When our thoughts are not our own: Investigating agency misattributions using the Mind-to-Mind paradigm

Lauren Swiney*, Paulo Sousa
Queen's University, Belfast, UK.

*Corresponding author at: Institute of Cognition and Culture, Queen's University Belfast, 2-4 Fitzwilliam Street, Belfast, BT7 1NN, UK. *Tel:* + 44 28 9097 3701.
E-mail address: lswiney01@qub.ac.uk (L. Swiney).

Abstract: At the core of the sense of agency for self-produced action is the sense that *I,* and not some *other agent*, am producing and directing those actions. While there is an ever-expanding body of empirical research investigating the sense of agency for bodily action, there has, to date, been little empirical investigation of the sense of agency for thought. The present study uses the novel Mind-to-Mind paradigm, in which the agentive source of a target thought is ambiguous, to measure misattributions of agency. Seventy-two percent of participants made at least one misattribution of agency during a five-minute trial. Misattributions were significantly more frequent when the target thought was an arousing negative thought as compared to a neutral control. The findings establish a novel protocol for measuring the sense of agency for thought, and suggest that both contextual factors and emotional experience play a role in its generation.

Keywords: Sense of agency; Thought insertion; Auditory hallucinations; Schizophrenia; Mind-to-Mind paradigm.

1. Introduction

In recent years there has been an explosion of interest in the cognitive underpinnings and phenomenology of what is widely termed the sense of agency, attracting the attention of researchers from a variety of disciplines, including philosophy, psychology, and neuroscience (Balconi, 2010; David, Newen, & Vogeley, 2008; Dijksterhuis, Preston, Wegner, & Aarts, 2008; Farrer et al., 2003; Gallagher, 2012). Fundamental to characterizing the sense of agency is the notion of agent attribution; when I act, I have the sense that *I,* and not some *other* agent, am causing and directing those actions. This self-attributional element of the sense of agency—the sense of self-agency—has been of particular interest to theorists and researchers in large part because of puzzling reports given by psychiatric patients indicating that movements of their body or thoughts in their mind are in some sense "not theirs". Such reports have been held by various theorists to constitute attributions of external agency for a given bodily movement or thought, and thus to indicate a disruption to the normal experience of agency. Specifically, the diagnostic symptom of delusions of control—in which the patient reports that parts of their body are being moved by another agent—is widely interpreted as indicating a disruption to the sense of agency for bodily action (e.g. Frith, 2005), while the delusion of thought insertion—in which the patient reports that another person's thoughts are being inserted into their mind—has been held by many to indicate a disruption to the sense of

agency for thought (e.g. Gallagher, 2004; Sousa & Swiney, 2011).[1] Additionally, many theorists argue that Auditory Verbal Hallucinations (AVH)—the experience of hearing a voice in the absence of appropriate stimuli—may be characterized as external attributions of agency for verbal thought (Jones & Fernyhough, 2007; Langland-Hassan, 2008; Stephens & Graham, 2003; Synofzik, Vosgerau, & Newen, 2008a). While the term 'thought' is used across the cognitive sciences to refer to many different types of mental event, the literature on the sense of agency broadly pertains to occurrent mental events, excluding perceptual states, of which we are first-personally aware. Various notions of verbal thought/inner speech have been particularly prominent (e.g. Jones & Fernyhough, 2007; Synofzik, Vosgerau, & Newen, 2008b; Vosgerau & Newen, 2007).

In the case of bodily action, the recent explosion of conceptual work and the explication of relevant cognitive mechanisms involved in the sense of agency has been accompanied and enhanced by the development of a variety of ingenious experimental paradigms (e.g. Blakemore, Oakley, & Frith, 2003; Farrer & Frith, 2002; Moore & Haggard, 2010; Wegner, Sparrow, & Winerman, 2004). By contrast, while conceptual and theoretical debates relating to the case of *mental* action have developed, there has been a noted dearth of accompanying empirical investigation, perhaps due to the difficultly of submitting mental action to experimental investigation (Frith, 2005; 2012). In fact, we are aware of only one attempt in the existing literature to experimentally investigate the sense of agency for thought.[2] Sugimori, Asai and Tanno (2011a) have recently argued that the existing Deese-Roediger-McDermott (DRM) memory paradigm (Roediger & McDermott, 1995), can be considered a measure of the sense of agency for thought. While this use of the DRM constitutes an inventive attempt to measure the sense of agency for thought, it provides a measure that is arguably problematic, and at best indirect. The DRM paradigm involves presenting individuals with a list of words which are closely related to a specific other word known as the *critical lure* (e.g. the words hill, climb valley, summit etc. are closely related to the critical lure "mountain") and then later asking participants if the critical lure was presented in the original list. Because the presentation of the original word list activates the semantic network, causing the critical lure to be more accessible, normal healthy controls often incorrectly recall being presented with the critical lure (Roediger & McDermott, 1995). Sugimori et al. argue that false memories of the critical lure can be taken as an indication of a weak sense of agency for thought, since the participant fails to recognize that the familiarity of the word is the result of earlier self-generated mental activity. But since the semantic activation of the critical lure during the initial word list presentation occurs "unconsciously and automatically" (Sugimori, Asai, & Tanno, 2011a, p. 693), it is unclear whether it constitutes thought in a sense relevant to understanding first person phenomenology. Moreover, differences in this activation of the semantic network that may affect memory of the critical lure need have nothing to do with the experience of agency; for instance, memory may be affected by the degree to which the semantic activation engages the acoustic components of the language system, an especially relevant possibility since the original word list was read aloud to participants, and one which might also explain their primary finding that false memories of the critical lure were associated with proneness to AVH. Finally, the viability of the measure rests on the proposal that a weak sense of agency over the original semantic activation will also lead to a weak "agency memory" at the moment of recall, but the theoretical framework on which this assumption rests has only been demonstrated for the bodily act of speech (Sugimori, Asai, & Tanno, 2011b). Overall, it is not clear that the DRM paradigm provides a suitable measure of the sense of agency for thought. It certainly does not allow the measurement of either occurrent or explicit agency attributions or the exploration of phenomenological dimensions of the sense of agency.

The primary aim of this article is to fill this gap in the experimental literature by developing a paradigm that allows direct measurement of explicit attributions of agency for specific episodes of conscious thought. A key challenge in developing such a protocol is to provide plausible ambiguity

[1] But see Bortolotti and Broome (2009) and Fernandez (2010) for alternative characterizations.

[2] Although there *are* existing experimental paradigms which investigate mechanisms which may contribute to the sense of agency, such as those which measure the ability to intentionally suppress memories (e.g. Paulik, 2008).

as to the agentive source of an experienced thought. After all, who else but me could think a thought in my mind? In fact, there are real-world settings in which just such plausible ambiguity is raised. A prior belief in some sort of supernatural agent is a unifying feature of religious and spiritual systems (Boyer, 2002), and such belief systems make room for a belief in a plausible mechanism for external agency experiences for thought, to a greater or lesser degree. In some cases a belief in God's ability to directly produce thoughts in a worshipper's mind is explicit, and in such contexts it is perfectly sensible to ask questions about the agentive source of a given thought (Luhrmann, 2005). While these examples rely on a supernatural mechanism for agentive disruption to thought (i.e. supernatural agents), they do suggest a possible approach for experimental investigation. The novel experimental protocol used in the present study will provide a supposedly scientifically plausible mechanism by which one person can experience thoughts that anther person is thinking. Participants wear a fake transcranial magnetic stimulation (TMS) headpiece which forms part of a phony 'Mind-to-Mind Interface' supposedly connecting their mind to that of another person using a combination of EEG and TMS technologies. The participant is asked to sit quietly and attend to their thoughts and to report, by clicking a mouse, whenever they judge that they have "experienced a thought that the other person is thinking". The report constitutes an explicit judgment related to the act of thinking the thought, and can be considered an attribution of agency. Since the Mind-to-Mind Interface is phony and there is no other participant, any attribution to external agency is in fact a *mis*attribution of agency.

