. Turk, D. J., Macdonald, L. M. & Macrae, C. N. (2008). Yours or mine? Ownership and memory. *Consciousness and Cognition, 17,* 312-318. doi: 10.1016/j.concog.2007.04.003.

Cunningham, S. J., Van den Bos, M., & Turk, D. J. (2011). Exploring the effects of ownership and choice on self-memory biases. *Memory, 19,* 449-461. doi: 10.3758/s13421-012-0279-0.

Frith C.D., & Frith U. (2012) Mechanisms of social cognition. Annual Review of Psychology, 63: 287-313.

Gillihan, S & Farrah, M. J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin, 131,* 76-9. doi:10.1037/0033-2909.131.1.76

Heatherton, T. F., Macrae, C. N., & Kelley, W. M. (2004). What the social brain sciences can tell us about the self. *Current Directions in Psychological Science, 13,* 190-193.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14,* 785-794. doi:10.1162/08989290260138672

Leshikar, E. D. & Duarte, A. (2012). Medial prefrontal cortex supports source memory accuracy for self-referenced items. *Social Neuroscience, 7,* 126-145. doi:10.1080/17470919.2011.585242

Leshikar, E. D. & Duarte, A. (2014). Medial prefrontal cortex supports source memory for self-referenced materials in young and older adults. *Cognitive, Affective, & Behavioral Neuroscience, 14,* 236-252.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14,* 647-654. doi: 10.1093/cercor/bhh025

McCLelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why are there complementary learning systems in the Hippocampus and Neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50,* 655-663.

Perfect, T.J., Field I., & Jones, R. (2009) Source credibility and idea improvement have independent effects on unconscious plagiarism errors in recall and generate-new tasks. J. Exp. Psychol. Learn. Mem. Cogn. 35, 267-74.

Ray, R. D., Shelton, A. L., Hollon, N. G., Michel, B. D., Frankel, C. B., Gross, J. J., & Gabrieli, J. D. E. (2009) Cognitive and neural development of individuated self-representation in children, *Child Development, 80,* 1232–1242.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35*, 677-688. doi:10.1037/0022-3514.35.9.677

Schacter, D.L., Guerin, S. A., & St Jacques, P.L. (2011). Memory distortion: an adaptive perspective. Trends in Cognitive Sciences, 15, 467-474.

Serbun, S. J., Shih, J. Y., & Gutchess, A. H. (2011). Memory for details with self-referencing. *Memory, 19*, 1004-1014.

Stark, L.J., & Perfect, T.J. (2007) Whose idea was that? Source monitoring for idea ownership following elaboration. Memory, 15, 776-83.

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1105.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*, 371-394. doi:10.1037/0033-2909.121.3.371

Thakral, P.P., Wang, T.H., & Rugg, M.D. (2015). Cortical reinstatement and the confidence and accuracy of source memory. *Neuroimage*, 109, 118-129.

Turk, D. J., van Bussel, K., Brebner, J. L., Toma, A., Krigolson, O., & Handy, T. C. (2011). When IT becomes MINE: Attentional biases triggered by object ownership. *Journal of Cognitive Neuroscience, 12,* 3725-3733. doi:10.1162/jocn_a_00101

Watkins, C. D., Jones, B. C. & DeBruine, L. M. (2010). Individual differences in dominance perception: Dominant men are less sensitive to facial cues of male dominance. *Personality and Individual Differences, 49,* 967-971.

Wheeler, R., Allan, K., Tsivilis, D., Martin, D. & Gabbert, F. (2013). Explicit mentalizing mechanisms and their adaptive role in memory conformity. *PLoS ONE*, 8, e62106.

Figures

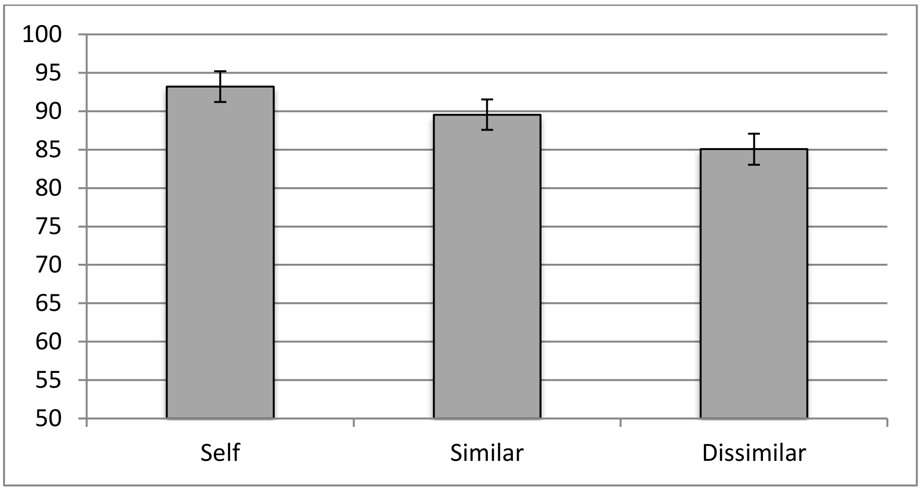

Figure 1: Percentage correct ability to recollect which of the two encoded objects had previously been chosen for self versus the similar and dissimilar targets (error bars +/- SEM).

