Title
Neural Mechanisms of Person-Specific Theory of Mind

Author
Welborn, Benjamin Locke

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Neural Mechanisms of Person-Specific Theory of Mind

A dissertation submitted in satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Benjamin Locke Welborn

2015

ABSTRACT OF THE DISSERTATION

Neural Mechanisms of Person-Specific Theory of Mind

by

Benjamin Locke Welborn

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2015

Professor Matthew D. Lieberman, Chair

Human beings are uniquely equipped with the capacity to reason in sophisticated ways about others' thoughts and feelings, an ability often referred to as a 'theory of mind' (Premack & Woodruff, 1978; Gopnik & Wellman, 1994; Leslie et al., 2004). Recent literature on theory of mind has expanded considerably, and contemporary theoretical work has sought to differentiate ways in which different component psychological processes may contribute to mental state reasoning (Schaafsma et al., 2015; Van Overwalle & Vandekerckhove, 2013). Several prominent accounts have sought to explain an empirical dissociation between mentalizing-related activity in the dorsal medial prefrontal cortex (DMPFC) and the ventral medial prefrontal cortex (VMPFC): the former region is characteristically implicated in reasoning about unknown others while the latter is consistently evoked with participants reason about the self and close others (see meta-analysis by Van Overwalle & Baetens, 2009). A 'similarity' account sees the activity of ventral MPFC as related to its role in simulating the minds of others by reference to our own thoughts

and feelings (Mitchell, Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006). According to this viewpoint, we can employ the same neuro-cognitive resources and mechanisms when thinking about others as when introspecting about our own mental states, provided that others are seen to be sufficiently similar to the self. A second theoretical view contends that VMPFC is responsible for encoding and reasoning about the minds of close others, who are members of our family, tribe, or another meaningful social group (Krienen, Tu, & Buckner, 2010).

In the present dissertation work a third alternative is proposed, according to which person-specific *knowledge* of others' unique personal characteristics allows for nuanced and individuated inferences about their transient mental states and enduring dispositional traits. We theorize that rich experiential knowledge of particular individuals can be leveraged into person-specific theories of mind, instantiated in the MPFC, which facilitate mental state reasoning that is tailored to the minds of those persons we know especially well. In Study 1, we provide neuroimaging (fMRI) results in favor of this 'person-specific theory-of-mind hypothesis', showing that ventral VMPFC is more responsive to well-known targets than to poorly-known targets, regardless of perceived similarity and affective closeness. Moreover, activity in medial prefrontal cortex is associated with trial-by-trial variation in the availability of person-specific knowledge, and with participants' willingness to attribute idiosyncratic traits to social targets. In a second fMRI study (Study 2), we explore the formation of person-specific theories of mind regarding initially unfamiliar targets, revealing a comparable pattern of hemodynamic activity.

This dissertation of Benjamin Locke Welborn is approved.

Naomi I. Eisenberger

Martin M. Monti

Marco Iacoboni

Matthew D. Lieberman, Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge the steadfast support of my advisor and mentor, Matthew Lieberman, as well as the thoughtful and helpful contributions of committee members Naomi Eisenberger, Martin Monti, and Marco Iacoboni. I would also like to express my gratitude for intellectual and moral support of current and former colleauges in the Social Cognitive Neuroscience and Social and Affective Neuroscience laboratories at UCLA, especially Meghan Meyer, Jared Torre, Stephanie Vezich, Ben Gunter, Tristen Inagaki, Keely Muscatell, Kate Haltom, Elliot Berkman, Emily Falk, and Robert Spunt. I would like to thank Eva Telzer, Jesse Rissman for their intellectual advice and encouragment.

I would like to acknowledge the generous fellowship funding received through the National Science Foundation Graduate Rearch Fellowship Program. I would also like to thank MIT press for permission to reproduce Paper 1 of this dissertation, which is published in the January 2015 issue of the *Journal of Cognitive Neuroscience* (Volume 27, pp. 1-12) and is © 2014 Massachusetts Institute of Technology.

Lastly I would like to thank Larry Welborn, Diane Welborn, Mark Welborn, and Natalia Konstantinovskaia for their constant inspiration and encouragement during my graduate studies and the pursuit of this dissertation research. Without them it would not have been possible.

Vita

Education:
2012 Candidate in Philosophy in Psychology, University of California – Los Angeles
2010 Master of Arts in Psychology, University of California – Los Angeles
2008 Bachelor of Arts in Cognitive Science, Magna Cum Laude, Yale University

Fellowships and Honors:
2015 Sage Center Junior Research Fellowship, UCSB
2013 UCLA Neuroimaging Training Program (NITP)
2012 National Science Foundation, Graduate Research Fellowship
2011 University of Michigan Training Course in fMRI
2010 UCLA Graduate Research Mentorship
2010 Graduate Summer Research Mentorship
2008 UCLA Chancellor's Prize/Stipend
2008 UCLA University Fellowship

Publications:

Welborn, B.L., Lieberman, M.D., Goldenberg, D., Fuligni, A.J., Galvan, A., and Telzer, E.H. (under review). Neural mechanisms of social influence in adolescents.

Welborn, B.L, Gunter, B.C., Vezich, I.S., and Lieberman, M.D. (under review). Neural correlates of the false consensus effect (FCE): Evidence for motivated projection and regulatory restraint.

Welborn, B.L., and Lieberman, M.D. (2015) Person-specific Theories of Mind in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience, 27,* 1-12.

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., & Lieberman, M. D. (2013). Creating buzz: The neural correlates of effective message propagation. *Psychological Science, 24,* 1234-1242.

Poore, J. C., Pfeifer, J. H., Berkman, E. T., Inagaki, T. K., Welborn, B. L., Lieberman, M. D. (2012). Prediction-error in the context of real social relationships modulates reward system activity. *Frontiers in Human Neuroscience, 6*, 218.

Welborn, B.L., Papademetris, X., Reis, D.L., Rajeevan, N., Bloise, S.M., Gray, J.R., (2009). Variation in orbitofrontal cortex volume: Relation to sex, emotion regulation, and affect. *SCAN*, *4*, 328-339.

General Introduction:

The human ability to represent, interpret, and predict the mental states of others is fundamental to social reasoning, and may be one of the unique capacities that distinguishes humans as a species (Premack & Woodruff, 1978). As a result, theory of mind has been a critical subject of investigation for social cognitive neuroscience, and has generated a considerable body of literature (Lieberman, 2010). Early research in this area identified a set of core brain regions consistently involved in human mentalizing using PET and fMRI (see Frith & Frith, 1999; Frith & Frith, 2003), including the medial prefrontal cortex (MPFC), precuneus/posterior cingulate, temporo-parietal junction (TPJ), and temporal poles (TP). However, out of the considerable literature on the neural basis of social reasoning, a relatively consistent empirical dissociation has emerged between the functional profiles of dorsal and ventral medial prefrontal cortex (roughly corresponding to Brodmann's areas 8/9 and 10, respectively; hereafter referred to as DMPFC and MPFC). Lieberman (2010) notes that while DMPFC activity is consistently associated with mentalizing tasks and only occasionally associated with self-reflection, MPFC demonstrates the opposite pattern of activity.

A number of theoretical proposals have been offered to explain the functional differences between neurocognitive processes carried out by DMPFC and MPFC, and each has accrued some measure of experimental support. The 'similarity' perspective holds that we reason about the minds of similar others in much the same way as we think about the self, and argues that VMPFC underlies both kinds of mentalizing (Mitchell, Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006). This research draws upon ideas in simulation theory (Harris 1992), which asserts that our own mental apparatus equips us with a ready-made model for understanding the mental states of others. In mentalizing, we frequently generate 'simulations' of

other minds, temporarily adopting what we perceive to be their beliefs, desires, and values and employing only information to which they would have access. We can then use the feelings and thoughts evoked by the simulation as a rough heuristic to aid the mentalizing process, on the assumption that others would respond approximately as we do, given similar circumstances. More recently, Krienen, Tu, & Buckner (2010) have proposed an integrative account of MPFC function centered around the idea that evolution has endowed us with unique processes and strategies for thinking about socially meaningful others. According to Krienen and colleagues, the MPFC is especially sensitive to those individuals who are close to us, whether through ties of kinship, friendship, or common group membership (e.g. clan or tribe).

The present dissertation research investigated a novel hypothesis regarding the function of MPFC in mentalizing: that it supports the implementation of *person-specific* theories of mind, extending the social thinker's abilities to make nuanced inferences about the thoughts, feelings, and actions of others by taking into account their unique character traits and personal idiosyncrasies. In the first study, the neural bases of person-specific theories of mind were explored by contrasting well-known to poorly-known social targets, while controlling for affective factors such as perceived similarity and felt closeness that often go hand-in-hand with rich, detailed knowledge of others people. The second study focused on the development and deployment of person-specific theories of mind about *novel* social targets, and assessed their effects on activity in regions associated with mentalizing, including especially the medial prefrontal cortex.

PAPER 1:

Person-specific Theory of Mind in Medial pFC

(Published in the *Journal of Cognitive Neuroscience*)

Abstract

While research on Theory of Mind has strongly implicated the dorsomedial prefrontal cortex (DMPFC, incuding medial BA8 and BA9), the unique contributions of medial prefrontal cortex (MPFC, corresponding to medial BA10) to mentalizing remain uncertain. The extant literature has considered the possibility that these regions may be specialized for self-related cognition or for reasoning about close others, but evidence for neither theory has been conclusive. We propose a novel theoretical framework: MPFC selectively implements 'person-specific theories of mind' (ToM_p) representing the unique, idiosyncratic traits or attributes of well-known individuals. To test this hypothesis, we used fMRI to assess MPFC response in Democratic and Republican participants as they evaluated more or less subjectively well-known political figures. Consistent with the ToM_p account, MPFC showed greater activity to subjectively well-known targets, irrespective of participants' reported feelings of closeness or similarity. MPFC also demonstrated greater activity on trials in which targets (whether politicians or oneself) were judged to be relatively idiosyncratic, making a generic theory of mind (ToM_g) inapplicable. These results suggest that MPFC may supplement the generic theory of mind process (ToM_g) with which DMPFC has been associated, by adding on mentalizing capacities tuned to individuated representations of specific well-known others.

Introduction

The capacity to make sense of the mental states and traits of others is an extraordinary ability that allows us to plot a course through the complexities of social life. Countless studies have identified regions of the medial prefrontal cortex involved in thinking about the mental states of others and of ourselves (Amodio and Frith, 2005). Within this region, an asymmetry has been observed (Lieberman, 2010; Van Overwalle, 2009) such that dorsomedial prefrontal cortex (DMPFC, Brodmann areas (BA) 8/9) is more often found in studies of mentalizing (i.e. thinking about the mental states and traits of others), whereas ventromedial prefrontal cortex (MPFC, BA 10) is more often found in studies of self-reflection (i.e. thinking about one's own states and traits). A number of recent investigations have attempted to clarify the ways in which MPFC might also contribute to mentalizing.

Early neuroimaging studies of trait self-knowledge, a form of self-reflection, typically included famous targets as controls (e.g. George Bush) and more often than not, the comparison of self to famous targets yielded activity in MPFC (Kelley et al., 2002). A few studies also included a well-known close other as a control and unlike the famous targets, the close others often produced MPFC activity similar to that of self-reference (Ochsner et al., 2005; Vanderwal et al., 2008). Research from Mitchell and colleagues (Mitchell, Banaji, and Macrae, 2005) provides a potential account of these results in terms of the perceived similarity of close others. If close others are perceived to be similar to the self, then reflecting on one's own reactions to a query and projecting this on to the other person would be an efficient strategy for estimating the other's reactions.

Consistent with this hypothesis, activity in MPFC during mentalizing judgments of a social target has been shown to vary parametrically with perceived similarity to the self

(Mitchell, Banaji, and Macrae, 2005). Social targets with a political orientation similar to the self (liberal/conservative) also elicit greater MPFC response during judgments of preferences than do politically dissimilar others (Mitchell, Macrae and Banaji, 2006). By this account, the essential function of MPFC is self-knowledge, but this self-knowledge can be used strategically to make inferences about similar others.

A second account suggests that the closeness of others to oneself is the key factor driving MPFC activity. More specifically, Krienen and colleagues (Krienen, Tu, and Buckner, 2010) have suggested that MPFC is primarily sensitive to the social relevance to or social distance from oneself, signaling friendship and kinship affiliation rather than abstract similarity. In their research MPFC was consistently more responsive to real-world friends than to unknown strangers, even when those strangers were judged to be more similar to the self than comparable friends. Thus, similarity and kinship accounts of MPFC contributions to mentalizing have both garnered a fair amount of empirical support.