A further aim of the present research is to use the novel Mind-to-Mind paradigm to shed light on the mechanisms underlying the sense of self-agency for thought in the general population. A variety of cognitive accounts have recently been outlined positing specific mechanisms implicated in the sense of agency for thought. While the differing predictions made by these accounts as to the circumstances under which self-produced thought will be misattributed to an external agent have received some empirical consideration (e.g. Carruthers, 2012), they have not been directly investigated experimentally. A distinction is often drawn between 'bottom-up' accounts, which implicate subpersonal cognitive mechanisms in giving rise to a *feeling* of agency, and 'top-down' accounts, which highlight the role of inferences and contextual factors at the level of *judgments* of agency for thought.[3] One type of bottom-up account of the sense of agency for thought is based on a leading account of the sense of agency for bodily action—the 'comparator model' (CM)—which posits that the subpersonal comparator mechanisms of the motor control system function not only to specify, initiate, control and adjust bodily action (Holst & Mittelstaedt, 1950), but also to provide the internal signals underlying the sense of agency (Frith, 2005; 2012; Frith, Blakemore, & Wolpert, 2000). While the possibility of extending a CM-style account based on the workings of the motor control system to the realm of mental action was part of Frith's original CM proposal and is regularly posited in unelaborated form across the literature (e.g. Carruthers, 2011; Mathalon & Ford, 2008; Vosgerau & Newen, 2007), precise details of such an extension remain controversial and open to criticism. The most popular approach has been to posit a direct extension, suggesting that the very same subpersonal mechanisms of the motor control system contribute to the sense of agency for thought and it's disruptions (e.g. Jones & Fernyhough, 2007; Langland-Hassan, 2008; Seal, Aleman, & McGuire, 2004).[4] While details of the extensions differ, this type of account hinges on the relevant notion of thought being inner speech, and the presumption that inner speech, like outer speech, involves the motor control system. Any judgement of self/other agency for thought is derived from the feeling of agency provided by these subpersonal mechanisms. In the case of patients making misattributions of agency for thought in schizophrenia, these mechanisms are proposed to be disrupted, giving rise to a deeply unusual feeling of agency. In the non-clinical population these mechanisms are supposed to work normally, meaning that misattributions of

[3] For an analysis of the sense of agency in terms of these two distinct functional and representational levels (*feeling* of agency and *judgement* of agency) see Synofzik et al. (2008a). For a similar analysis see Bayne and Pacherie (2007). In the present article both 'sense' and 'experience' are used in a broad sense to refer to both of these levels.

[4] An alternative type of account does not implicate the actual motor control system, but instead takes the general idea of comparator mechanisms and posits a similar but separate mechanism which deals with thought (e.g. Proust, 2006).

agency based on such unusual experiences of thought should not occur (or be very rare, if the authors subscribe to a continuum view of the schizotypy phenotype (van Os, Linscott, Myin-Germeys, Delespaul, & Krabbendam, 2009)) in the non-clinical population, and by extension, in the present experimental context.

Recently Synofzik and colleagues have argued that CM-type approaches cannot fully explain the sense of agency and its disruptions, and have proposed the multifactorial weighting model (MWM) of agency, which can be considered to combine top-down and bottom-up processes (Synofzik & Vosgerau, 2012; Synofzik & Voss, 2010; Synofzik, Vosgerau, & Lindner, 2009a; Synofzik, Vosgerau, & Newen, 2008a; 2008b; 2009b). According to this model (which also applies to bodily action) a wide variety of cues are dynamically integrated to generate the sense of agency for thought. While non-conceptual feelings of agency (such as those provided by the comparator mechanisms of the motor control system) are one type of cue and may form the basis for a separate judgment of agency in some cases, cues can also include conceptual level factors such as background beliefs. The process in which the multiple factors are combined follows the framework of optimal cue integration and is similar to the process that allows robust perception of the world in the face of multiple information channels (Synofzik, Vosgerau, & Lindner, 2009a); cues are weighted according to context and reliability, and the process of integration—which can be either conscious or unconscious—resolves differences to come up with a coherent picture of the self (Synofzik, Vosgerau, & Newen, 2008b). Under this account, the sense of agency consists of the registration that we are the initiators of our actions, and its representational content encompasses action initiation, intention, performance and consequences of actions (Synofzik, Vosgerau, & Newen, 2008a). The MWM would predict that even in the absence of deficits in the motor control system, there could be cases in which contextual factors and prior beliefs may lead thoughts to be misattributed to external agency. A similar prediction would be made by the purely top-down approach advanced by Stephens and Graham (2003), which holds that attributions of agency are the result of conceptually-laden interpretive processes incorporating the agent's existing intentional states.

A third aim of the present research is to investigate a specific hypothesis regarding the effect of negative thought content on attributions of agency. The MWM of the sense of agency for thought highlights the role of emotional reactions to negative thought content, arguing that negative thought contents are hard to reconcile with the self, and thus give rise to a "strange feeling" that the thought does not "fit" (Synofzik, Vosgerau, & Newen, 2008a, p. 235; 2009b, p. 552). The rationalization module responsible for integrating the various agency cues has several options for dealing with this strange feeling, including suppressing it or, importantly, attributing the thought to another agent. This cue forms an important plank of their account of how disruptions in the MWM give rise to reports of thought insertion. Based on evidence that schizophrenia patients show increased intensity of negative emotion experiences compared to healthy subjects (Aleman & Kahn, 2005), they hold that patients will have "an abnormally strong emotional aversion against certain thought contents" (Synofzik, Vosgerau, & Newen, 2008a, p. 235), making it especially likely that the thought will be attributed to another.[5]

The Mind-to-Mind Interface protocol allows investigation of the effect of emotional thought content on attributions of agency. In the present study, participants are informed that the (supposed) other person connected to the interface is thinking a specific verbal thought consisting of a single word (the target thought); each time the participant experiences the target thought they are asked to make an explicit agentive attribution to either self or other for that thought. We manipulated the valance and arousal of the target word to test the hypothesis that more misattributions would be associated with negative than with neutral target words. In order to avoid carry-over effects between these conditions, we used a between rather than within-subject design. The MWM makes no

[5] In fact, the MWM makes a functional distinction between verbal thought and other thought (broadly specified as intentions etc.), and this proposal about the role of negative emotional reaction is discussed by Synofzik and colleagues only in terms of its effect on the sense of agency for *non*-verbal thought. However, it is not at all clear why a similar effect would not occur in the case of verbal thought.

prediction about the role of positive emotional content on attributions of agency. While recent experimental work on attributions of agency for bodily action suggest that valence affects agency attributions on a gradient, with positive actions *least* likely to be misattributed (Wilke, Synofzik, & Lindner, 2012), there is also theoretical reason arising from the literature on cognitive dissonance theory (Cooper, 2007) to suspect that strongly positive thought content may, like negative content, also provoke a feeling of "not fitting" and therefore might be *more* likely to be misattributed than neutral (Morrison, 2001). The inclusion of a third condition in which the target thought has strongly positive valence combined with high arousal allows exploration of these possibilities.

2. Method

2.1. Participants

Fifty-five participants were recruited from the Queen's University Student Union and participated in exchange for monetary compensation. All potential participants were screened for exclusion criteria including neurological disorder and diagnosis of schizophrenia. Subjects were naive to the purpose of the study. All participants gave their informed consent. Technical issues meant that one participant could not complete the study, two participants voluntarily withdrew, and debriefing revealed that four participants might have guessed the true aims of the study. The data for these seven participants has been excluded from the analysis. Data from two participants had to be discarded as they appeared to be critical outliers (see Section 3.0). The reported results are from the remaining 46 participants (21 females). Their mean age was 20.7 years (SD = 2.24; range: 18-30 years). Ethical approval was obtained from the School Research Ethics Committee (History and Anthropology), Queen's University, Belfast.

2.2. Procedure and equipment

Participants were randomly assigned to one of three between subject conditions (see Section 2.3). Upon entering the lab, participants were shown a brief technical information sheet on what we called the 'Mind-to-Mind Interface'. The sheet indicated that the interface could "transfer thoughts", allowing one person to "experience the thoughts that another person is thinking". A diagram depicted the interface as consisting of two interconnected headpieces: the first an EEG cap worn by the 'thinking' participant, the other a Transcranial Magnetic Stimulation (TMS) helmet worn by the 'experiencing' participant. After this explication of the general experimental setup, participants provided their informed consent. They were then directed to read an information sheet that provided all further instructions.

The sheet informed participants that they would wear the TMS helmet (i.e. be the 'experiencing' participant) while another person in an adjoining (unseen) lab would wear the EEG headpiece (i.e. be the 'thinker'). Visible in the lab was a mock TMS helmet. The helmet consisted of a white ski helmet with built-in headphones which had been adapted so that four black electrical wires ran from the top of the helmet and disappeared into a nearby white housing unit. Participants were told that they would wear the TMS helmet for a single five-minute trial, and that the person in the adjoining lab would be thinking a specific word (the *target thought*) throughout this five-minute trial. Participants were informed what this target thought would be (see Section 2.3). Finally, participants were told that their task during the trial was to report, by clicking a mouse, when they judged that they had "experienced a thought that the other participant was thinking". Thus, the participant was led to believe that a mechanism existed by which they could experience thoughts (specifically, the *target thought*) that were being thought by another person, and gave them an opportunity to make immediate attributions of external agency for the target thought via the click of a mouse. In reality, there was no other participant, and the mock TMS helmet had no electromagnetic capabilities.