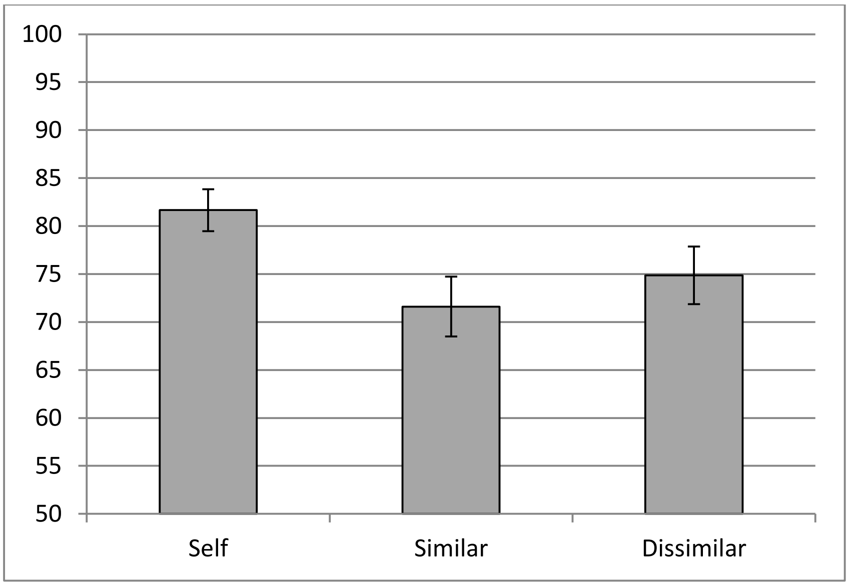

Figure 2a: Percentage accurate source memory for the owner of correctly identified chosen-objects for self versus the similar and dissimilar targets (error bars +/- SEM).

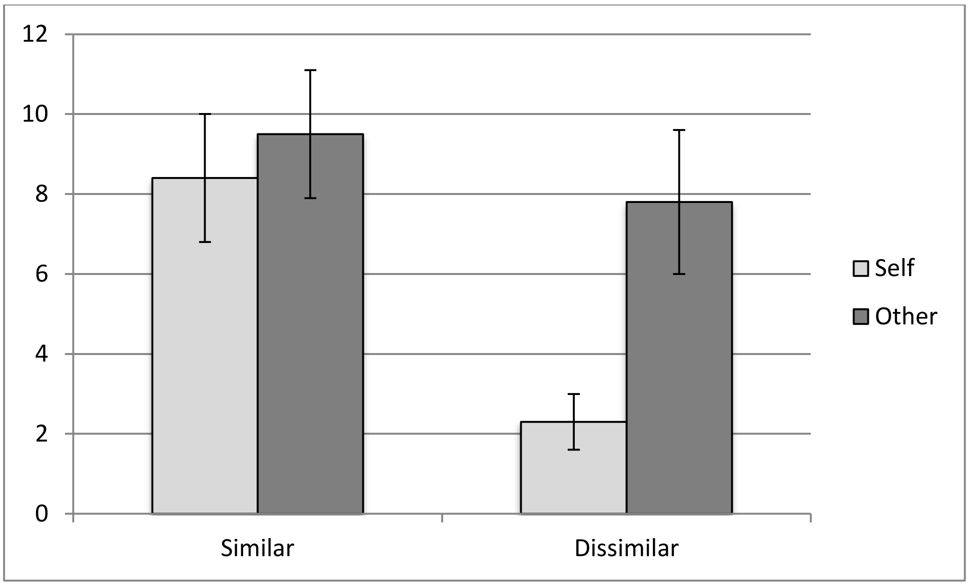

Figure 2b: Percentage rates of incorrect owner judgements for correctly identified objects belonging to the similar versus dissimilar targets (Self = light grey; Other = dark grey, error bars +/- SEM).

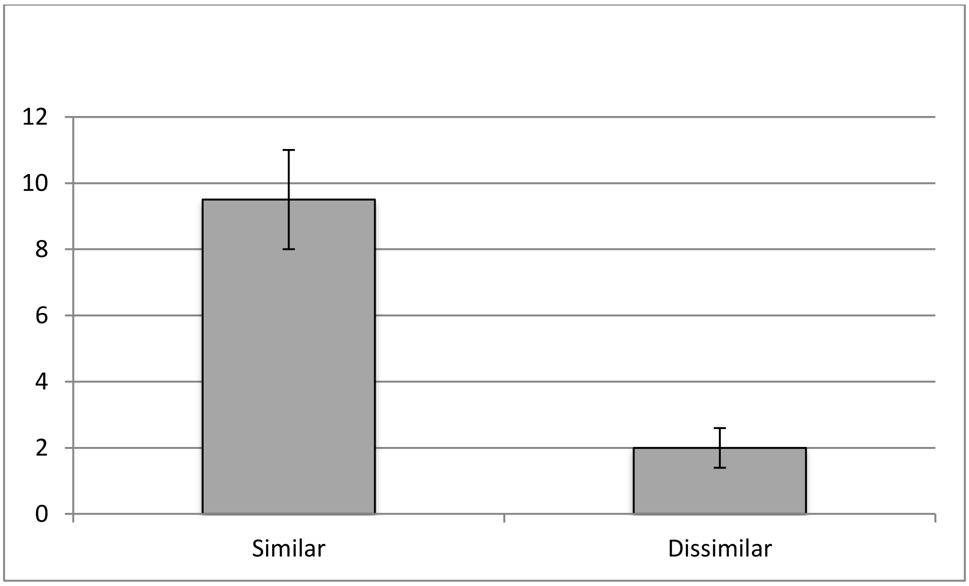

Figure 2c: Percentage rates of incorrect owner judgements for correctly identified objects belonging to self (error bars +/- SEM).