In the current research, we propose a novel characterization of MPFC's contribution to mentalizing that would simultaneously account for existing data on self-reflection, similarity, and kinship. DMPFC has often been characterized as supporting our Theory of Mind (Wimmer and Perner, 1983), a generic model of how minds react to various situations and experiences. This *generic theory of mind* (ToM_g) can be applied to anyone in any real or imagined context -- allowing us, for example, to confidently predict how a typical adult male with a gun to his head would respond to a request to express his undying love of Justin Bieber's music (compared to when the gun is absent from the same scene). Nearly all neuroimaging studies of Theory of Mind have focused on strangers or imaginary characters for which a ToM_g is sensible to apply.

However, in our daily life we interact with friends, family, and co-workers repeatedly and often learn that their distinctive personalities mean that our ToM_g does not always apply. Instead, we may generate *person-specific theories of minds* (ToM_p) that are tailored to particular individuals. Whereas there is a considerable body of evidence showing that DMPFC supports generic mentalizing (ToM_g), we hypothesize that MPFC supports ToM_p and aimed to test this notion empirically. More specifically, we predicted that MPFC would be more active for social targets about whom we have extensive knowledge, particularly when this knowledge is both idiosyncratic to that target and relevant to a judgment to be made.

The ToM_p hypothesis can account for results associated with the theoretical perspectives discussed above. Close relationships with well-known others naturally furnish us with a diverse array of interpersonal experiences, from which we can generate a unique, person-specific theory of mind (ToM_p). In the case of similar others, we may draw upon an especially rich person-specific theory, that of the self. Thus, a person-specific ToM account of MPFC function is congruent with the findings associated with the similarity and closeness approaches. This approach also suggests that representations of the self are not qualitatively distinct from other person-specific representations, but rather are the most well-developed exemplars of the kinds of social representations handled by MPFC more broadly.

The ToM_p hypothesis can also be distinguished from these other approaches on empirical grounds. Only the ToM_p hypothesis predicts that MPFC will be recruited when thinking about individuals about whom we have a great deal of idiosyncratic knowledge, but who are also *disliked* and perceived as *dissimilar* from us. To test this hypothesis, we asked individuals with a strong political affiliation (Democrat/Republican) to make trait judgments about four political figures, two in the participant's own political party (*Own Party*) and two in the opposing party

(*Opposition Party*). We assumed that the political targets in the participant's own party would be seen as both more similar and closer to oneself than the political targets in the opposing party. Critically, we also manipulated the political targets used such that the amount of prior knowledge about them varied (see Figure 1). Within each political party, participants nominated one political figure about whom they knew a considerable amount (high knowledge targets) and one political figure about whom they knew relatively little (low knowledge targets).

We hypothesized that regardless of political affiliation, participants would generate greater MPFC activity to well-known political targets, due to the recruitment of a ToM_p, than to the less well-known political targets, for which only the ToM_g would be available. We also hypothesized that for each trait judgment, the extent to which the political target is judged to be distinctive from both the self and the typical person reflects the likelihood that a ToM_p is being applied to a particular trait, because neither projection from the self or a ToM_g would be applicable. Thus, MPFC should be more active during judgments of more *idiosyncratic* traits for a given target. Finally, we also hypothesized that MPFC during the retrieval of self-knowledge reflects the use of one ToM_p among several, rather than a "self" mechanism, per se. Consequently, we predicted that MPFC would be more active during self-judgments on traits for which the self is judged to be idiosyncratic (i.e. distinct from the typical person).

Methods:

Participants

Sixteen participants (8 female) were recruited by a combination of email solicitations and in-person presentations at undergraduate Democratic and Republican clubs at UCLA. All participants:

a) indicated strong affiliation with either the Democratic or Republican party by selecting either 0-2 or 8-10 respectively on an 11-point Likert scale anchored on either end at either "Strongly Republican" and "Strongly Democratic",
b) indicated that they considered themselves to be at least moderately knowledgeable regarding current American politics (>5 on an 9-point Likert scale anchored at "1 - not at all knowledgeable", "5 – moderately knowledgeable", and "9 - extremely knowledgeable").
c) met target-selection criteria detailed below.

Participants were judged ineligible for participation if they did not meet the above criteria. In addition, participants were ineligible if they were left-handed, using psychoactive medications or drugs, had been diagnosed with a neurological or psychiatric disorder, were pregnant, had a history of claustrophobia, or presented any other condition that would render participation in fMRI research hazardous.

Participants were all young adults between 18 to 29 years of age (M=22.1, SD=3.2). All participants were compensated $40 for their contribution to this research. Participants provided written informed consent approved by the UCLA Institutional Review Board. One participant's data are not included in these analyses due to partial data acquisition failure.

Target Selection

Four political figures (one high- and one low-knowledge target from each party) were selected ideographically for each participant based upon a screening questionnaire that queried participants about each of 50 contemporary politicians. All potential targets included in the screening questionnaire were active political figures who presently or formerly (<5 years prior) served in one more of the following offices: President of the United States, Senator,

congressional Representative, or state Governor. No targets were included that might be well-known for reasons unrelated to politics and governance (e.g. Arnold Schwarzenegger, Jesse Ventura, or Al Franken), or who had been associated with highly-publicized scandals or controversies (e.g. Bill Clinton or Larry Craig).

For each screening target, participants indicated their degree of knowledge and liking on 9-point Likert scales and reported the target's party affiliation. Responses on both scales were used to select 4 targets for the scanning session (one high-knowledge and one low-knowledge target from each party), subject to the following conditions:

1) Each participant's high-knowledge targets were rated between "Very knowledgeable" and "Extremely knowledgeable" (knowledge scale values 7-9)
2) Each participant's low-knowledge targets were rated between "Slightly knowledgeable" and "Moderately knowledgeable" (knowledge scale values of 2-5), and at least 4 points lower than both high-knowledge targets.
3) Each participant's own-party candidates were liked (liking scale values >5). Opposition-party candidates were disliked (liking scale values <5).
4) All targets' party affiliations were correctly identified.

These criteria accomplished the following objectives: a) ensure appropriate like/dislike attitudes toward Own party and Opposite party targets, b) provide targets with desired variation in knowledge for the political targets, and c) accommodate individual differences in the use of screening scales (see Figure 1 for example targets for a hypothetical Democratic participant). Participants were excluded if appropriate targets could not be identified.

This individualized target selection procedure subjected each participant to a 2 (target politician knowledge, High vs. Low) x 2 (target politician party affiliation, Own Party vs. Opposition Party) within-subjects factorial design, yielding four cells (see Figure 1). As detailed below, participants judged the applicability of trait words to each political figure while undergoing fMRI. Planned comparisons between hemodynamic activity associated with assessments of target politicians in different experimental conditions (e.g. High Knowledge > Low Knowledge, Own Party > Opposition Party) allowed for a direct test of our primary hypothesis regarding the function of MPFC.

Behavioral Measures:

Given that self-reports of knowledge assessments may be biased and/or self-serving, participants were asked (after scanning) to write essays demonstrating their political knowledge of each target. Ratings of these essays by independent evaluators blind to experimental hypotheses were used as an alternative (unbiased) measure of target knowledge.

In their essays, participants were instructed to describe the target's activities in politics and government, stances on contemporary political issues, and important accomplishments. These essays were rated by trained evaluators (research assistants) who were blind to the participant's party affiliation and the experimental condition of the target political figure. Evaluators were instructed to judge how knowledgeable the participant was about the target in the domain of politics and government, using a 7-point Likert scale anchored at "not at all knowledgeable" and "extremely knowledgeable". Essay-rated knowledge scores reflecting raters' independent judgments of participants' target-specific knowledge were thus available for all four political figure targets. Evaluators did not possess any special political expertise, and

were not affiliated with either political party. Parametric modulation of activity in MPFC by essay-rated target knowledge is employed as a complementary test of the ToM_p hypothesis. Lastly, participants completed self-report measures indicating their degree of overall similarity, closeness, and connectedness to each of the four political targets selected for the scanner task (range 0 to 10 inclusive). Participants also completed a measure of perceived personal overlap with each of the political targets (range 1 to 7, inclusive).

fMRI Paradigms

While undergoing fMRI, participants completed a trait-judgment task ("Politician Judgment Task") in which they rated the applicability of 30 personality traits to each of the four target political figures over three functional runs. In two separate functional runs, participants rated the applicability of these 30 traits to the self as well as to the ordinary American ("Self/ordinary American Judgment Task"). A case-judgment control task (uppercase/lowercase) using the same trait words and phrases was included in all runs as an experimental control.

Trait words and phrases were selected for the above tasks from a larger list on the basis of pilot testing within the UCLA undergraduate population, ensuring that items were comprehensible and meaningful to participants. All trait items were relevant to the domain of politics and governance (e.g. "patriotic", "able to take command", "opportunistic"). A mixed block/event-related design was employed in order to best explore hypotheses concerning both target-level and trial-level effects, with trial events grouped into super-ordinate blocks based on target identity. During each trait-judgment trial, participants used an on-screen 9-point Likert-type scale to indicate the applicability of the specified trait to the current target.

Trait-judgments of political figures were spread over three functional runs, with each run containing 8 experimental blocks and 2 control blocks. Block order within and between runs was

randomized for each participant, with the following constraints: 1) the same target was never selected for two consecutive blocks, 2) each political figure appeared at least once in each run, 3) no political figure appeared more than twice in any single run, and 4) no more than two blocks featuring targets from the same party were presented consecutively.

Trait-judgments regarding the Self and the ordinary American were spread over two functional runs, with each run containing 6 experimental blocks (3 Self and 3 ordinary American) and 2 control blocks. Block order for Self/ordinary American judgments was similarly randomized, with no two blocks of the same type occurring consecutively. In total, participants completed 6 blocks of trait-judgment for each political figure, for the Self, and for the ordinary American, and 10 control blocks.

Each block consisted of a 2 second introduction specifying the target, followed by 5 trait-judgment trials, and concluding with a 5 second rest period between blocks. During each trait-judgment trial, participants used an on-screen 9-point Likert-type scale to indicate the applicability of the specified trait to the current target. Trait words and phrases appeared on-screen for 5 seconds, during which participants moved the scale to the appropriate response value and confirmed their selection. Scale-movement and stimulus duration were determined on the basis of pilot testing, such that participants could comfortably make their judgments, move the on-screen scale, and confirm their responses in the allotted time. Trait words and phrases, as well as the scale selection indicator were removed from the screen following response selection. Each trial was followed by a jittered inter-stimulus interval drawn from an exponential random distribution with a mean of 2 seconds. Stimulus presentation was identical for Self/ordinary American runs, except that the targets of judgment differed. During the control task, the same 30 trait words were presented in either all upper-case or all lower-case letters, and participants were

required to make a binary case judgment. This task was designed to control for basic perceptual and motor processing associated with the use of the on-screen scale, as well as spontaneous lexico-semantic processing unrelated to the trait-judgments themselves.

fMRI Data Acquisition

All imaging data was acquired using a 3.0-Tesla Siemens Trio scanner at the Ahmanson-Lovelace Brain Mapping Center at UCLA. Across 5 functional runs, 761 T2*-weighted echo-planar images were acquired during completion of experimental tasks described above (slice thickness=3 mm, gap=1 mm, 36 slices, TR=2000 ms, TE=25 ms, flip angle=90°, matrix=64 x 64, field of view=200 mm). An oblique slice angle was used in order to minimize signal drop-out in ventral medial portions of the brain. In addition, a T2-weighted, matched-bandwith anatomical scan was acquired for each participant (TR=5000 ms, TE=34 ms, flip angle=90°, matrix=128 x128; otherwise identical to EPIs). Lastly, we acquired a T1-weighted magnetically-prepared rapid acquisition gradient echo anatomical image (slice thickness=1 mm, 176 slices, TR=2530 ms, TE=3.31 ms, flip angle=7°, matrix=256 x 256, field of view=256 mm).

fMRI Data Preprocessing and Analysis

Preprocessing and Region of Interest (ROI) definition

Functional data were analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Within each functional run, image volumes were corrected for slice acquisition timing, realigned to correct for head motion, segmented by tissue type, and normalized into standard MNI stereotactic space (resampled at 3 x 3 x 3 mm) using a diffeomorphic registration algorithm (DARTEL; Ashburner, 2007). Finally, images were smoothed with an 8 mm Gaussian kernel, FWHM.

Given our specific hypotheses regarding the role of MPFC in deploying person-specific

theories of mind, all principal analyses were conducted on a predefined MPFC region of interest. This ROI was constructed using the Automated Anatomical Labeling (AAL) toolbox (Tzourio-Mazoyer et al., 2002) of the Wakeforest University Pickatlas (Maldjian, Laurienti, Kraft, and Burdette, 2003), encompassing Brodmann's area 10. This base region was dilated and constrained to the medial aspect ($-15<x<15$). All ROI analyses reported below interrogate only voxels within this region.