It was stressed to participants that they should not click the mouse every time they experienced the target thought, but only when they judged that they had experienced the *other person* thinking it. In other words, they were free to think the target thought themselves. In order to reduce demand characteristics, participants were also told that, as part of the experimental design, the interface may

be set to low effectiveness or may not be activated at all during their trial. Thus, it was made clear to them that it might be perfectly appropriate not to click the mouse at all during the trial.

<Insert Figure 1 about here>

After reading the instructions, participants were positioned in a comfortable armchair, and fitted with the mock TMS headpiece (see Figure 1.). They were provided with a computer mouse over which their dominant hand could rest comfortably. Both the mouse and the helmet's built-in headphones were connected to a laptop that sat on the adjacent housing unit. The experimenter then initiated the trial on the laptop and left the lab. After 10s, the participant heard three beeps through the helmet's built-in headphones, which indicated the immanent start of the trial. Immediately after this, a continuous white noise began and continued for five minutes, indicating the duration of the trial and ensuring that all participants had the same auditory experience. Participants were instructed to keep their eyes fixed on a white wall approximately 1.5m in front of them. Any mouse clicks were automatically recorded on the laptop. At the end of the five-minute trial the white noise stopped and the experimenter returned and removed the headpiece. The participant then moved to a desk to answer a post-trial questionnaire. In order to control for diffusion effects, participants were fully debriefed as to the true aims of the experiment and the phony nature of the interface by email only once all data had been gathered.

2.3. Experimental Design

In a between-subjects design, the emotionality of the target thought was varied across three conditions according to its valence (1 = pleasant, 9 = unpleasant) and arousal (1 = low, 9 = high), as rated by the Affective Norms for English Words (ANEW) database (Bradley & Lang, 1999). In the *neutral* condition the target thought was a word (elbow) with medium valence (5.12) and low arousal (3.81). In the *negative* and *positive* conditions, the target thought was a word with either low or high valence matched for distance from neutral valence (betray, 1.68; thrill, 8.05) and matched high arousal (betray, 7.24; thrill, 8.02). All words were matched for frequency of occurrence (frequency ≤ 10, Bradley & Lang, 1999) and approximate word length.

2.4. Measures

The number of times the participant clicked the mouse during the trial provided the primary dependent measure of *misattribution frequency*. In addition, the timing (in seconds) of the first click provided a measure of *misattribution latency.* Whether or not participants clicked at least once during the trial provided a categorical measure to indicate the prevalence of misattributions across the sample. A more general measure of the sense of external agency was provided by asking participants to rate how successful they felt the interface had been at transferring thoughts (*interface success rating*) on a scale from 1 (not at all successful) to 5 (completely successful).

In addition to these measures related to the sense of agency, those who made misattributions of agency were also asked to indicate their (mis)*attribution confidence* on a scale from 1 (not at all confident) to 5 (completely confident). Those participants who did not make any misattributions of agency were asked to indicate if they suspected "there may have been some thought transfer at some point", on a scale from 1 (definitely none) to 5 (definitely some). Participants who either made misattributions or suspected that thoughts had been transferred were additionally asked to describe "what it was about the thoughts or thinking process" that lead them to make (or consider making) an external attribution. The responses to this open question provided descriptions of the unique characteristics of the misattributed thought, which were then coded according to six characteristics. The characteristics (see Table 3.) were decided prior to the study on the basis of pilot data ($N = 20$) and theoretical concerns coming from the literature. The following are examples of thought descriptions and how they were coded: "It came alongside my own thoughts and just felt different to my own" (*strange phenomenology*); "They seemed to interrupt other thoughts and my head had a weird throbbing sensation on those occasions" (*out of keeping with current thought; physical/perceptual accompaniment*); "They kept coming into my mind when I was trying to think

of something totally different" (*lack of intentional control*). Inter-rater reliability between the first author and an independent coder was more than satisfactory (Cohen's Kappa = .94). Any discrepancies were discussed and resolved.

We identified two possible confounds that might result from our manipulation of the valence and arousal of the target thought and affect the primary measure of misattribution frequency. First, evidence from the thought suppression literature indicates that emotional content can affect thought frequency (Rassin, 2005, p. 7). Thus it seemed plausible that there may be systematic differences in the number of times participants thought the target thought across conditions, a difference that could go on to impact the frequency of misattributions of agency. We therefore included a post-trial measure of overall *target thought frequency* in order to ascertain if there was a systematic difference between conditions. Similar post-trial self-assessment measures have been used in the thought suppression literature to measure thought frequency and found to be sufficiently accurate for present purposes (Rassin, 2005, p. 42). Piloting ($N = 20$) revealed that participants were comfortable using a five-point scale to describe target thought frequency (1 = *Never, 2 = Once or twice, 3 = Around 5 times, 4 = Around 10 times, 5 = More than 20 times)*.

Additionally, it seemed plausible that negative thought content would be more likely to be spontaneously suppressed than neutral. Suppression may lead to any occurrences of the target thought seeming more involuntarily, which may in turn affect the experience of agency. Borrowing again from the thought suppression literature—this time a measure normally used as a manipulation check—a measure of *target thought suppression* was included, asking participants to indicate the degree to which they tried not to think the target thought. Participants provided their rating on a visual analogue scale (*0 = Not at all, 100 = To a very large extent*).

Finally, in order to check participant's belief that the Mind-to-Mind interface was plausible, we asked participants how confident they were that the interface could function as described on a scale from 1 (*not at all confident*) to 7 (*completely confident*), irrespective of their particular experiences. This rating allowed the experimenter to further probe those participants who provided the lowest possible rating ($N = 7$). The data of any participants who appeared to have completely ruled out the possibility that the interface was capable of operating as described ($N = 4$) was excluded from analysis (see Section 2.1).

2.5. Statistical analysis

Statistical analyses were carried out using the Statistical Package for the Social Sciences (SPSS), version 18. Prior to analysis, data were screened for outliers ($M \pm 3SDs$) on the primary dependent measure (*misattribution frequency*), and data from two participants were excluded on this basis. Due to the non-normal nature of the data across the primary dependent measures, and non-interval nature of other measures (*target thought frequency*) non-parametric tests were used for all analyses.

3. Results

3.1. Descriptive statistics

Means and standard deviations for all measures are provided in Table 1. Most participants (71.7%, $N = 33$) made at least one misattribution of agency during the single five-minute trial, and 56.5% ($N = 26$) made more than one. Furthermore, 17.4% ($N = 8$) of participants did not make any misattributions but did have some suspicion that they had experienced a thought that the other person was thinking. Among those who made at least one misattribution, the mean frequency of misattributions was 3.96 ($SD = 2.31$).

Table 1
Means (and standard deviations) for all variables.

Misattribution frequency	Misattribution latency[b]	Interface success ratings[a]	Misattribution confidence[a]	Target thought frequency[a]	Target thought suppression[d]	Belief in interface[c]

2.87 (3.42)	107.61 (77.34)	2.15 (0.89)	2.84 (0.77)	2.38 (0.77)	61.13 (21.30)	3.24 (1.35)

$N = 46$, except misattribution latency and misattribution confidence $N = 33$.
[a]5-point scale; [b] Seconds; [c] 7-point scale; [d] Analogue scale, 1-100.

3.2. Correlations

Correlations are summarized in Table 2. The results revealed the expected pattern of relationships among the various experimental measures related to misattributions of agency. There was a significant positive correlation between misattribution frequency and ratings of interface success. Misattribution frequency was significantly *negatively* correlated with misattribution latency, indicating that those who made the most misattributions also made them sooner. There was also a significant negative correlation between interface success ratings and misattribution latency. Finally, there was a significant positive relationship between misattribution confidence and both misattribution frequency and interface success ratings.

Belief that the Mind-to-Mind interface could function as described was significantly positively correlated with misattribution frequency, misattribution confidence and interface success ratings, and had a significant negative relationship with misattribution latency.

Neither target thought suppression nor target thought frequency was significantly correlated with misattribution frequency, nor with any of the other measures relating to agency misattributions, though they were significantly correlated to one another.