Two analytic strategies were employed to assess the role of MPFC in person-specific and theory of mind. First, we used selective averaging for hypothesis testing on the ROI considered as a whole functional unit. Given the relatively large volume of the MPFC, we also use an alternative exploratory approach, searching within the same ROI for significant clusters and correcting for multiple comparisons within that search-space. Monte Carlo simulations implemented in 3dClustSim (from AFNI; Cox et al., 1996) were used to determine appropriate cluster-size thresholds given the smoothness of the images (16 contiguous voxels) to ensure overall false discovery rate (FDR) of less than 0.05, when combined with a voxel-wise significance threshold of $p<0.005$. All results reported exceed these joint voxel-wise and cluster-extent thresholds.

Target-level Analyses

For both the Politician Judgment Task and the Self/ordinary American Judgment Task we defined a GLM for each participant. Blocks were modeled as variable epochs spanning the duration between the onset of the first trial and the offset of the final trial, convolved with the canonical (double-gamma) HR. For the Politician Judgment Task, five regressors of interest were modeled (High Knowledge, Own Party; High Knowledge, Opposition Party; Low Knowledge, Own Party; Low Knowledge, Opposition Party; Case-Judgment Control). For the Self/ordinary

American Judgment Task, three regressors of interest were modeled (Self; ordinary American; Case-Judgment Control). All analyses controlled for 18 motion parameters (3 translations and rotations, as well as their squares and first-order derivatives). The time series was high-pass filtered using a cutoff period of 128s and serial autocorrelations were modeled as an AR(1) process. Contrast images were averaged across runs for each participant, and entered into a mixed effects analysis at the group level. Parameter estimates were extracted from the MPFC ROI using MarsBaR (Brett, Anton, Valabregue, and Poline, 2002) and subjected to a repeated-measures ANOVA. In the results presented below and the accompanying figures, parameter estimates reflect contrasts between the appropriate experimental condition and the case-judgment control task, unless otherwise noted.

Target-level factors (overall similarity, closeness, connectedness, personal overlap, scored knowledge, mean similarity, mean idiosyncrasy, and mean positivity) not subject to experimental manipulation were analyzed as parametric modulators of activity in the MPFC ROI. The effects of each factor (when appropriate, controlling for other factors, see below) were assessed based upon the appropriate parameter estimates from a GLM, identical to the above except for the different regressors of interest (that is, regressors reflected continuous scores on target-level variables, rather than target identity).

Trial-by-trial Trait-level Analyses

The effects of trial-to-trial trait-level perceived similarity, idiosyncrasy, and positivity were assessed using analyses of parametric modulation of the hemodynamic response to each of three trial-level predictors (similarity, idiosyncrasy, and positivity) implemented using an event-related general linear model. Similarity was operationalized as the absolute value of the difference between self-judgments and political target judgments on a given trait, with small

values indicating high similarity. Idiosyncrasy was computed as the absolute value of the difference between judgments of the political target and judgments of the ordinary American, with large values indicating high idiosyncrasy. Judgment positivity simply reflected the participant's rating of a given political target on a specific trait (reverse-coded for trait words or phrases judged by a pilot sample to be negative). For self-judgments, we also computed an index analogous to target idiosyncrasy, reflecting the distinctiveness of the self on a given trait, relative to the ordinary American. This self-idiosyncrasy was operationalized as the absolute value of the difference between judgments of the self and judgments of the ordinary American on a particular trait, with large values indicating high self-idiosyncrasy. The GLMs used to assess the effects of trait-level factors differed from those used to assess target-level factors in that specific trials (rather than the super-ordinate block) were modeled as discrete events, using a variable epoch spanning the duration from trial onset to response.

Results:

Behavioral results

Eight participants identified as strong members of the Democratic Party, while 7 identified as strong members of the Republican Party. See Table 1 for a summary of descriptive data discussed below.

Target differences in Self-reported Knowledge, Closeness, Connectedness, Personal Overlap, and Overall Similarity:

Participants indicated that they possessed greater knowledge of High-Knowledge than Low-Knowledge targets (M_{high}=7.90, M_{low}=3.43, t=19.31, p<0.001). This result is direct consequence of the target selection procedures employed (see *Methods* above for details). Self-reported knowledge did not differ significantly between Own Party and Opposition Party targets

(M_{Own}=5.77, M_{Opp}=5.57, t=1.71, *ns*).

Participants judged themselves to be more similar (M_{Own}= 5.44, M_{Opp}=0.66, t=16.51, p<0.001), closer (M_{Own}= 4.61, M_{Opp}=0.14, t=10.67, p<0.001) and more connected (M_{Own}= 5.28, M_{Opp}=0.25, t=17.09, p<0.001) to Own Party than to Opposition Party targets. In addition, participants perceived greater personal overlap with Own Party relative to Opposition Party targets (M_{Own}=5.00, M_{Opp}=1.50, t=14.34, p<0.001). Participants did not see themselves as more similar to High-Knowledge than to Low-Knowledge targets, and did not perceive greater overlap (all *p*s *ns*). However, participants did feel closer (M_{high}=3.00, M_{low}=1.75, t=5.49, p<0.001) and more connected (M_{high}=3.32, M_{low}=2.21, t=3.60, p=0.003) to high-knowledge targets than to low-knowledge targets. For this reason, relevant target-level analyses reported below control for closeness and connectedness to target.

Target Differences in Knowledge, assessed by Individual Essays:

Participants' essays regarding each political figure were coded by three research assistants who were blind to the target's experimental condition. The Spearman-Brown reliability of these ratings across raters is 0.75, suggesting that they represent a reasonably reliable measure of target knowledge. Consistent with self-reports, participants demonstrated greater essay-rated knowledge of High-Knowledge relative to Low-Knowledge targets (M_{high}=4.08, M_{low}=3.31, t=6.55, p<0.001). In contrast, Own Party and Opposition Party targets did not differ in essay-rated knowledge (M_{Own}=4.13, M_{Opp}=3.92, t=1.00, *ns*).

Target differences in Trait-level Idiosyncrasy, Similarity, and Positivity:

On the basis of participants' trait ratings of political figure targets during the scanner session, aggregate indices of target idiosyncrasy, positivity, and similarity were computed (see *Methods*). On average, participants judged High-Knowledge targets to be more idiosyncratic

than Low-Knowledge targets (M_{high}=2.33, M_{low}1.77, t=5.15, p<0.001), but not more positive (M_{high}=0.69, M_{low}=0.59, t=0.64, *ns*). Low-knowledge targets were judged to be more similar to the self than high-knowledge targets (M_{high}=5.76, M_{low}=6.11, t=2.61, p=0.02). Own Party targets were not judged to be more idiosyncratic than Opposition Party targets (M_{Own}=2.12, M_{Opp}=1.98, t=0.62, *ns*). However, as expected, on a trial-by-trial basis participants judged Own Party targets to be more similar to the self than Opposition Party targets (M_{Own}=6.67, M_{Opp}=5.20, t=5.97, p<0.001) and more positive overall (M_{Own}= 1.83, M_{Opp}=-0.55, t=8.58, p<0.001).

fMRI Results

Target-Level Experimental Factors: Knowledge and Party Affiliation

Our primary hypothesis was that MPFC activity would be greater for High-Knowledge targets than Low-Knowledge targets as the former would rely on ToM_p more than the latter. Additionally, we hypothesized that this effect would not be moderated by Own/Other Party status of the targets. Repeated-measures ANOVA indicated that MPFC was more responsive to High-Knowledge than to Low-Knowledge targets (F(1,13) =16.799, p=0.001), but not more responsive to Own Party than to Opposition Party targets (F(1,13)=0.025, *ns*), as shown in Figure 2 (One participant was dropped from target-level analyses. One ordered pair [essay-rated target knowledge, MPFC ROI parameter estimate] from this subject was identified as an extreme multivariate outlier. Cook's D for this case was 1.00, exceeding our threshold of 0.07.)

No interaction between target Knowledge and Party Affiliation was observed (F(1,13)=0.130, *ns*)). Follow-up comparisons between specific cells showed greater activity to High-Knowledge Own Party targets than to Low-Knowledge Own Party targets (t(13)=2.705, p=0.018), and to High-Knowledge Opposition Party targets than to Low-Knowledge Opposition Party targets (t(13)=3.004, p=0.010). Each of these effects was also identified when we searched

within the MPFC mask for significant clusters (see Table 2).

Participants also showed greater activity in the MPFC to Self trials than to Ordinary American trials ($t(13)=4.06$, $p=0.001$), replicating self-reference effects observed in previous research (Mitchell, Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006; Heatherton et al., 2006; Denny et al., 2012). The MPFC did not differentiate between the Self and High-Knowledge Own Party targets ($t(13)=0.718$, *ns*) or High-Knowledge Opposition Party targets ($t(13)=0.434$, *ns*). However, this region was significantly more active to Self trials than to Low-Knowledge Own Party targets ($t(13)=2.22$, $p=0.0451$) and Low-Knowledge Opposition Party targets ($t(13)=2.490$, $p=0.027$). This pattern of results is sensible if MPFC is responding to available knowledge concerning the target (including the Self), but is more difficult to explain if MPFC is encoding information concerning similarity or party affiliation.

Target-level Non-Experimental Factors:

Consistent with our primary hypothesis, essay-rated target knowledge demonstrated a significant linear relationship with activity within the MPFC ROI when controlling for closeness to the target ($t(13)=3.339$, $p=0.005$), but closeness did not significantly predict activity in this region ($t(13)=1.584$, *ns*). Separate analyses show a consistent positive relationship between essay-rated target knowledge and activity in the MPFC ROI, controlling separately for connectedness to target ($t(13)=3.157$, $p=0.008$), perceived personal overlap ($t(13)=4.244$, $p<0.001$), and overall similarity ($t(13)=3.472$, $p=0.004$). In contrast, there was no relationship between MPFC activity and target-level closeness ($t(13)=1.584$, *ns*), connectedness ($t(13)=1.192$, *ns*), overlap ($t(13)=0.238$, *ns*), or overall similarity ($t(13)=0.721$, *ns*). As can be seen in the plots of essay-rated target knowledge against parameter estimates from the MPFC ROI (target > control, Figure 3), MPFC activity is positively associated with essay-rated target knowledge for

political figures from Own and Opposition Parties. Each of these effects were also identified when we searched within the MPFC ROI for significant clusters (see Table 2).

These results are not easily explained by similarity and closeness accounts of MPFC function. The MPFC showed sensitivity to within-subjects variation in target knowledge (both as assessed using experimental contrasts and using independent, essay-rated target knowledge) that was not attenuated by controlling for self-reported closeness, connectedness, personal overlap, or overall similarity. In contrast, none of the other target-level factors significantly predicted activity in this region.

Trial-by-trial trait-level analyses of idiosyncrasy

Trial-by-trial analyses allow characterization of MPFC response to factors such as idiosyncrasy and similarity that vary from trait to trait both within and across targets. High idiosyncrasy trait-judgments cannot depend upon the deployment of ToM_g, as the target is perceived to differ from the ordinary American on the trait in question, and therefore ought instead to depend on a ToM_p. In order to rule out other explanations of hemodynamic response associated with our trial-by-trial index of target idiosyncrasy, analyses reported below employ statistical controls for perceived similarity to self and reaction time. These analyses were run across all trials, ignoring the identity of the target and target-level factors (e.g. closeness, party affiliation). The MPFC ROI demonstrated a significant linear relationship with trait-level idiosyncrasy ($t(14)=2.257$, $p=0.040$), controlling for trait similarity, positivity and reaction time, such that greater activity was observed for trials on which targets were judged to be more idiosyncratic. Similarity and positivity were each marginally significant as unique predictors of MPFC activity ($t(14)=1.919$, $p=0.076$ and $t(14)=1.798$, $p=0.094$, respectively), but were not significant controlling for idiosyncrasy.

Conjunction Analysis: Target-level essay-rated Knowledge and Trait-level Idiosyncrasy

To examine the relationship between essay-rated target knowledge and trait-level idiosyncrasy, we performed a conjunction analysis using the minimum statistic (Nichols et al., 2005) from these analyses (described above), constrained to the MPFC ROI. A significant cluster was detected within this region (peak MNI: -6, 47, 7; t=4.53, k=41; see Figure 4), suggesting that MPFC supports both person-specific mentalizing (ToM_p) as well as judgments of particularly idiosyncratic traits, regardless of overall target knowledge.

Self-idiosyncrasy

The ToM_p hypothesis of MPFC suggests that this region is not functionally devoted to self-processes, per se. Rather, according to the hypothesis, the representation of oneself is typically the most idiosyncratic person representation one has and thus, accessing self-knowledge recruits this region quite reliably. If this is the case, MPFC should be more active during self-judgments to the extent that a person views himself idiosyncratically on a particular trait. Contrary to expectations, a parametric modulation of self-judgments by trait-level self-idiosyncrasy (controlling for reaction time) failed to produce a significant response in MPFC ROI as a whole (t=1.397, *ns*). However, when we searched within the ROI we did observe a cluster whose activity was significantly associated with self-idiosyncrasy (MNI: -15, 48, -12; t=4.45, k=16; see Figure 5).