Table 2
Kendall's Tau correlations between variables under investigation.

Variable	1	2	3	4	5	6	7
Misattribution frequency	—	-.51***	.69***	.36*	.17	.17	.48***
Misattribution latency		—	-.29*	-.17	-.14	-.04	-.27*
Interface success ratings			—	.47**	.21	.14	.59***
Misattribution confidence				—	.22	-.17	.54***
Target thought frequency					—	.33**	.26*
Target thought suppression						—	.052
Belief in Mind-to-Mind Interface							—

Note: $N = 46$, except comparisons with misattribution latency and misattribution confidence, $N = 32$.
* $p < .05$
** $p < .005$
*** $p < .001$

3.3. Effect of target thought content

As predicted, those in the negative condition made more frequent misattributions of agency than those in the neutral condition ($Mdn_{neg} = 3.00$; $Mdn_{neu} = 1.00$; $U = 59.5$, $z = 2.24$, $p = .013$, $r = .32$, one-tailed). Also in line with our hypothesis, interface success ratings were higher in the negative condition as compared to the neutral condition ($Mdn_{neg} = 3.00$, $Mdn_{neu} = 2.00$; $U = 61.5$, $z = -2.23$, $p = .017$, $r = .32$, one-tailed). While those in the negative condition made their first misattribution more quickly than those in the neutral condition ($Mdn_{neg} = 66$, $Mdn_{neu} = 126$), this difference was not significant ($U = 41.5$, $z = -1.136$, $p = .13$, one-tailed). Finally, while a greater proportion of those in the negative condition made at least one misattribution of agency (87%) than those in the neutral condition (60%), this difference was not significant ($p > .15$, one-tailed Fisher's exact test).

We did not have a specific hypothesis regarding the relationship between the neutral and positive conditions on the primary dependent measures of the sense of agency. Exploratory analysis revealed that while those in the positive condition made more misattributions than those in the neutral condition ($Mdn_{pos} = 2.50$), this difference was not significant ($U = 87.00$, $z = -1.34$, $p = .16$, two-tailed). Interface success ratings in the positive condition were the same, on average, as the

neutral condition (Mdn_{pos} = 2.00), and the proportion of participants who made at least one misattribution (69%) only slightly greater than in the neutral condition (60%; $p > .60$, two-tailed Fisher's exact test). Misattribution latency did not differ significantly between the neutral and positive conditions *(*Mdn_{pos} = 92, Mdn_{neu} = 126, U = 33.0, z = -1.25, $p > .20$, two-tailed).

Kruskall-Wallace tests revealed that there was no significant difference in either target thought frequency (Mdn_{neu} = 3.00, Mdn_{neg} = 2.00, Mdn_{pos} = 2.00; $H(2) = 1.87$ $p > .30$) or target thought suppression (Mdn_{neu} = 62, Mdn_{neg} = 58, Mdn_{pos} = 51); $H(2) = 0.89$, $p > .60$) between conditions, suggesting that neither of these variables can account for the observed differences between the negative and neutral conditions on the primary dependent measures of agency.

3.4. Descriptions of misattributed thoughts

Of forty responses provided, thirty-four were considered for coding (the others did not respond to the question; e.g. gave reasons why thoughts were *not* misattributed). All demonstrated at least one characteristic and only four responses demonstrated two characteristics. See Table 3 for characteristics and descriptive statistics.

Table 3
Descriptions of misattributed thought (N = 34).

Coding category	Examples	Percentage (frequency)
Contrary to current thought	"out of the blue" "came from nowhere"	38% (13)
Strange phenomenology	"felt a bit odd" "just felt different"	23% (8)
Lack of intentional control	"beyond my control" " as if I was being made to think"	18% (6)
Auditory dimension	"heard" "shouted"	9% (3)
Physical/perceptual accompaniment	"throbbing sensation" "physical feeling...electric pulse"	9% (3)
Emotion	"felt some emotion of the word betray"	3% (1)

4. Discussion

The present research demonstrates the viability of what we believe to be the first experimental task allowing direct measurement of attributions of agency for thought. We provided a plausible mechanism for disrupting the agentive source of thoughts—the Mind-to-Mind Interface—and asked participants to indicate when they "experienced a thought that the other person was thinking". Across conditions a strikingly high proportion of participants misattributed their own self-produced thought to another agent at least once during a single five-minute session, and confidence in these misattributions was relatively high. While the prevalence and timing of misattribution was unaffected by thought content, the results supported our hypothesis that negative thought content would be more frequently misattributed than neutral.

4.1. Prevalence and confidence of misattributions

The fact that nearly three-quarters of participants attributed their own ordinary verbal thought to an external source in a single five-minute trial—most of them more than once—was surprising. One possible explanation is that the high prevalence and frequency of agency misattributions in the present sample is merely a reflection of demand characteristics. However, instructions that the Mind-to-Mind Interface might be set to low or be completely off during any particular session should have reduced this possibility (see Section 2.0) and questioning during debriefing indicated

that these instructions had been understood. Furthermore, the relatively high confidence of the misattributions suggests that the results reflect genuine judgments of external agency. These findings appear to run counter to the widely held view that the normal experience of self-agency for thought is both straightforward and robust. More specifically, the high prevalence of external agency attributions in the present sample and context is difficult to reconcile with some cognitive accounts of the sense of agency for thought. According to the CM-based account outlined by Jones and Fernyhough (2007), the processing of inner speech in the motor control system should normally give rise to a subpersonal feeling of self-agency which accompanies all verbal thought.[6] The present results suggest that if there is such a feeling, it does not automatically lead to a corresponding *judgment* of self-agency. In addition, both Jones and Fernyhough (2007) and Langland-Hassan's (2008) CM-based accounts propose that misattributions of agency arise as the result of specific deficits in the motor control system which place verbal thought "in a very peculiar phenomenological category" (Langland-Hassan, 2008, p. 391).[7] But since these subpersonal deficits are supposed to occur only in clinical patients (and even then only intermittently), they cannot possibly account for the misattributions in the present sample. The results, therefore, demonstrate that misattributions of agency for verbal thought can occur even in the absence of any deficits in the motor control system. Moreover, it is clear that in order to characterize the etiology of the misattributions of agency across the present sample, appeal will have to be made to cognitive mechanisms other than the motor control system.

The finding of high prevalence and frequency of misattributions in this non-clinical sample is easier to reconcile with accounts which posit a role for top-down factors in the sense of agency. In contrast to CM-type approaches, the MWM account posits that the sense of agency will be impacted by a variety of factors including prior expectation, sensory information and post-hoc beliefs, all of which could be included as relevant cues. Moreover, it holds that these cues are weighted according to their relevance in the specific context (Synofzik & Vosgerau, 2012). While MWM appeals to specific deficits in the experience of emotional thought content to explain pathological misattributions in schizophrenia, it should also allow that there are other contexts in which judgments of external agency would arise in relation to perfectly ordinary episodes of thought (including verbal thought).

In the present study the participants had a prior belief that the Mind-to-Mind Interface was capable of causing them to experience thoughts that another person was thinking. The results raise the intriguing possibility that such a prior belief not only made external attributions a logically reasonable option, but was also sufficient for ordinary thoughts to be misattributed to an external agent. Such an interpretation would be in line with experimental work on the sense of agency for bodily action showing that attributions of agency rely on higher-order causal inferences on the basis of belief states and intentional stances (Aarts, Custers, & Wegner, 2005; Wegner et al., 2004), and demonstrating that prior belief can affect the sense of agency in profound and unexpected ways, even affecting subpersonal processes of attenuation of motor signals (Desantis, Weiss, Schütz-Bosbach, & Waszak, 2012). The finding that level of belief in the functioning of the Mind-to-Mind Interface was strongly correlated with misattribution frequency, misattribution latency, misattribution confidence, and interface success ratings is certainly consistent with the idea of a

[6] We are describing Jones and Fernyhough's proposal in terms of the distinction between feeling and judgement (see Synofzik et al. 2008a) but it should be noted that Jones and Fernyhough themselves refer to this feeling as an 'emotion' of self-authorship. It should also be noted that Langland-Hassan's (2008) alternative CM-based account (to be discussed in what follows) explicitly rejects the idea that there is an ongoing sense of self-agency accompanying verbal thought (p. 391).