Discussion:

A variety of hypotheses have been advanced to explain the unique contribution of MPFC processes to human mentalizing (Lieberman, 2012). While each of the approaches adopted thus far are plausible, the results of the present research argue in favor of the ToM_p account of MPFC. This account is consistent with a wide range of prior results and integrates them under a more

coherent umbrella account of MPFC's social cognitive functions. Crucially, the MPFC was more responsive to High-Knowledge than to Low-Knowledge targets, regardless of whether the target was seen as similar or close to oneself (i.e. in one's own political party or the opposing party). In more fine-grained analysis, MPFC was more active when judging a target on a particular trait to the extent that the target was seen as idiosyncratic on that dimension, differing from both the typical person and the participant making the judgment. Finally, MPFC was also more active when judging the self on a particular trait to the extent that the self was seen as more idiosyncratic, consistent with the view that MPFC isn't a self-knowledge region, *per se*. The self, on this view, is merely the paradigmatic instance of an intimately well-known social target, and self-relevant cognition is therefore extremely likely to be associated with MPFC response.

Prior similarity findings are therefore consistent with the ToM_p account of MPFC. Personal similarity may lead the social thinker to project her own attributes and preferences on to others, and engage in mental simulation where appropriate when gauging similar others' mental states (Mitchell, Banaji, and Macrae, 2005; Mitchell, Macrae, and Banaji, 2006). According to the ToM_p account, this social reasoning strategy relies on the most person-specific representation one has: the self. However, it is just one of many person-specific mental models represented by MPFC and in other cases we may recruit this region when relying on another other non-self ToM_p. Unlike the similarity account, the ToM_p account also predicts that MPFC can be involved in projecting from individuals other than the self. If a new target is deemed similar to a friend or family member for whom we have ToM_p ("you remind me of my mother"), then MPFC should be recruited as that ToM_p is projected on to the novel target.

Prior closeness findings (Krienen et al., 2010) are also consistent with the ToM_p account of MPFC, insofar as we typically represent the minds of close others with detailed, idiosyncratic

models. However only the ToM_p account can explain why MPFC is more active when thinking about targets who are well-known but *neither* similar *nor* close to oneself, such as a well-known politician in an opposing political party. In addition, of the three accounts, only the ToM_p account also predicts that thinking about more idiosyncratic aspects of the self would differentially recruit MPFC.

Indeed, prior research has suggested that information about specific dispositional traits, which may serve as a scaffold for person-specific theories of mind, is encoded in the MPFC in the form of a trait code (Ma, Vandekerckhove, & Van Overwalle, 2010; Ma, Baetens, Vandekerckhove, Kestemont, Fias, & Van Overwalle, 2013; Ma, Baetens, Vandekerckhove, Van der Cruyssen, & Van Overwalle, 2013). Rather than simply encoding the valence of a given judgment, this research implies that traits are represented in a discrete fashion within MPFC. The availability of trait-specific information is a necessary precursor for the formation of a person-specific theory of mind, which in addition requires the association of specific traits with specific targets. The present study suggests that MPFC does in fact pair trait information with specific targets in order to produce consistent representations of the minds of specific individuals. Evidence concerning the existence of a trait code in the MPFC enhances the plausibility of the person-specific ToM account.

Person-specific ToM and group size

The ToM_p model of MPFC function is also consistent with findings inspired by Dunbar's (Dunbar, 1998) social brain hypothesis, which links brain size across primate species with group size. Although the group effects are often conceptually linked to the mentalizing network, including DMPFC, most of the group size findings observed in human and primate neuroimaging instead implicate MPFC. For example, a number of recent studies show a linear relationship

between MPFC volume and social network size (Powell et al., 2010; Powell et al., 2012; Lewis et al., 2011). Such results imply that MPFC may be essential for encoding information about multiple individuals in complex social arrangements. Successfully navigating a relatively large social network may depend upon the ability to deploy a diverse repertoire of person-specific ToMs, which may be facilitated by greater MPFC volume.

Most strikingly, Sallet and colleagues (Sallet et al., 2011) recently observed increased grey matter in MPFC when macaques were moved from smaller to larger living groups. If successful group living depended solely on generic forms of social cognition, it would scale easily to any group size because the same generic knowledge applies to all individuals. Instead, successful group living is partially about keeping track of idiosyncratic social information about each relevant individual and the distinctive relationships between individuals in the group. Thus, the canonical mentalizing system may support a basic capacity for group living, whereas MPFC may drive the size of the manageable group. Groups with complex social dynamics are characteristic of the human species, and indeed BA10 within MPFC is one of the only frontal regions known to be disproportionately larger in humans than other primates (Semendeferi, 2001).

Why a second system?

Evolutionary pressures may have expanded the volume of MPFC in primates and humans, supplying them with more robust social cognitive resources for living in increasingly complex groups. However, such considerations do not explain why person-specific and generic forms of mentalizing ought to be functionally and structurally differentiated in distinct subregions of the medial prefrontal cortex to begin with. Insight into this issue may be gained by consideration of analogous distinctions in the cognitive neuroscience of declarative memory.

The memory literature has long distinguished the roles of neocortical versus hippocampal systems in the formation, consolidation, and retrieval of memories over time (Squire, Stark, and Clark, 2004; Alvarez, and Squire, 1994). Long-term, relatively permanent representations of a semantic nature are thought to depend primarily on the neocortex, and to change only slowly over time. In contrast, the hippocampus serves a crucial but time-limited function in declarative memory, conjoining the distributed neocortical representations that together constitute the memory as a whole (Squire, 1992).

While the existence of multiple memory systems and their interaction during learning has thus been well characterized, the cognitive utility of encoding memory in distinct, overlapping forms has not always been evident. Using computational models, McClelland and colleagues (McClelland, McNaughton, and O'Reilly, 1995) presented an intriguing account of the complementary contributions of hippocampal and neocortical memory to the formation of categorical knowledge. They found that the attempt to encode information about the category membership of novel exemplars in a single, unitary connectionist system produced "catastrophic interference". In brief, if modeled knowledge structures were too malleable, they were ineffective at forming long-term categorical representations and integrating information from different observations over time. Instead, they were disproportionately biased by the unique features of each newly-presented exemplar. In contrast, more rigid knowledge structures were unable to evolve and incorporate novel information, because inflexible representations were insufficiently sensitive to new or atypical instances (e.g. learning that a penguin is a bird, despite phenotypic dissimilarity). McClelland et al. theorized that the hippocampal formation provides an intermediary system, capable of storing new information in a form that does not interfere destructively with established knowledge. Over time, the aggregation of relevant information

then shapes enduring category representations in the more slowly-adapting neocortex.

We propose that a mechanism of this type might help to explain the utility of different systems for generic and person-specific mentalizing processes. The social thinker cannot help but take for granted her extensive and sophisticated understanding of how the minds of others form beliefs, experience emotions, and make decisions. This wealth of information constitutes our generic theory of the human mind, and as such must be applicable to a diversity of situations. Precisely because its content has to apply broadly to countless situations, its representational features ought to be more fixed. By abstracting from the idiosyncrasies of any particular individual or experience, our generic theory of mind retains an essential generalizability that facilitates the understanding of novel mentalizing episodes. ToM_g ought therefore to develop slowly, and change only when many experiences considerably reshape our fundamental assumptions and expectations regarding human thought, feeling, and action. Consistent with this notion, prior research has suggested DMPFC in abstract mentalizing contexts in which the application of a generic ToM is appropriate (Spunt, Falk, & Lieberman, 2010; Spunt, Satpute, & Lieberman, 2011; Baetens, Ma, Steen, & Van Overwalle, 2013).

In contrast, a ToM_p would allow for the efficient and rapid encoding of unique, idiosyncratic information regarding particular, important individuals. Whether concerning friends or enemies, similar or dissimilar others, the use of ToM_p may allow more precise inferences concerning their mental states, and more accurate predictions of their future behavior. Separate systems for person-specific and generic mentalizing would allow us to take advantage of idiosyncratic knowledge in thinking about those we know well, without resulting in destructive interference with our more general theory of mind.

Conclusions:

The results reported above substantiate a novel account of MPFC function, according to which this region implements person-specific theory of mind (ToM_p) and enables nuanced, adaptive responding to the idiosyncratic characteristics of particular, well-known individuals. Person-specific ToM can account for the empirical findings associated with other approaches to mentalizing (e.g. similarity, closeness), but also provides new insights and testable predictions that go beyond these perspectives. Future research might elucidate the circumstances under which person-specific and generic ToM are deployed, and better characterize the psychological and neural correlates of the formation of person-specific theories. In addition, further work may help to uncover the relationship between person-specific ToM and other sources of mental state inference.

Tables:

Table 1

Descriptive behavioral data summarized for each of four political figure targets

		High Knowledge		Low Knowledge	
		Own Party	Opposition Party	Own Party	Opposition Party
Self-Report:	Knowledge (0-10)	8.00	7.80	3.53	3.33
	Closeness (0-10)	5.93	0.07	3.29	0.21
	Connectedness (0-10)	6.50	0.14	4.07	0.36
	Overall Similarity (0-10)	5.86	0.39	5.02	0.93
	Personal Overlap (1-7)	5.29	1.43	4.71	1.57
Rated:	Knowledge (1-7)	4.86	4.62	3.40	3.21
Trial-based:	Idiosyncrasy (0-8)	2.41	2.26	1.83	1.71
	Similarity (0-8)	6.57	4.95	6.77	5.44
	Positivity (-4 to 4)	2.17	-0.79	1.48	-0.31

Note: Self-report items reflect participant responses to questionnaire items. Scored knowledge data are derived from coder ratings of individual essays on the accomplishments, positions, values, and career of each political figure target. Trial-based items represent aggregate means of idiosyncrasy, similarity, and positivity across trait-judgments completed during the scanning session (see *Methods* for details regarding these indices).

Table 2

Summary of search-within analyses for the MPFC region-of-interest

Test Effect	x	y	z	t	k
Target-level analyses					
High > Low Knowledge (both parties)	-9	56	19	6.01	320
	-12	50	-11	5.08	
	-3	53	-2	4.73	
High > Low Knowledge (Own party)	6	59	13	4.17	59
	-6	53	-5	3.98	28
High > Low Knowledge (Opp party)	-3	56	13	6.7	165
Essay rated target knowledge	-6	50	13	7.27	173
Trial-by-trial Trait-level analyses					
Trait-level political-figure idiosyncrasy)	-6	42	3	5.29	73
Trait-level self-idiosyncrasy	-15	48	-12	4.45	16

Note: All results are FDR corrected $p<0.05$ with combined voxel-wise p and cluster-size thresholds (see *Methods*). Coordinates reported are from local maxima separated by at least 20 mm. x, y, and z: Montreal Neurological Institute coordinates in left-right, anterior-posterior, and inferior-superior dimensions; peak t: t-statistic value at the each local maxima; k: cluster voxel extent; cluster p(FWE): cluster-level FWE probability.

Figure 1: Example targets for a hypothetical Democratic participant, reflecting an experimental crossing of the Knowledge and Party Affiliation factors. Targets were selected independently for each participant, based on self-reported knowledge of Own Party and Opposition Party political figures. See Table 1 for descriptive data related to each target.

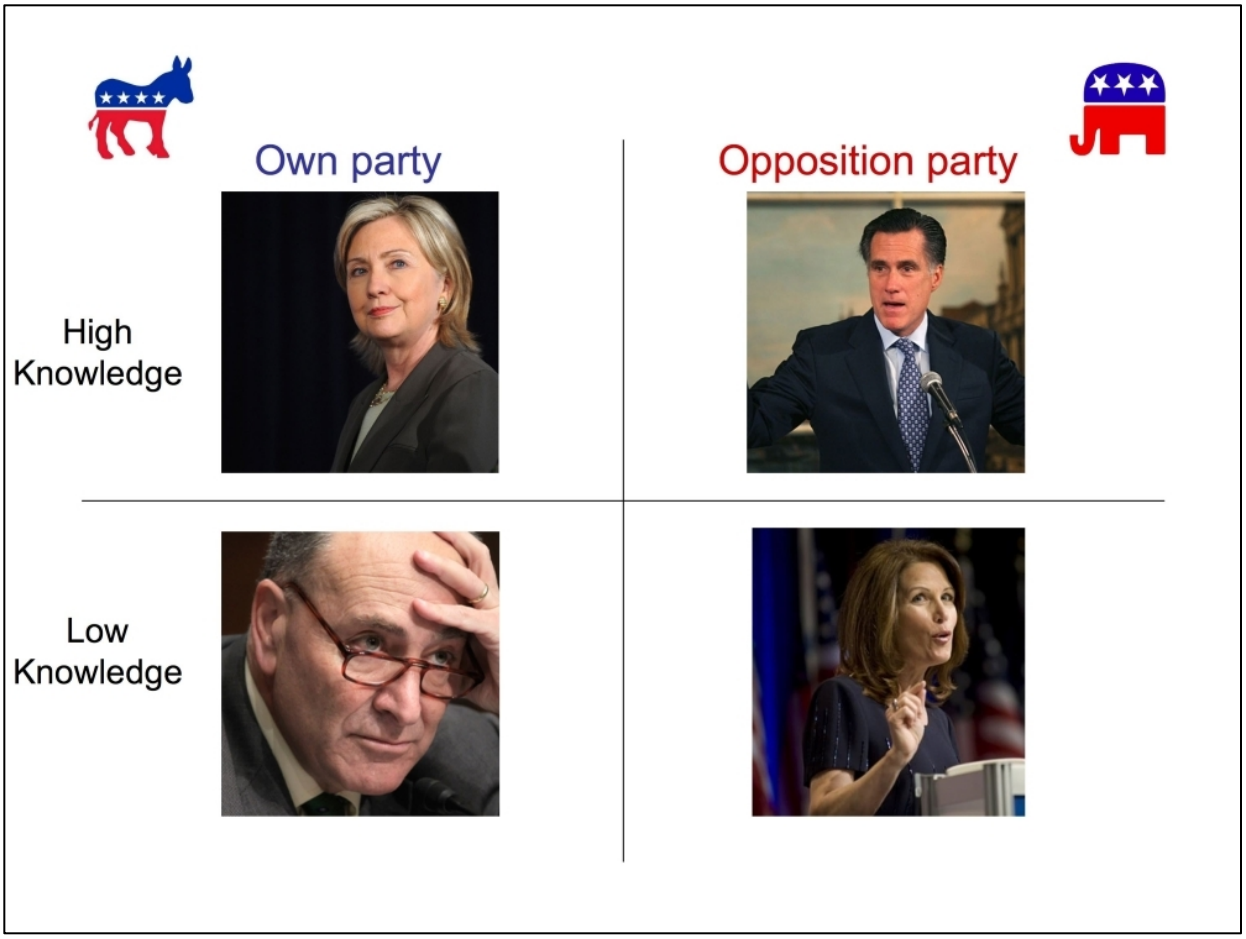


Figure 2. MPFC demonstrated greater activity to High-knowledge than to Low-knowledge targets. There was no significant effect of Party Affiliation on activity in the MPFC and no interaction between Party Affiliation and target Knowledge. MPFC also demonstrated greater activity to Self trials than to Ordinary American trials. Parameter estimates (relative to control) are plotted separately for each target. Error bars reflect the SEM. See also Table 2 for associated search-within analysis.

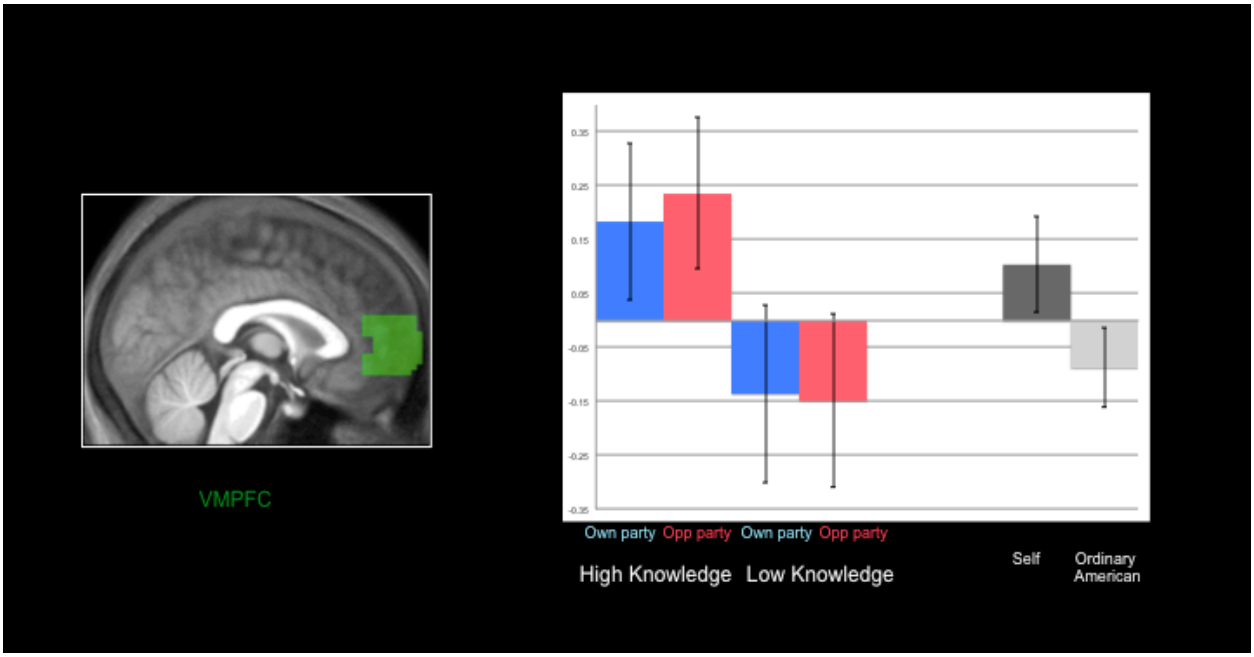


Figure 3. Activity within the MPFC ROI demonstrated a significant linear relationship with essay-rated target knowledge for both Own Party (A) and Opposition Party (B) political figures. To present this relationship graphically, we plot MPFC parameter estimates (p.e.) from the block target > control comparison against essay-rated target knowledge. 'X' marks represent self-reported High-Knowledge targets; 'O' marks represent self-reported Low-Knowledge targets. See also Table 2 for associated search-within analysis.

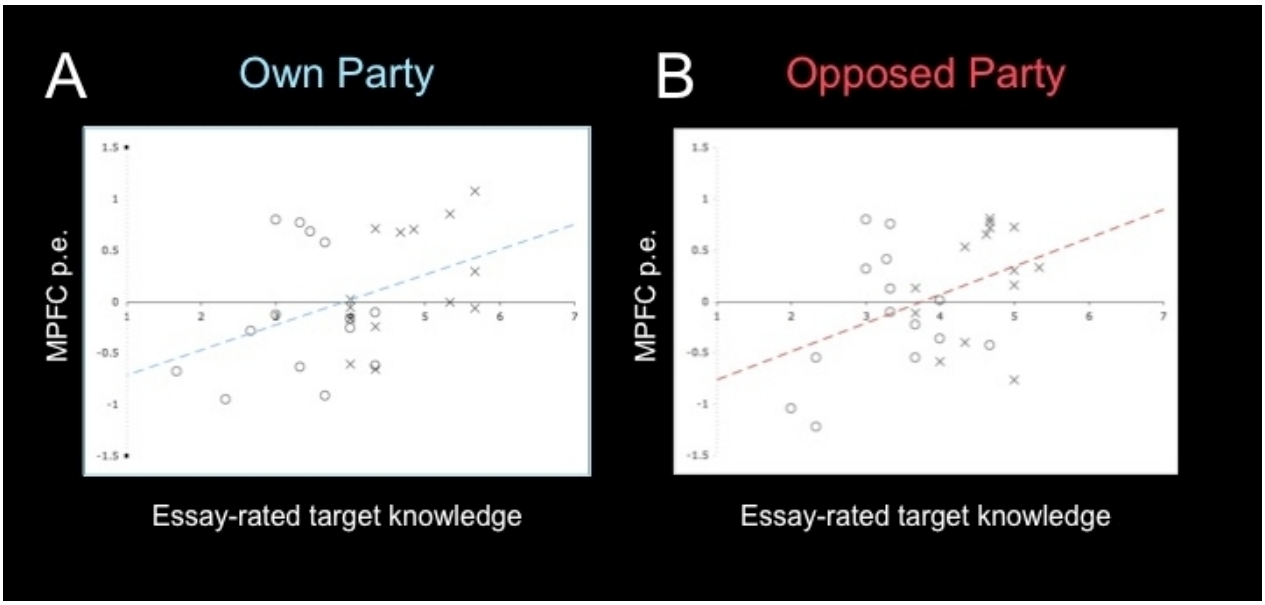


Figure 4. Essay-rated target knowledge and trait-level idiosyncrasy demonstrate overlapping neural correlates. Results of a conjunction analysis constrained to the MPFC are displayed, showing activity associated with essay-rated target knowledge (green), trait-by-trait idiosyncrasy (red), and their conjunction (in yellow). The conjunction cluster consisted of 41 voxels, with peak at MNI -6, 47, 7. See also Table 2 and Figure 5.

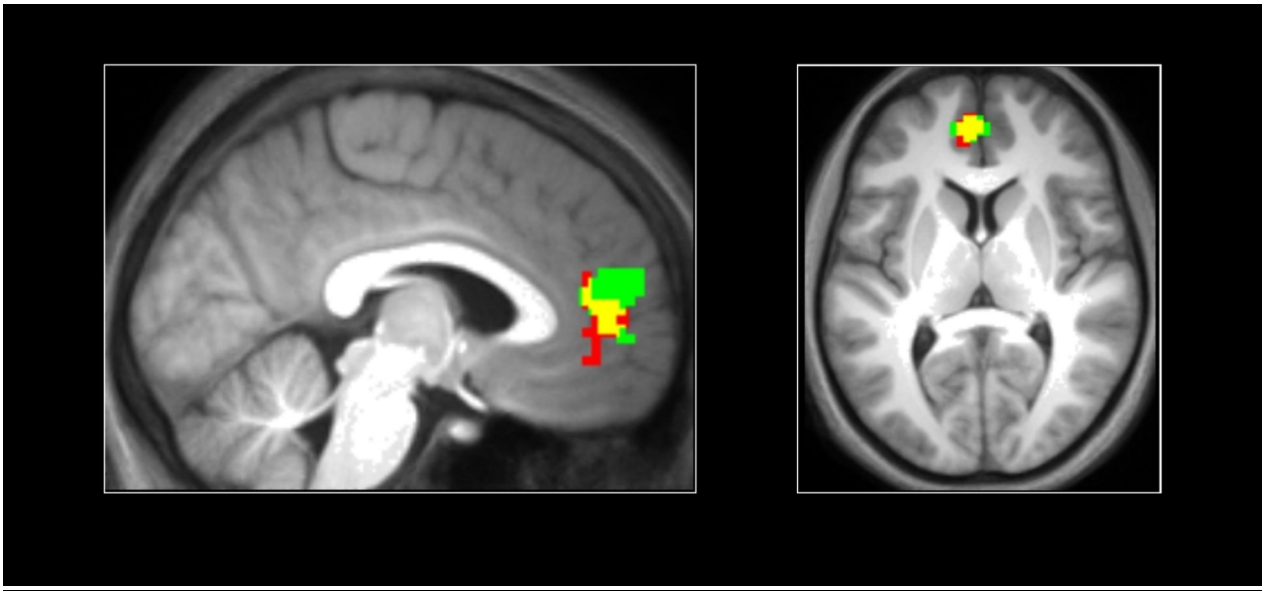
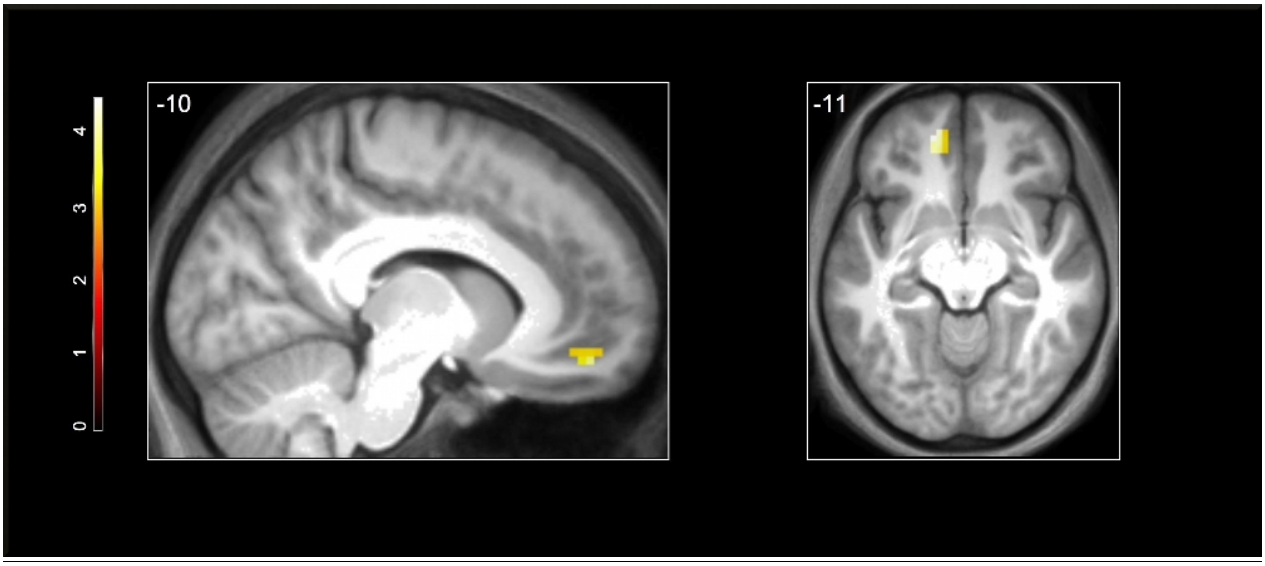


Figure 5. Trial-by-trial self-idiosyncrasy (difference from ordinary American) predicted activity in the VMPFC region-of-interest (while controlling for judgment positivity) when assessing the applicability of trait words to the self. Peak MNI: -15, 48, -12, FDR $p<0.05$.