[7] Note that while this characterization suffices for present purposes, it glosses over important functional differences between these accounts, notably whether the comparator acts to attenuate (Jones and Fernyhough) or filter (Langlan-Hassan) inner speech. In addition, for Jones and Fernyhough (2007) the "other" is represented at the level of feeling, meaning that the subsequent judgment is just an endorsement to this feeling, but leaving open to criticism how such mechanisms could give rise to conceptual content (Bayne & Pacherie, 2007). For Langland-Hassan the structural differences in the disrupted phenomenology are sufficient, and the connection between perceptual phenomenology and belief arguably direct enough for the feeling to directly give rise to the external attributions (Langland-Hassan, 2008).

causal link between belief in a mechanism and external agency attributions, though it should be noted that the measure of belief was taken at the *end* of the experimental protocol. The proposal is also consistent with anthropological research indicating that misattributions of agency for thought occur in religious contexts where a similar prior belief in a (supernatural) mechanism for agency disruption exists, though corresponding quantitative data is not yet available to indicate the prevalence of misattributions in these contexts (Luhrmann, 2005).

If, as the results suggest, contextual factors such as a prior belief in a plausible mechanism might be sufficient to generate misattributions of agency for thought in the general population, then it is possible that a similar causal account may also stand for at least some clinical cases. Such a proposal runs counter not only to many accounts from the literature on the sense of agency outlined above, which posit subpersonal disruptions to emotional (Synofzik, Vosgerau, & Newen, 2008a) or motor (Jones & Fernyhough, 2007; Synofzik, Vosgerau, & Newen, 2008a) processing at the heart of clinical symptoms, but also to accounts from the wider literature on delusional beliefs which hold that in order to fully account for the content of 'bizarre' beliefs like thought insertion, appeal must be made to a deeply unusual experience of thought in the first instance (Coltheart, Langdon, & Mckay, 2011).[8] The possibility that a higher-order account may suffice in clinical cases is supported by several additional strands of evidence. Many patient reports include an elaboration of the mechanism by which thought insertion/AVH occurs and it is possible that these beliefs may arise *prior* to the experience of external agency, rather than subsequently as is widely assumed. A belief in a mechanism for agentive disruption of thought, such as the "Air Loom" machine so vividly described by one 18th century patient (Jay, 2012) or the electrical device operated by a cabal of doctors described by another (Tausk & Feigenbaum, 1992), could arise as the result of the unusual reasoning style demonstrated by delusional patients (Fletcher & Frith, 2009; Garety & Hemsley, 1997) and as part of a broader web of interconnected delusional beliefs which characterize delusional ideation. Certainly there is evidence that modern technologies are readily incorporated into delusional complexes (Chakraborty, 1964). There is also evidence that such beliefs about mechanisms circulate within the pathological community online (e.g. online forums discussing FBI mind control devices (Bell, Maiden, Muñoz-Solomando, & Reddy, 2006)). Since patients also tend to refute evidence that runs counter to their prior beliefs (Woodward, Moritz, Menon, & Klinge, 2008), and show an exaggerated confirmation bias (Bentall & Young, 1996), a belief in a plausible mechanism for agentive disruption of thought could take on an important cognitive role, generating new beliefs in relation to the source of relatively ordinary individual thoughts. The possibility is also in line with a finding by Linney and Peters (2007) which demonstrated that underlying beliefs about the permeability of mind were different in those who had thought interference symptoms and those who did not, and with research suggesting that beliefs about mind may constrain the emergence of symptoms of thought insertion (Barrett, 2003). It is also consistent with the proposal that beliefs about the existence of types of causal forces (e.g. spirits) may drive the development of psychotic symptoms (Morrison, 2001).

4.2. Effect of emotional thought content

Participants in the negative condition made more frequent misattributions than those in the neutral condition and also rated the Mind-to-Mind Interface as more successful, indicating that negatively-valenced high-arousal thought affects the sense of agency for thought. While this finding demonstrates for the first time that thought content can affect attributions of agency, it leaves open the question of what cognitive mechanisms account for the effect.

Interestingly, target thought frequency (i.e. how many times participants thought the target thought overall, regardless of attribution) neither differed significantly between conditions nor correlated with frequency of misattributions, suggesting that the greater frequency of misattributions in the negative condition cannot be explained by a higher baseline occurrence of the

[8] 'Bizarre' delusional beliefs are distinguished by the fact that they could not be true under any circumstances (contrast with the 'ordinary' delusional belief that one is being followed by the FBI, which may run counter to available evidence, but could possibly be true).

target thought. Similarly, since target thought suppression neither differed across conditions nor correlated with misattribution frequency, suppression of negative target thoughts appears not to have acted as a confounding variable. Thus, it seems that the emotional response to the negative thought content was directly responsible for the higher frequency of agency misattributions in the negative condition. A direct causal role for negative thought content in misattributions of agency is consistent with the MWM model of the sense of agency, which holds that one important cue in the multifactoral weighting process giving rise to the sense of agency for thought is a feeling of the degree to which the thought "fits" based on whether or not it is negatively valenced (Synofzik, Vosgerau, & Newen, 2008a; 2009b). Broadly, this causal role for negative thought content is also consistent with the large body of research on the so-called self-serving bias, which suggests that negative action outcomes are likely to be casually attributed to another source (Mezulis, Abramson, & Hyde, 2004), though it should be noted that a wide variety of possible mechanisms have been proposed to account for these earlier findings, including motivational accounts which hold that negative outcomes are attributed to another source in order to maintain a positive view of the self (Shepperd, Malone, & Sweeny, 2008). The finding that negative thought content affected the frequency of misattributions of agency appears to go against Langland-Hassan's CM-based account, which explicitly rejects content as a causal factor and "predict[s] no special link between thought content and a lack of a sense of agency." (Langland-Hassan, 2008, p. 394).

While exploratory analysis revealed that the difference in frequency of misattributions between neutral and *positive* conditions was not significant, it is possible that the present study had insufficient power to detect an effect here. Certainly the direction of the difference is interesting, with positive thought being *more* misattributed than neutral. If future research should reveal this to be a consistent effect, it would suggest that the effect of emotional thought content is not the result of motivational factors related to maintenance of self image, and give support to cognitive accounts which posit a role for both positive and negative thought in generating misattributions. Stephens and Graham's (2003) top-down account of the sense of agency for thought centers on the role of thought content in generating misattributions, but rejects the idea that motivational factors play a role. Instead, the account holds that thoughts are misattributed because the subject finds them inexplicable in terms of their underlying intentional states and self-conception. It would certainly be consistent with such an account that strongly emotional thoughts might strike the subject as particularly "contextually unsuitable and personally uncharacteristic" and be more likely to be misattributed (Stephens & Graham, 2003, p. 173). Similarly, while the MWM account of the sense of agency stresses the role of negative thought in generating misattributions, later clarifications seem to allow the possibility that positive valence may also play a causal role; the feeling that a thought does not "fit" with the self is argued to arise from a lack of "emotional and content wise concordance with background beliefs, especially with self-image, and the general line of thoughts" (Synofzik, Vosgerau, & Newen, 2009b, p. 522) and such lack of emotional concordance could presumably arise in relation to strongly positive thought. If it transpires that positive thought is also more likely to be misattributed than neutral, this would be consistent with a leading account of AVH (recently extended to explain thought insertion (Linney & Peters, 2007)), which holds that cognitive dissonance caused by both strongly negative *and* positive thought is one factor leading patients to attribute the source of their inner speech to another person (Morrison, 1998; 2001; Morrison, Haddock, & Tarrier, 1995).

Results from previous studies looking at the effect of valence on attributions of *overt* speech using similarly varied single-word stimuli have been mixed, with some finding greater external attributions for both positive and negative words (Baker & Morrison, 1998; Ensum & Morrison, 2003; Morrison & Haddock, 1997), while others found the effect only for negative words (Johns, Gregg, Allen, & McGuire, 2006; Johns & McGuire, 1999; Johns, Rossell, Frith, & Ahmad, 2001). Interestingly, these experiments found that while word valence affected source attributions in a patient sample, no such effect was found for the non-patient control group (Allen, Freeman, Johns, & McGuire, 2006; Johns et al., 2001; Morrison & Haddock, 1997). The findings have been interpreted to suggest that a fundamental difference in emotional reaction to word content could

explain the occurrence of pathological misattributions, or at least their particular content. The present findings stand somewhat in contrast to these earlier results and interpretations, suggesting that, in the case of *inner* speech, the effect of emotional thought content on attributions is also found in non-patient samples. The discrepancy highlights the importance of using experimental protocols that directly measure agency attributions for thought rather than relying on findings from research on bodily action.

4.3. Descriptions of misattributed thoughts

Despite the known difficulty people have in parsing their experience of thought to provide accurate description of a given occurrent thought (Hurlburt & Heavey, 2004), participants expressed no difficulty in providing descriptions of the misattributed thoughts. Overall, participants' descriptions of the key characteristics of the thought or thinking process that led them to misattribute agency (or to contemplate it) were strikingly vivid and suggested quite unusual experiences of thought (e.g. "It was as if the thoughts were being planted in my conscious thoughts. It was clear to me they were not my own"). Many participants reiterated their powerful experiences, unprompted, during the debriefing interview. It should be noted that the present approach to analyzing participants' descriptions was exploratory; given the complexity of the phenomenology of thought it undoubtedly leaves some dimensions of the experience of the misattributed thought unexplored. Furthermore, it is possible that participant's responses may reflect culturally shared beliefs about thought (Lillard, 1998) and folk psychological distinctions about the nature of thought and thinking (Boyer, 2011) as much as first-person experience.

The most frequently cited characteristic was that the inserted thought content was *out of keeping* with current thought. While this characteristic was defined in terms of the relationship of thought content to the ongoing stream of thoughts (e.g. "they were not consistent with my train of thoughts at the time"), some individuals provided additional descriptive phenomenological information about the sudden arrival of the thought in the steam of consciousness (e.g. that it "popped", or was "out of the blue"). Interestingly, this very same identifying characteristic (inconsistent, 'out of the blue' thought) has been cited by individuals who believe that they have identified God's thoughts in their own mind during the course of prayer (Luhrmann, 2005). While quantitative data on the normal experience of thought is hard to come by (but see Hurlburt & Akhter, 2008) it is anecdotally clear that such sudden and unexpected changes in thought content are a relatively common part of mental life, suggesting that this ordinary feature of thought has gained special significance in the particular experimental setting.

The next most frequently cited characteristic was of strange or alien phenomenology accompanying the thought. Of these, three were specific, citing particular phenomenological qualities (e.g. "clearer and more direct"), while the remaining four gave a more general assessment ("just felt different", "did not feel like my own", "almost like alien thoughts"). Patient reports of thought insertion in schizophrenia often indicate a similarly hard-to-describe phenomenology of 'strangeness' or 'otherness', which has been taken by some as motivation for a subpersonal-deficit account (e.g. Langland-Hassan, 2008). The fact that such similar verbal reports were provided by the present non-patient sample suggests that such characteristics can occur in the absence of any deficit in subpersonal mechanisms.

Some participants indicated features related to the absence of intentional control of thought. In fact, in four of the responses coded positive for this characteristic the participant explicitly states that the thought occurred even though they were trying not to think it. Although correlational analysis did not reveal a relationship between levels of intentional suppression of the target thought (as measured by the post-trial measure of suppression) and frequency of agency misattributions, these open responses indicate that the inability to intentionally suppress the target thought was connected to misattributions of agency for some participants, at least in their post-hoc explanations.

Just three participants indicated that it was auditory characteristics that identified the thought, and in one of those cases the thought was misattributed because it was *quieter* than normal. In the present protocol great effort was taken to measure misattributions of agency and *not* hallucinatory experiences of 'hearing voices' (for which a variety of protocols exist e.g. van de Ven &

Merckelbach, 2003). The finding that auditory/voice-like dimensions did not play a major role in identifying external thought in the present sample suggests not only that this effort was successful, but also that the two can be separated at both the level of phenomenology and the level of verbal report.

Just three participants indicated that they had misattributed the thought because of some physical/perceptual accompaniment, and two of these reports related to electrical-type sensations in the head. This characteristic was likely to be a direct result of the helmet feature of the Mind-to-Mind interface, and suggests that general expectation can have a powerful effect on even perceptual experience. While not a primary interest of the present study, this effect is interesting in light of existing work with functioning TMS headpieces aiming to induce religious experiences (Persinger, 2001). Recent work has suggested that many reports of perceptual experiences in this context are the result of expectation related to the setup of the experiment and may be related to levels of suggestibility (Granqvist et al., 2005).

Only one response mentioned the emotional dimension of the thought content as a reason for misattribution. Since between-condition analysis revealed that there were significant differences in frequency of misattribution between the negative and neutral condition, this absence in the verbal reports indicated that individuals did not have insight into the effect of emotional content on their attributions. The idea that individuals may not be aware of the true reasons for their choices is well-established psychological phenomenon (Nisbett & Wilson, 1977). More specifically, the lack of insight into the role of emotional experience on attributions of agency is entirely in keeping with the particular characterization of the effect of emotional content on the sense of agency for thought under the MWM account, where the effect of emotional content occurs at a subpersonal level leaving the individual only with a 'strange feeling' (Synofzik, Vosgerau, & Newen, 2008a).

4.4. Limitations

Since the sample size in the present study was modest, the overall findings presented here should be viewed as preliminary. Furthermore, the sample consisted largely of undergraduate students and may not be representative of the wider population in Northern Ireland let alone populations from other cultures (Henrich, Heine, & Norenzayan, 2010).

One of the aims of the present study was to examine the role of emotional thought content on attributions of agency. The results show clear support for the hypothesis that those in the negative condition would make more frequent misattributions than those in the neutral condition, an effect discussed above in terms of the emotional dimension of valence. However, there are a number of issues that may bear on this interpretation. Because the study did not include a measure of mood we do not know if levels of positive and negative affect differed across conditions as a result of the presentation of the target thought. Furthermore, the effect of valence is complicated by the additional emotional dimension of arousal, which also differed across conditions. While extremes of valence and arousal are often combined, as here, to create highly emotional stimuli (e.g. Fox, Russo, Bowles, & Dutton, 2001), under the present design it is impossible to conclude whether the significant effects detected in frequency of misattributions were due to valence alone or an interaction between valence and arousal. In addition, valence itself may not have a consistent effect on the sense of agency, making extrapolation from the present findings to other settings difficult. For instance, the effect of valence may vary between individuals, with those with a depressive profile being more likely to *self*-attribute negative action outcomes (Mezulis et al., 2004). Moreover, the effect of negative valence on agency judgments may vary across *emotions*; thoughts which differentially evoke the emotions of *fear* and *shame* may both be of low valence, but given that one relates to external events and the other to internal events, it is possible that they may have quite different effects with respect to action perception and judgement (Berninger & Döring, 2012; Wilke et al., 2012). Some of these complexities relating to the effect of valence and arousal may help to explain the ambiguous results in terms of the effect of positive valence on the sense of agency, and should be taken into account in future research. Finally, while the experimental design allowed two key confounds to be ruled out (target thought frequency and suppression) it remains possible that features of the target thought words unrelated to their valence may have affected

frequency of misattribution. For instance, while all of the words can be used as verbs, only 'elbow' can also be used as a noun. It is at least plausible that the degree to which the word is interpreted as related to action could affect attributions of agency. Replication of the present results with a different set of words would help to reinforce the interpretation offered here in relation to the effect of negative thought content.

Another aim of the present study was to shed light on the viability of various cognitive etiological models of the sense of agency for thought. One of the complexities involved in discussing the present findings in relation to theoretical models relates to the precise notion of thought at stake in cognitive terms. The present study related to the mental occurrence of single words, and thus had some similarities to existing protocols examining inner speech, such as fMRI studies where participants are asked to mentally repeat single words (Shergill et al., 2001). It would be interesting to establish if the same results would obtain using more elaborated forms of inner speech. The theoretical landscape regarding notions of thought is incredibly complex, with little consensus on how to carve up mental activity at either the levels of phenomenology (Hurlburt & Schwitzgebel, 2007), or cognitive mechanisms (Carruthers, 2011). This complexity is reflected in cognitive accounts of the sense of agency for thought, some of which make a distinction between verbal and other types of thought (Synofzik, Vosgerau, & Newen, 2008a; Vosgerau & Newen, 2007), while others specifically relate only to inner speech (e.g. Jones & Fernyhough, 2007). Future empirical research will need to advance alongside conceptual developments to continue to disentangle these various notions of thought and probe whether the sense of agency varies across them. This is especially important as there is evidence that emotional response varies in intensity across modes of thought (Holmes & Mathews, 2010). The Mind-to-Mind protocol could easily be adapted to investigate attributions of agency for other types of thought—e.g. visual imagery—in order to provide empirical contributions to this debate.