REFERENCES

Alvarez, P., Squire, L.R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. Proc. Natl. Acad. Sci. USA 91(15), 7041-7045.

Amodio, D.M., and Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. 7, 268-277.

Ashburner, J. (2007). A Fast Diffeomorphic Image Registration Algorithm. Neuroimage 38, 95-113.

Brett, M., Anton, J-L., Valabregue, R., and Poline, J-B. (2002). Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in Neuroimage 16(2).

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162-173.

Dunbar, R.I.M. (1998). The social brain hypothesis. Evol. Anthropol. 6(5), 178-190.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., and Heatherton, T.F. (2002). Finding the self? An event-related fMRI study. J. Cogn. Neurosci. 14, 785–94.

Krienen, F.M., Tu, P-C., and Buckner, R.L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. J. Neurosci. 30(41), 13906-13915.

Lewis, P.A., Rezaie, R., Brown, R., Roberts, N., and Dunbar, R.I.M. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. Neuroimage 57(4), 1624-1629.

Lieberman, M.D. (2010). Social Cognitive Neuroscience. In Handbook of Social Psychology (5th ed.), S.T. Fiske, D.T. Gilbert, and G. Lindzey, eds. (New York, NY: McGraw-Hill), pp. 143-193.

Lieberman, M.D. (2012). Self-knowledge: From philosophy to neuroscience to psychology. In Handbook of Self-knowledge, S. Vazire and T.D. Wilson, eds. (New York, NY: Guilford), pp. 63-76.

Maldjian, J.A., Laurienti, P.J., Burdette, J.B., and Kraft, R.A. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19, 1233-1239.

McClelland, J.L., McNaughton, B.L., O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. 102(3), 419-457.

Mitchell, J.P., Banaji, M.R., and Macrae, C.N. (2005). The Link between Social Cognition and Self-referential Thought in the Medial Prefrontal Cortex. J. Cogn. Neurosci. 17(8): 1306-1315.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. Neuron 50, 655-663.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J-B. (2005). Valid conjunction inference with the minimum statistic. Neuroimage 25(3), 653-660.

Ochsner, K.N., Beer, J.S., Robertson, E.R., Cooper, J.C., Gabrieli, J.D., Kihlstrom, J.F., and D'Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. Neuroimage 28, 797–814.

Powell, J.L., Lewis, P.A., Dunbar, R.I.M., Garcia-Finana, M., and Roberts, N. (2010). Orbital prefrontal cortex volume correlates with social cognitive competence. Neuropsychologia 48, 3554-3562.

Powell, J.L., Lewis, P.A., Roberts, N., Garcia-Finana, M., and Dunbar, R.I.M. (2012). Orbital prefrontal cortex volume predicts social network size: an imaging study of individual differences in humans. Proc. Biol. Sci. 279(1736), 2157-2162.

Sallet, J., Mars, R.B., Noonan, M.P., Andersson, J.L., O'Reilly, J.X., Jbabdi, S., Croxson, P.L., Jenkinson, M., Miller, K.L., and Rushworth, M.F.S. (2011). Social Network Size Affects Neural Circuits in Macaques. Science 334(6056), 697-700.

Semendeferi, K., Armstrong, E., Schleicher, A., Ziles, K., and Van Hoessen, G.W. (2001). Prefrontal cortex in humans and apes: a comparative stud of area 10. Am. J. Phys. Anthropol. 114(3), 224-241.

Squire, L.R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychol. Rev. 99(2), 195-231.

Squire, L.R., Stark, C.E., and Clark, R.E. (2004). The medial temporal lobe. Annu. Rev. Neurosci. 27, 279-306.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273-89.

Van Overwalle, F. (2009). Social Cognition and the Brain: A Meta-Analysis. Hum. Brain Mapp. 30, 829-858.

Vanderwal, T., Hunyadi, E., Grupe, D.W., Connors, C.M., and Schultz, R.T. (2008). Self, mother

and abstract other: An fMRI study of reflective social processing. Neuroimage 41, 1437-1446.

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of false beliefs in young children's understanding of deception. Cognition 13, 103-128.

PAPER 2:

Formation and Deployment of Person-Specific Theories of Mind

Abstract

The research reported above (in Study 1: Welborn & Lieberman, 2015) introduced the notion that person-specific theories of mind – individuated mental models facilitating nuanced, idiosyncratic inferences about others' transient mental states and enduring dispositional traits – might depend upon regions of ventral medial prefrontal cortex (corresponding to medial Brodmann's Area 10). Hemodynamic activity in this region was found to vary with participants' overall knowledge of a given social target as well as their trait-specific perception of the target's unique idiosyncrasies. Moreover, these effects did not seem to depend upon the perceived similarity of the target to the self, felt closeness or connectedness, or the target's party affiliation (despite the strong political commitments of our participants). While an important first step towards an understanding of person-specific theories of mind, this study analyzed mentalizing processes in an atypical sample (strongly committed political partisans) in a domain in which they were likely highly motivated to reason about the target's beliefs and attitudes. In addition, this first attempt characterized the neural correlates of established, pre-existing mental models, and the mechanisms involved may differ considerably during the formation of new person-specific theories of mind and their subsequent deployment. For these reasons, we sought to replicate and extend our prior work, investigating the formation and use of person-specific theories of mind regarding novel, previously unfamiliar social targets.

In the present study, mentalizing regions demonstrate greater response to social targets about whom participants have recently acquired a person-specific theory of mind than to targets about whom participants have not received detailed personal information. Moreover, when assessing the personality traits of well-known and poorly-known targets, MPFC exhibited elevated activity when participants confidently judged the target to differ from the ordinary

person. Taken together, these results suggest that: 1) person-specific theories of mind can be rapidly acquired for previously unknown targets, and 2) that their subsequent deployment in mental state reasoning is associated with engagement of mentalizing regions, including especially MPFC.

Introduction:

A number of important studies in field of social cognitive neuroscience have explored differences in hemodynamic responses to well-known targets as compared to less familiar targets. However, as noted above, the major limitation of these early endeavors is the clear presence of confounding factors, including especially perceived personal similarity as well as affective factors such as felt closeness (Krienen, Tu, & Buckner, 2012). Other neuroimaging studies have focused on the neural bases of impression formation, often with very limited or superficial information about social targets (Kang et al., 2015) or sought to understand the neural processes that support category-based judgments or stereotypes (Kaul et al., 2014). While this research is quite important in its own right, and pertain to phenomena of considerable import to social psychology and social neuroscience, they do not provide discriminating data that directly bears upon the hypothesis that human mentalizing may involve the selective use of person-specific mental models. To our knowledge, no previous neuroimaging study has introduced participants to novel social targets using repeated, extended exposure to revealing person-specific information, comparable to the experiences that inform our knowledge of real-world social targets.

In our second neuroimaging study on person-specific theory of mind, we thus introduced participants to novel (previously unfamiliar) social targets through materials drawn from contemporary television dramas. Using these materials, participants were exposed to varying degrees of information, with the aim of creating distinct levels of knowledge concerning three social targets, about whom participants would possess: 1) high knowledge, 2) intermediate knowledge, and 3) no knowledge. In order to induce the formation of a person-specific theory of mind regarding a 'high knowledge' target, each participant was assigned to watch five episodes

of a selected television serial in which the protagonist's personality was firmly individuated. Participants learned about an 'intermediate knowledge' target during the scanning session itself, during which they viewed a brief video montage which introduced a new, previously unknown character. Participants received no information at all regarding the 'no knowledge' target prior to the experimental tasks. In this way, we hoped to contrast person-specific and generic mentalizing processes in an experimental context in which participants would not possess antecedent attitudes or knowledge about the targets.

While undergoing functional magnetic resonance imaging, participants completed both a mental state reasoning task (the well-established Why-How task employed in, e.g. Spunt & Lieberman, 2012 and Spunt, Falk, & Lieberman, 2010) as well as a trait-judgment task similar to that used in Welborn and Lieberman (2015). In both of these tasks, we predicted that MPFC activity would covary with the availability of person-specific knowledge about the target characters, showing greatest and least responsiveness respectively to the high-knowledge and no-knowledge targets. Moreover, we sought to test whether or not participants' confidence in making person-specific evaluations of target traits would correlate with engagement of the MPFC. We predicted an interaction between trial-level confidence and trial-level idiosyncrasy, such that participants would recruit MPFC most when they felt they had great confidence in judging the target to be distinct from the ordinary individual.

Methods:

Participants

Participants were 17 young adults (age M=23.2) recruited at UCLA through a variety of methods, including email solicitation and posted advertisements. In order to ensure that participants did not have antecedent knowledge or attitudes about the target characters,

participants were ineligible to participate if they had viewed any episodes of any of the television serials used as stimuli (see below). Participants were also excluded if they were left-handed, currently using psychoactive medications or drugs, had been diagnosed with a psychiatric or neurological illness, were pregnant, had experienced claustrophobia, and under any circumstances in which safety considerations precluded participation in fMRI research. All participants provided written, informed consent prior to all research activities as approved by the UCLA Institutional Review Board, and were compensated $50.00 for the participation in our research.

Television Serial Selection:

The television serials employed as training stimuli for this experiment were *House of Cards, The Newsroom,* and *Scandal*; the focal characters were these shows' respective protagonists, Frank Underwood, Will McAvoy, and Olivia Pope. Each are on-going and popular serials presented by major television broadcasting or online streaming services. These television serials were selected based upon four principal factors: 1) complex, multifaceted characterization of protagonists, 2) comparable narrative structure, 3) similarity of themes and subject-matter, and 4) roughly equal duration of episodes.

First, we wished to induce participants to form person-specific theories regarding the protagonists of the television serials employed, but did not necessarily desire for participants to develop feelings of closeness or connectedness to these characters, or to perceive these characters to be similar to themselves. For this reason, we selected serials with complex protagonists, who generally possessed both desirable and undesirable personality traits, engaged in both prosocial and aggressive behaviors, and expressed a range of positive and negative emotions. Second, these serials each depict their protagonists from a number of distinct vantage

points, illustrating their different social roles and varying relationships. Each protagonist experiences conflict in both personal and professional domains, and must tailor their self-presentations to different audiences or constituencies. Third, these serials all share a thematic focus on politics and public policy, exploring in different ways the relationships between government, media, and the population at large. While they differ substantially in emphasis, tone, and ideological orientation, we felt that this similarity in theme would help to limit the effect of extraneous differences and somewhat mitigate the possibility of confounds. None of these shows involves fantastic or supernatural characters, events, or locations, and all are set in contemporary American society. Lastly, episodes of each serial lasted between 43 and 52 minutes, provide (over the course of the five episodes watched) upwards of four hours of exposure to the central characters.

Overview of Experimental Procedure and Tasks Design:

For each participant, the three protagonists were randomly assigned to serve as the high-knowledge, intermediate-knowledge, or no-knowledge targets. As described below, high-knowledge targets were introduced through exposure to five episodes of the respective television serial, intermediate-knowledge targets were introduced through a brief video montage, and no-knowledge targets were *not* introduced to participants prior to completion of the experimental tasks during fMRI scanning. Participants were unaware of the identity of the intermediate-knowledge and no-knowledge targets before the scanning component of the study commenced.

Prior to the scanning session, participants were instructed to watch (all and only) the first five episodes of one of the three television serials described above (*House of Cards, The Newsroom,* or *Scandal*). Participants watched each episode at home, and at their convenience, subject to the following restrictions: 1) episodes were viewed in the correct order in an

environment free of distractions, 2) participants watched no more than one episode per day, 3) no more than a single day elapsed between viewing of successive episodes, 4) participants did not watch any of the episodes more than once, 5) participants did not watch any of the other television serials to be used as experimental stimuli during the scanning session. Participants were scheduled for a scanning session within one week after complete viewing of all five episodes of the assigned television serial.

During the scanner session, participants were exposed to a brief (~5 minute) video montage that served to introduce a novel (intermediate-knowledge) target: the protagonist of one of the other (un-watched) television serials. Extracted video clips from the first five episodes of the relevant serial were employed in the creation of these brief montages, and did not overlap in content with any of the stimuli employed in the person-specific Why/How task (see below; Why/How stimuli are derived from episodes 6+ for each serial). Each montage aimed to acquaint the participant with the cardinal features of the focal character's personality by depicting typical dialogue or actions. However, insofar as possible, the montage segments did not reveal crucial plots elements from the un-watched televisions shows, and only included incidental depictions of non-focal characters.

Person-Specific Why/How Task:

While undergoing functional magnetic resonance imaging (fMRI) participants completed two tasks in which they made judgments regarding the mental states and personality traits of each character. The first task was based upon the Emotional Scenes version of the Why/How task employed by Spunt & Lieberman (2012). In this variation of the Why/How paradigm participants had been presented with brief video segments from the television serial *Gossip Girl* in which characters expressed strong emotions through their facial features, gestures, or bodily

movements. On separate trials, participants had been asked to assess either 'Why' the target character was experiencing the emotion depicted in the video-clip or 'How' the target character was expressing his or her feelings. Spunt and Lieberman (2012) found that this task manipulation evoked significant shifts in hemodynamic activity within brain regions associated with mentalizing (during 'Why' trials) and action understanding (during 'How' trials). The Why/How task employed in the present study is similar to that described by Spunt and Lieberman (2012), except that the brief video-clips are excerpted from the television serials *House of Cards, The Newsroom,* and *Scandal* instead of *Gossip Girl.* It should be noted, however, that participants in Spunt and Lieberman (2012) were not familiar with *Gossip Girl* and did not therefore have access to person-specific knowledge that might assist them in the evaluation of the target characters' emotional expressions. In the present study, participants have extensive information about the high-knowledge target from having viewed five episodes of the selected television serial, and have limited information about the intermediate-knowledge target from having viewed the brief montage.

The Why/How task used in the present study was presented over two functional runs, with 4 blocks of 'Why' trials and 4 blocks of 'How' trials for each target (high-knowledge, intermediate-knowledge, and no-knowledge). During control blocks, participants were asked to perform a shape-matching task, selecting which of two similar candidate geometric figures matched a target. Therefore, taken together, there were a total of seven conditions during the Why/How task: High-Knowledge 'Why', High-Knowledge 'How', Intermediate-Knowledge 'Why', Intermediate-Knowledge 'How', No-Knowledge 'Why', No-Knowledge 'How', and Shape-Matching control.

Twenty video segments (without audio) of approximately 4 seconds duration each were extracted for each target from episodes participants would not previously have viewed (i.e. no video-clips were extracted from the first 5 episodes of any of the television serials), for a total of 60 total video segments. These segments were blocked in groups of five for presentation during the Why/How task, each preceded by an instruction prompt (e.g. 'Why is Frank Underwood feeling it?' or 'How is Frank Underwood expressing his feelings?'). Within each block, individuals segments were separated by brief, jittered fixation inter-trial intervals (ITIs) drawn from an exponential distribution with a mean of 2 seconds. Blocks were presented in a pseudo-randomized order that maximized contrast efficiency, subject to the restriction that no combination of target and prompt could be repeated sequentially. All video segments for the Why/How task featured the target character as the primary focus (usually in 'close-up'), and were designed to minimize cuts as well as dialogue. Given the absence of contextual information or other cues as to the causes of the target's emotional experience, participants would need to rely upon either their specific knowledge of the target characters' dispositions or deploy a generic theory of mind to successfully complete the task.

Importantly, each video segment was present twice (once per functional run). The assignment of video segments for each target to the 'Why' or 'How' conditions was counterbalanced across runs such that if a video segment was included in a 'Why' block in the first functional run, it would be included in a 'How' block during the second functional run. Thus, the visual stimuli for the 'Why' and 'How' blocks for each target were identical across the course of the experiment; only the instructional prompt and the cognitive set thereby induced distinguished these stimuli. Given that the order of the video segments used was randomized across subjects, we did not feel that order effects would be problematic (i.e. participants were

just as likely to view a segment in the 'How' condition first as to view the same segment in the 'Why' condition first).

Person-Specific Trait-Judgment Task:

The second scanning-session task was very similar to the trait-judgment task included in Study 1, from which it was derived. In this task, participants used on-screen nine-point Likert-type scales to rate the applicability of 30 trait words (e.g. 'intelligent', 'dishonest', 'vengeful', 'altruistic') to each of the three targets. The on-screen scale was anchored at '1 – Not at All' applicable and '9 – Completely' applicable, respectively. A mixed block-event related design was used for the trait-judgment task, with trials organized into blocks of 5 items each by target type. Across two separate functional runs, participants completed 6 blocks of 5 trials each for the high-knowledge, intermediate-knowledge, and no-knowledge targets, for a total of 30 trials each per target and 90 trials total. Each trial was constrained to a maximum duration of 5 seconds, followed by a jittered ITI drawn from an exponential distribution with a mean of 2 seconds. If participants did not manipulate the scale and confirm their response selection within the trial duration, the next trial would commence automatically after the ITI. During each trial, the trait word and the name of the target remained on-alongside the scale until participants confirmed a response. Each block was also preceded by a 2 second introduction, displaying the name of the target character for the up-coming block.

The trait-judgment task also included a basic visuo-motor control condition in which participants indicated whether trait words were presented in all lower-case or all upper-case letters. The same trait words were used for the case-judgment control trials as in the trait-judgment experimental conditions, and words appeared randomly in upper- or lower-case with equal frequency. Case-judgment trials were grouped similarly into blocks consisting of 5 trials

separated by a jittered ITI and preceded by a 2-second introduction indicating that the up-coming block would require judgments of case. Each trait was presented only once in the case-judgment control condition, for a total of 30 trials across 2 functional runs, divided equally between 6 blocks.

After completion of scanning, participants performed a number of additional tasks. Importantly, participants rated the self and the 'ordinary person' on each of the 30 trait items used in the person-specific trait-judgment task using identical 9-point Likert-type scales. On the basis of these ratings, trait-level (dis)similarity scores were computed as | target rating – self rating | and idiosyncrasy scores were computed as | target rating – ordinary person rating |. These definitions are identical to those used in Study 1, and were designed to facilitate parametric modulation analyses comparable to those presented above. In addition, participants indicated retrospectively their subjective confidence in the accuracy of each of the 90 trait-judgments they had made during the scanning session, using a nine-point Likert-type scale anchored at '1 – Not at All Confident' and '9 – Extremely Confident'. These ratings were intended to measure indirectly the availability of person-specific knowledge relevant to making each trait-judgment. While it is possible that participants might be able to draw upon category-based information, stereotypes, or a generic theory of mind in order to confidently make trait-judgments in some instances, we felt that participants would be especially likely to employ person-specific theory of mind processes on trials in which they confidently judged the target individual to differ from the ordinary person (i.e. exhibit high idiosyncrasy).

Lastly, participants completed a number of post-scanning questionnaires, including verions of Narrative Transport Scale adapt for use with the television serial and the montage stimuli, and the Empathy Subscale of the Interpersonal Reactivity Index. Participants also

assessed their holistic similarity to each target character, as well as feelings of closeness and connectedness, using a 0 – 10 scale (inclusive), and evaluated the degree of inter-personal overlap with each target character using a 1 – 7 scale (inclusive). These items were strictly analogous to those employed in Study 1.

fMRI Data Acquisition:

All imaging data was acquired using a 3.0-Tesla Siemens Prisma scanner at the Ahmanson-Lovelace Brain Mapping Center at UCLA. Across 5 functional runs, approximately 2,880 T2*-weighted were acquired during completion of experimental tasks described above using a pulse sequence derived from the Human Connectome Project's multiband acquisition parameters (slice acceleration factor = 8, slice thickness=2 mm, gap=0 mm, TR=720 ms, TE=33 ms, flip angle=90°, matrix=64 x 64, field of view=200 mm). An oblique slice angle was used in order to minimize signal drop-out in ventral medial portions of the brain. In addition, we acquired a T1-weighted magnetically-prepared rapid acquisition gradient echo anatomical image (slice thickness=1 mm, 176 slices, TR=2530 ms, TE=3.31 ms, flip angle=7°, matrix=256 x 256, field of view=256 mm).

fMRI Preprocessing and Analysis:

Functional data were analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Within each functional run, image volumes were realigned to correct for head motion and coregistered to the high-resolution MPRAGE structural image. MPRAGE images were segmented by tissue type (grey matter, white matter, and cerebro-spinal fluid), and this segmentation served as the basis for normalizing structural and functional images into standard MNI stereotactic space (resampled at 2 x 2 x 2 mm). Finally, images were smoothed

with an 6 mm Gaussian kernel, FWHM.

For the majority of the analysis reported below, whole-brain univariate statistical analyses were performed at each voxel and corrected for multiple comparisons by using a combination of voxel-wise p and cluster-size thresholds to limit the false discovery rate (FDR) to less than 0.05. Monte Carlo simulations implemented in 3dClustSim (from AFNI; Cox et al., 1996) were used to determine appropriate cluster-size thresholds given the smoothness of the images (60 contiguous voxels) to ensure overall false discovery rate (FDR) of less than 0.05, when combined with a voxel-wise significance threshold of $p<0.005$. All results reported (except where indicated) exceed these joint voxel-wise and cluster-extent thresholds.

Analysis of the Person-Specific ToM Why/How Task:

For the person-specific theory of mind Why/How task, primary hypotheses concerned contrasts between targets and between the Why and How processing instructions for each target. As noted above (see *Overview of Experimental Procedure and Task Design*), the combination of target and processing instructions yielded a total of seven conditions during this task: 1) 'Why' high-knowledge, 2) 'How' high-knowledge, 3) 'Why' intermediate-knowledge, 4) 'How' intermediate-knowledge, 5) 'Why' no-knowledge, 6) 'How' no-knowledge, and 7) Shape-Matching control. A general linear model (GLM) was defined base upon the block and trial orders in which these conditions occured for each participant. 'Why/How' trials were modeled as fixed epochs spanning the ~4 second duration between the onset and offset of each video segment, convolved with the canonical (double-gamma) hemodynamic response function (HRF). Shape-Matching trials were modeled as variable epochs spanning the duration between the onset of each trial and participant response (or stimulus offset in the absence of response), convolved with the HRF. In total, five regressors of interest were modeled (one for each condition), and 18

motion regressors (3 translations and rotations, as well as their squares and first-order derivatives) for purposes of statistical control. The time series was not high-pass filtered because the large number of conditions and block design employed created substantial temporal gaps between subsequent repetitions of some conditions. The use of a high-pass filter might therefore have removed variance of interest. Serial autocorrelations were modeled as an AR(1) process. Contrast images were averaged across runs for each participant, and entered into a mixed effects analysis at the group level.

In some instances, parameter estimates for selected contrasts were extracted from statistically significant clusters using MarsBaR (Brett, Anton, Valabregue, and Poline, 2002). These parameter estimates are presented purely for illustrative purposes, as any statistical tests conducted on these ROIs are not independent of the whole-brain contrasts that yielded the clusters.

Analysis of the Person-Specific Trait Judgment Task:

In the person-specific trait-judgment task, participants evaluated the applicability of various trait words to each of the target characters using an on-screen scale, and completed a case-judgment control task. As in the person-specific Why/How task above, contrasts between hemodynamic responses to each target were of considerable interest, but for this task we also sought to explore the ways in which trial-by-trial variation in the evaluation of each trait would modulate response. For this reasons, we also conducted selective analyses of parametric modulation. In the basic model, four regressors of interest (one for each condition: high-knowledge, intermediate-knowledge, no-knowledge, and case-judgment control) were included in a GLM defined for each participant. Each event was modeled as a variable epoch spanning the duration between stimulus onset and response selection (a maximum of 6 seconds). In additional

GLM models, parametric modulators reflecting judgment confidence and perceived trait idiosyncrasy were included as additional regressors of interest. As above, all analyses controlled for 18 motion parameters and modeled serial autocorrelations as an AR(1) process. Contrast images were averaged across runs for each participant, and entered into a mixed effects analysis at the group level.