It remains unclear what aspects of the experimental setup contributed to the unexpectedly high prevalence of external agency attributions in the present sample. We have discussed the possibility that a prior belief in a plausible mechanism may be largely sufficient to explain misattributions, and this possibility could be explored in future research by measuring or manipulating levels of belief in a plausible mechanism prior to measuring agency attributions. Moreover, there are likely to be other contextual factors that contributed to the high prevalence in the present sample, such as expectation of specific thought content and the act of quietly concentrating on thoughts. These would be particularly interesting to explore in future research given their relevance to religious contexts of worship and prayer.

Finally, the present study did not explore the impact of individual differences on levels of agency misattribution. The theoretical literature suggests several fruitful lines of inquiry here, such as a possible connection with tendency to absorption (Luhrmann, 2005), and connection to metacognitive beliefs (Morrison, 2001). It is also possible that the tendency to misattribute thoughts in the context of the Mind-to-Mind interface is related to biases in decision making in ambiguous contexts and certain reasoning styles such as the tendency to 'jump to conclusions', which are in turn associated with delusion-proneness (Fine, Gardner, Craigie, & Gold, 2007). In fact, both theoretical models of the sense of agency for thought and existing empirical research (Allen et al., 2006; M. Startup, Startup, & Sedgman, 2008; Sugimori, Asai, & Tanno, 2011a) suggest that there is likely to be a connection between levels of schizotypal personality traits and tendency to misattribute thoughts to an external source. Investigation of this possibility could contribute to the debate on the nature of the continuum of both pathological symptoms and underlying cognitive architecture in the general population (Freeman, Pugh, & Garety, 2008). It is also possible that the effect of emotional thought content on agency attributions might interact with schizotypy (Larøi, Linden, & Marczewski, 2004).

4.5. Concluding remarks

Given the novelty of the present experimental task and the notable paucity of previous experimental work on the sense of agency for mental action, the present results provide a valuable contribution to the literature on the sense of agency for thought, suggesting that prior belief in a

plausible mechanism may be largely sufficient to generate misattributions of agency and giving support to cognitive models which posit a role for contextual factors and prior belief in generating the sense of agency for thought. The findings also demonstrate that negative thought content increases the frequency of agency misattributions, and raise the possibility that this effect may extend to positive content, though more research is needed. There remain a myriad of open questions as to both the cognitive and phenomenological nature of the sense of agency for thought, and the Mind-to-Mind Interface protocol provides a powerful and flexible tool with which to empirically investigate these questions.

Acknowledgements

This work was supported by a grant to the first author from the European Office of Aerospace Research and Development (EOARD) and Queen's University, Belfast. The funding sources had no involvement in any aspect of the research. We would like to thank Nora Parren for her help in collecting data, Sarah Allen for allowing her photograph to be used, and two anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Aarts, H., Custers, R., & Wegner, D. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition*, *14*(3), 439–458.

Aleman, A., & Kahn, R. S. (2005). Strange feelings: do amygdala abnormalities dysregulate the emotional brain in schizophrenia? *Progress in Neurobiology*, *77*(5), 283–298. doi:10.1016/j.pneurobio.2005.11.005

Allen, P., Freeman, D., Johns, L., & McGuire, P. (2006). Misattribution of self-generated speech in relation to hallucinatory proneness and delusional ideation in healthy volunteers. *Schizophrenia Research*, *84*(2-3), 281–288.

Baker, C., & Morrison, A. (1998). Cognitive processes in auditory hallucinations: attributional biases and metacognition. *Psychological Medicine*, *28*(05), 1199–1208.

Balconi, M. (2010). *Neuropsychology of the Sense of Agency*. Springer.

Barrett, R. J. (2003). Kurt Schneider in Borneo: Do first rank symptoms apply to the Iban? In R. J. Barrett & J. H. Jenkins (Eds.), *Schizophrenia, Culture, and Subjectivity - the Edge of Experience* (pp. 87–109). Cambridge University Press.

Bayne, T., & Pacherie, E. (2007). Narrators and comparators: The architecture of agentive self-awareness. *Synthese*, *159*(3), 475–491.

Bell, V., Maiden, C., Muñoz-Solomando, A., & Reddy, V. (2006). "Mind control" experiences on the internet: implications for the psychiatric diagnosis of delusions. *Psychopathology*, *39*(2), 87–91. doi:10.1159/000090598

Bentall, R. P., & Young, H. F. (1996). Sensible hypothesis testing in deluded, depressed and normal subjects. *The British Journal of Psychiatry*, *168*(3), 372–375. doi:10.1192/bjp.168.3.372

Berninger, A. A., & Döring, S. S. (2012). Emotion and perception of one's own actions - A comment on Wilke, Synofzik and Lindner. *Consciousness and Cognition*, *21*(1), 46–47. doi:10.1016/j.concog.2011.08.008

Blakemore, S., Oakley, D., & Frith, C. (2003). Delusions of alien control in the normal brain. *Neuropsychologia*, *41*(8), 1058–1067.

Bortolotti, L., & Broome, M. (2009). A role for ownership and authorship in the analysis of thought insertion. *Phenomenology and the Cognitive Sciences*, *8*(2), 205–224. doi:10.1007/s11097-008-9109-z

Boyer, P. (2002). *Religion Explained*. Basic Books.

Boyer, P. (2011). Intuitive expectations and the detection of mental disorder: A cognitive background to folk-psychiatries. *Philosphical Psychology*, *24*(1), 95–118. doi:10.1080/09515089.2010.529049

Bradley, L., & Lang, P. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1. *The Center for Research in Psychophysiology, University of Florida.*, 1–49.

Carruthers, G. (2012). A metacognitive model of the sense of agency over thoughts. *Cognitive Neuropsychiatry*, *17*(4), 291–314. doi:10.1080/13546805.2011.627275

Carruthers, P. (2011). *The Opacity of Mind*. OUP Oxford.

Chakraborty, A. (1964). An analysis of paranoid symptomatology. *Transcultural Psychiatric Research*, 103–106.

Coltheart, M., Langdon, R., & Mckay, R. (2011). Delusional belief. *Annual Review of Psychology*, *62*, 271–298. doi:10.1146/annurev.psych.121208.131622

Cooper, J. M. (2007). *Cognitive Dissonance: 50 Years of a Classic Theory*. Sage Publications Ltd.

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, *17*(2), 523–534.

Desantis, A., Weiss, C., Schütz-Bosbach, S., & Waszak, F. (2012). Believing and perceiving: authorship belief modulates sensory attenuation. *PLoS ONE*, *7*(5), e37959. doi:10.1371/journal.pone.0037959

Dijksterhuis, A., Preston, J., Wegner, D., & Aarts, H. (2008). Effects of subliminal priming of self and God on self-attribution of authorship for events. *Journal of Experimental Social Psychology*, *44*(1), 2–9.

Ensum, I., & Morrison, A. (2003). The effects of focus of attention on attributional bias in patients experiencing auditory hallucinations. *Behaviour Research and Therapy*, *41*(8), 895–907.

Farrer, C., & Frith, C. D. (2002). Experiencing Oneself vs Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *NeuroImage*, *15*(3), 596–603. doi:10.1006/nimg.2001.1009

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *NeuroImage*, *18*(2), 324–333.

Fernández, J. (2010). Thought Insertion and Self-Knowledge. *Mind and Language*, 66–88.

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry*, *12*(1), 46–77. doi:10.1080/13546800600750597

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. doi:10.1038/nrn2536

Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology-General*, *130*(4), 681–700.

Freeman, D., Pugh, K., & Garety, P. (2008). Jumping to conclusions and paranoid ideation in the general population. *Schizophrenia Research*, *102*(1-3), 254–260.

Frith, C. D. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition*, *14*(4), 752–770.

Frith, C. D. (2012). Explaining delusions of control: The comparator model 20years on. *Consciousness and Cognition*, *21*(1), 52–54. doi:10.1016/j.concog.2011.06.010

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *355*(1404), 1771–1788. doi:10.1098/rstb.2000.0734

Gallagher, S. (2004). Agency, ownership and alien control in schizophrenia. In D. Zahavi & J. Parnas (Eds.), *The Structure and Development of Self-Consciousness* (pp. 89–104). John Benjamins Pub Co.

Gallagher, S. (2012). Multiple aspects in the sense of agency. *New Ideas in Psychology*, *30*(1), 15–31. doi:10.1016/j.newideapsych.2010.03.003

Garety, P. A., & Hemsley, D. R. (1997). *Delusions*. Psychology Pr.

Granqvist, P., Fredrikson, M., Unge, P., Hagenfeldt, A., Valind, S., Larhammar, D., & Larsson, M. (2005). Sensed presence and mystical experiences are predicted by suggestibility, not by the application of transcranial weak complex magnetic fields. *Neuroscience Letters*, *379*(1), 1–6. doi:10.1016/j.neulet.2004.10.057

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi:10.1017/S0140525X0999152X

Holmes, E. A., & Mathews, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, *30*(3), 349–362. doi:10.1016/j.cpr.2010.01.001

Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Current Psychiatry Reports*, *37*(20), 464–476. doi:10.1007/BF00622503

Hurlburt, R., & Akhter, S. (2008). Unsymbolized thinking is a clearly defined phenomenon: A reply to Persaud. *Consciousness and Cognition*, *17*(4), 1376–1377.

Hurlburt, R., & Heavey, C. (2004). To beep or not to beep: Obtaining accurate reports about awareness. *Journal of Consciousness Studies, 11*, *7*(8), 113–128.

Hurlburt, R., & Schwitzgebel, E. (2007). *Describing inner experience?* The MIT Press.

Jay, M. (2012). *Influencing Machine: James Tilly and the Air Loom*. Strange Attractor.

Johns, L. C., Gregg, L., Allen, P., & McGuire, P. (2006). Impaired verbal self-monitoring in psychosis: effects of state, trait and diagnosis. *Psychological Medicine*, *36*(04), 465–474.

Johns, L., & McGuire, P. (1999). Verbal self-monitoring and auditory hallucinations in schizophrenia. *The Lancet*, *353*, 469–470.

Johns, L., Rossell, S., Frith, C., & Ahmad, F. (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychological Medicine*, *31*, 705–715.

Jones, S., & Fernyhough, C. (2007). Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and Cognition*, *16*(2), 391–399.

Langland-Hassan, P. (2008). Fractured phenomenologies: thought insertion, inner speech, and the puzzle of extraneity. *Mind and Language*, *23*(4), 369–401.

Larøi, F., Linden, M., & Marczewski, P. (2004). The effects of emotional salience, cognitive effort and meta-cognitive beliefs on a reality monitoring task in hallucination-prone subjects. *British Journal of Clinical Psychology*, *43*(3), 221–233. doi:10.1348/0144665031752970

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, *123*, 3–32.

Linney, Y., & Peters, E. (2007). The psychological processes underlying symptoms of thought interference in psychosis. *Behaviour Research and Therapy*, *45*(11), 2726–2741.

Luhrmann, T. (2005). The art of hearing God: Absorption, dissociation, and contemporary American spirituality. *Spiritus*, *5*(2), 133–157.

Mathalon, D., & Ford, J. (2008). Corollary discharge dysfunction in schizophrenia: evidence for an elemental deficit. *Clinical Eeg and Neuroscience*, *39*(2), 82–86.

Mezulis, A., Abramson, L., & Hyde, J. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*(5), 711–747.

Moore, J. W., & Haggard, P. (2010). Intentional binding and higher order agency experience. *Consciousness and Cognition*, *19*(1), 490–491. doi:10.1016/j.concog.2009.11.007

Morrison, A. (1998). A cognitive analysis of the maintenance of auditory hallucinations: are voices to schizophrenia what bodily sensations are to panic? *Behavioural and Cognitive Psychotherapy*, *26*(04), 289–302.

Morrison, A. (2001). The interpretation of intrusions in psychosis: an integrative cognitive approach to hallucinations and delusions. *Behavioural and Cognitive Psychotherapy*, *29*(03), 257–276.

Morrison, A. P., Haddock, G., & Tarrier, N. (1995). Intrusive Thoughts and Auditory Hallucinations: A Cognitive Approach. *Behavioural and Cognitive Psychotherapy*, *23*(03), 265–280. doi:10.1017/S1352465800015873

Morrison, A., & Haddock, G. (1997). Cognitive factors in source monitoring and auditory hallucinations. *Psychological Medicine*, *27*(03), 669–679.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Paulik, G., Badcock, J. C., & Maybery, M. T. (2008). Dissociating the components of inhibitory control involved in predisposition to hallucinations. *Cognitive Neuropsychiatry*, *13*(1), 33–46. doi:10.1080/13546800701775683

Persinger, M. A. (2001). The Neuropsychiatry of Paranormal Experiences. *Journal of Neuropsychiatry*, *13*(4), 515–524. doi:10.1176/appi.neuropsych.13.4.515

Proust, J. (2006). Agency in schizophrenia from a control theory viewpoint. *The Disorders of Volition*, 87–118.

Rassin, E. (2005). *Thought suppression*. Elsevier Science.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi:10.1037/0278-7393.21.4.803

Seal, M., Aleman, A., & McGuire, P. (2004). Compelling imagery, unanticipated speech and deceptive memory: Neurocognitive models of auditory verbal hallucinations in schizophrenia. *Cognitive Neuropsychiatry*, *9*(1-2), 43–72. doi:10.1080/13546800344000156

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring Causes of the Self-serving Bias. *Social and Personality Psychology Compass*, *2*(2), 895–908. doi:10.1111/j.1751-9004.2008.00078.x

Shergill, S., Bullmore, E., Brammer, M., Williams, S., Murray, R., & McGuire, P. (2001). A functional study of auditory verbal imagery. *Psychological Medicine*, *31*(02), 241–253.

Sousa, P., & Swiney, L. (2011). Thought insertion: Abnormal sense of thought agency or thought endorsement? *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-011-9225-z

Startup, M., Startup, S., & Sedgman, A. (2008). Immediate source-monitoring, self-focused attention and the positive symptoms of schizophrenia. *Behaviour Research and Therapy*, *46*(10), 1176–1180.

Stephens, G. L., & Graham, G. (2003). *When Self-Consciousness Breaks*. The MIT Press.

Sugimori, E., Asai, T., & Tanno, Y. (2011a). Sense of agency over thought: external misattribution of thought in a memory task and proneness to auditory hallucination. *Consciousness and Cognition*, *20*(3), 688–695. doi:10.1016/j.concog.2010.12.014

Sugimori, E., Asai, T., & Tanno, Y. (2011b). Sense of agency over speech and proneness to auditory hallucinations: the reality-monitoring paradigm. *Quarterly Journal of Experimental Psychology (2006)*, *64*(1), 169–185. doi:10.1080/17470218.2010.489261

Synofzik, M., & Vosgerau, G. (2012). Weighting models and weighting factors. *Consciousness and Cognition*, *21*, 55–58.

Synofzik, M., & Voss, M. (2010). Disturbances of the Sense of Agency in Schizophrenia. In M. Balconi (Ed.), *Neuropsychology of the Sense of Agency* (pp. 145–154). Neuropsychology of the Sense of Agency.

Synofzik, M., Vosgerau, G., & Lindner, A. (2009a). Me or not me – An optimal integration of agency cues? *Consciousness and Cognition*, *18*(4), 1065–1068. doi:10.1016/j.concog.2009.07.007

Synofzik, M., Vosgerau, G., & Newen, A. (2008a). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239.

Synofzik, M., Vosgerau, G., & Newen, A. (2008b). I move, therefore I am: A new theoretical framework to investigate agency and ownership. *Consciousness and Cognition*, *17*(2), 411–424.

Synofzik, M., Vosgerau, G., & Newen, A. (2009b). Reply to Carruthers. *Consciousness and Cognition*, *18*, 521–523.

Tausk, V., & Feigenbaum, D. (1992). On the origin of the "influencing machine" in schizophrenia. *The Journal of Psychotherapy Practice and Research*, *1*(2), 184–206.

van de Ven, V., & Merckelbach, H. (2003). The role of schizotypy, mental imagery, and fantasy proneness in hallucinatory reports of undergraduate students. *Personality and Individual Differences*, *35*(4), 889–896.

van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., & Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. *Psychological Medicine*, *39*(02), 179–179. doi:10.1017/S0033291708003814

Vosgerau, G., & Newen, A. (2007). Thoughts, motor actions, and the self. *Mind and Language*, *22*(1), 22–43.

Wegner, D., Sparrow, B., & Winerman, L. (2004). Vicarious Agency: Experiencing Control Over the Movements of Others. *Journal of Personality and Social Psychology*, *86*(6), 838–848.

Wilke, C., Synofzik, M., & Lindner, A. (2012). The valence of action outcomes modulates the perception of one's actions. *Consciousness and Cognition*, *21*(1), 18–29. doi:10.1016/j.concog.2011.06.004

Woodward, T. S., Moritz, S., Menon, M., & Klinge, R. (2008). Belief inflexibility in schizophrenia. *Cognitive Neuropsychiatry*, *13*(3), 267–277. doi:10.1080/13546800802099033