As mentioned above (see *Overview of Experimental Procedure and Task Design*), confidence scores reflected participants self-reported (retrospective) confidence in the accuracy of each trait judgment made for each target, and were measured on a 1 – 9 Likert-type scale. Idiosyncrasy scores were calculated based upon the absolute value of the difference between ratings of the target on a given trait and ratings of the ordinary person on that trait (i.e. idiosyncrasy score = | target rating – ordinary person rating |), as in Study 1. Scores were mean-centered within-subjects prior to all analyses.

Results:

Behavioral Results:

As expected, participants exhibited much greater confidence in their judgments of the high-knowledge relative to intermediate-knowledge targets ($t=3.11$, $p=0.007$) and low-knowledge targets ($t=4.49$, $p<0.001$). In addition, participants reported greater confidence in the accuracy of their judgments regarding the intermediate-knowledge (montage) targets than the low-knowledge targets ($t=2.43$, $p=0.027$). High-knowledge targets were also judged to be more idiosyncratic than intermediate-knowledge targets ($t=2.97$, $p=0.008$). Contrary to expectations, high-knowledge targets were not judged to be more idiosyncratic than low-knowledge targets ($t=1.60$, $p=0.12$). However, this test may not be accurate due to the presence of an extreme outlier, whose mean low-knowledge target idiosyncrasy scores was 3.15 standard deviations

above the mean (Grub's test for a single outlier was significant, with a G of 4.20 relative to a critical value of 2.74). With the exclusion of this outlier, participants rated high-knowledge targets to be more idiosyncratic than low-knowledge targets ($t(15)=3.51$, $p=0.002$). Participants did not judge the intermediate-knowledge targets to be more idiosyncratic than the low-knowledge targets, whether the outlier mentioned above was included ($t(16)=0.085$, *ns*) or not ($t(15)=1.37$, *ns*). Taken together, these results suggest that watching the television serial and montage stimuli increased participants' confidence in the judgments of the relevant targets, and facilitated making more idiosyncratic judgments about the personality traits. For the high-knowledge targets especially, we view these results as a successful manipulation check, indicating that participants have likely formed a person-specific mental model with which they can confidently differentiate the high-knowledge target from the others in mental state reasoning. Overall, confidence and idiosyncrasy were not especially correlated on a trial-by-trial basis ($r=0.12$, *ns*), suggesting that these constructs may capture different aspects of person-specific mentalizing.

In contrast, participants did not rate themselves (on average) to be more similar to the high-knowledge targets than to either the intermediate-knowledge ($t=0.83$, *ns*) or the low-knowledge targets ($t=0.30$, *ns*).

fMRI results:

Person-specific Why/How task:

Previous research using the Why/How paradigm has revealed robust activity in the dorsal medial prefrontal cortex during the 'Why' trials, consistent with the recruitment of this region during mental state reasoning (Spunt and Lieberman, 2012; Spunt, Falk, & Lieberman, 2010). First, we sought to determine whether or not the Why > How comparison, collapsing across

distinctions between different targets, would reveal a similar pattern of activation with in the present variation of this task. Indeed, the Why > How contrast across all targets reveals robust activity in the DMPFC, as well as other portions of the mentalizing network, including the posterior cingulate cortex(PCC)/precuneus, left and right temporo-parietal junctions (LTPJ and RTPJ), and the temporal pole bilaterally (see Figure 1A, and Table 2). The reverse contrast showed selective activation during the 'How' trials in the left and right supramarginal gyri (SMG), the left precentral gyrus, and the pars triangularis of the left inferior frontal gyrus (IFG) (see Figure 1B and Table 2). These regions have been implicated in action planning and action understanding as part of the human mirror neuron system (Iacoboni et al., 2005) and replicate prior results with the Why/How paradigm (Spunt & Lieberman, 2012; Spunt, Falk, & Lieberman, 2010).

Next, we sought to characterize regions that would show greater hemodynamic activity when reasoning about the mental states of the high-knowledge target, for whom we hypothesized that our participants would employ a person-specific theory of mind. Comparison of 'Why' and 'How' trials for the high-knowledge targets indicated that, as anticipated, participants recruited MPFC (within Brodmann's area 10) as well as DMPFC, PCC/precuneus, left and right TPJ, and bilateral temporal pole (See Figure 2A, and Table 2). Parameter estimates extracted from the BA10 MPFC cluster show the expected pattern of activity across targets (see Figure 2C). Comparison of these parameter estimates cannot be evaluated as a statistical test independent of the Why > How contrast for the high-knowledge target, but are presented merely to illustrate an overall pattern that we hope will be statistically significant in its own right with a more complete sample.

For the no-knowledge target, the Why > How contrast reveals significant differences in activity in several mentalizing regions, including the precuneus, MPFC, and DMPFC (see Figure 2B, and Table 2. Parameter estimates for the Why > How contrast were extracted from the DMPFC cluster for each target, but do not exhibit any significant differences (see Figure 2D). We had anticipated that DMPFC might show *greater* activity for the no-knowledge relative to high-knowledge (or intermediate-knowledge) targets, because participants have to rely more upon the neurocognitive resources associated with a generic theory of mind. However, present evidence does not yet support this conjecture.

Trait Judgment Task:

In the trait judgment task, participants rated the applicability of personality traits to each of the target characters. Insofar as participants would be more likely to recruit a person-specific theory of mind in assessing the personality traits of the high-knowledge target relative to the other targets, we anticipated that this condition would show elevated activity in mentalizing regions, especially MPFC. Indeed, when trait-judgment trials regarding the high-knowledge target were compared to trait-judgment trials regarding the other targets, greater activity was observed to the high-knowledge targets in a number of mentalizing regions, including DMPFC, MPFC (BA10), VMPFC (BA11), precuneus, and LTPJ. See Table 3 and Figure 3A for details and additional clusters. These result suggest that person-specific theories of mind may depend upon the recruitment of medial prefrontal in particular and mentalizing regions more generally, and may be operative even when participants have only learned trait information about a social target relatively recently.

While comparing high-knowledge, intermediate-knowledge, and low-knowledge targets is informative, we also sought to understand how trial-by-trial variation in retrospective

confidence and trait idiosyncrasy would predict activity in MPFC and other mentalizing regions. In order to characterize the neural correlates of trial-by-trial variation in these factors, we employed parametric modulation analysis. As described in the *Methods*, analyses of parametric modulators generally collapsed over distinctions between target characters. As reported above, participants showed greater confidence (on average) in their judgments of the high-knowledge targets than the intermediate-knowledge or no-knowledge targets, and showed greater confidence in their judgments of the intermediate-knowledge targets than the no-knowledge targets. A similar relationship (though somewhat weaker) held true for idiosyncrasy. Thus, to control for differences between target in our parametric modulation analyses would likely remove most of the variance associated with the parametric modulator. We therefore do not interpret our parametric modulation analyses as reflecting a separate phenomenon from the target contrasts reported above, but rather an alternative means of characterizing the neural correlates of person-specific mentalizing.

The confidence parametric modulator was positively associated with activity in several regions, including MPFC, DMPFC, and the left middle temporal gyrus (see Table 3 and Figure 3B). This association suggests that neural response in these regions may contribute to participants' sense that their judgments are justified in light of the relevant information available regarding the target characters. Contrary to our predictions, no regions demonstrated activity that was positively associated with the idiosyncrasy parametric modulator at FDR-corrected thresholds. While a person-specific theory of mind is one plausible process that might lead participants to judge a given target to be relatively idiosyncratic on a particular trait, it is certainly not the only such contributing factor. Category-based knowledge and stereotypes frequently lead to judgments that a target individual is different from the 'ordinary person', but

quite obviously do not involve the formation of individuated person-specific theories of mind. However, such category-based information may plausibly be most useful under circumstances in which detailed person-specific knowledge is unavailable or felt by the social thinker to be contextually inapplicable.

For this reason, we felt it prudent to investigate the interaction between confidence and idiosyncrasy, with the hypothesis that participants will be especially likely to have employed person-specific mentalizing processes on trials in which they *confidently* judged a target to be *different* from the ordinary person. We therefore conducted a third parametric modulation analysis, in which the effect of the cross-product term (confidence X idiosyncrasy) on trial-by-trial variation in hemodynamic response was estimated. This model also included confidence and idiosyncrasy parametric modulators separately, so that the variance predicted by the cross-product term would uniquely reflect the interaction between the two variables rather than their independent effects on activity. Controlling for these first-order parameters, the interaction term uniquely predicted activity in a MPFC cluster near the boundary between BA10 and BA9, as well as more dorsally (see Table 3 and Figure 3C). These results are consistent with trial-by-trial recruitment of person-specific mentalizing processes subserved by MPFC during judgments in which participants confidently attributed to the target relatively unique or idiosyncratic traits. Taken together with the target-based differences in recruitment of MPFC during the Why/How task described above, this outcome provides converging evidence that MPFC response is associated with participants' capacity to deploy person-specific theories of mind during mental state reasoning and trait attribution.

Discussion:

The present study provides novel neuroimaging evidence that exposure to rich, detailed personal information about social targets through video stimuli can evoke changes in hemodynamic activity within mentalizing regions during mental state reasoning and trait attribution tasks. These results are consistent with the person-specific theory of mind hypothesis advanced in Study 1, insofar as the greater recruitment of mentalizing regions when thinking about high-knowledge targets (when they are not perceived to be similar to the self) suggests that participants may have deployed person-specific information in making their judgments. First, participants showed elevated activity in mentalizing regions to high-knowledge targets across tasks, which indicates that such effects are not limited to the specific social/cognitive manipulation to which participants were subjected. Second, clusters within MPFC and other mentalizing regions were found to be positively associated with participants' willingness to confidently judge targets to have idiosyncratic traits. In the absence of person-specific information, participants rationally ought not to confidently attribute idiosyncratic traits to social targets (although, of course, they may still do so, as e.g. when exhibiting the fundamental attribution error).

The present study employed dramatic television serials because of their in-depth portrayal of multifaceted characters, which we hoped would facilitate the formation of person-specific theories of mind. However, the precise circumstances under which we learn best about the mental states and dispositional traits of others, and the underlying neural mechanisms that support such processes, have yet to be thoroughly explored. One recent study by Corradi-Dell'Acqua et al. (2015) introduced participants to novel social targets through photographs and brief textual descriptions, and subsequently found differences between dorsal and ventral medial

PFC based upon the content of participants' judgments. When assessing targets preferences, participants gradually increased their engagement of DMPFC over time, but found that a ventromedial PFC cluster became increasing responsive to the impact of rules on targets' behaviors. Interestingly, an MPFC cluster within BA10 was selectively active when participants had learned that a target was likely to *violate* norms or rules prescribed to him. While the others do not offer such an interpretation, it seems natural to consider this effect from the standpoint of person-specific mentalizing. Learning that a target will behave atypically might plausibly be associated with the formation of a person-specific theory of mind, rather than with the application of generic mentalizing processes. Another recent neuroimaging study investigated mental state reasoning after minimal exposure to novel social targets, and found above-chance accuracy at inferring others' preferences to be associated with activity in the DMPFC (Kang et al., 2015). This is entirely consistent with the notion that DMPFC is essential for applying a generic theory of mind in the absence of unique, individuating social information. It is certainly conceivable that participants in this study may very rapidly pick up on subtle features of the target that indicate his or her preferences, but these features are very unlikely to be unique or idiosyncratic attributes.

This second empirical investigation of person-specific theory of mind complements the first by elucidating the mechanisms that support person-specific mental state reasoning about novel, previously unfamiliar targets. Future work might profitably consider the relationship between person-specific theory of mind and other forms of social knowledge, including especially intermediate representations such as stereotypes. It would also be advantageous to know what particular features of a social target trigger the need to form a person-specific theory

of mind, and facilitate the encoding of information relevant to assessing others' mental states in an individuated manner.

Table 1

Descriptive behavioral data summarized for High-knowledge, Intermediate-knowledge, and no-knowledge targets:

	High-knowledge	Intermediate-knowledge	No-knowledge
Idiosyncrasy (1-9)	2.78	2.25	1.93
Confidence (1-9)	7.32	5.70	4.43
Similarity (1-9)	2.77	2.69	2.15

Table 2:

Summary of whole-brain analysis: Why/How Task

Test Effect/Anatomical Region	t	x	y	z	k
Why > How, All Targets:					
Left Calcarine Sulcus	5.1472	-9	-82	4	268
	9.5708	17	-77	4	268
Left Middle Temporal Gyrus	7.1361	-54	-10	-16	927
	4.7891	-63	-36	-2	927
	8.7278	-46	15	-31	927
DMPFC	8.1212	-3	49	27	527
Left TPJ	7.7299	-51	-66	34	679
Left Hippocampus	7.3003	-20	-21	-17	104
Right Middle Temporal Gyrus	6.9297	67	-20	-12	918
	7.2157	52	8	-23	918
Precuneus	7.0085	0	-57	36	1310
Left Calcarine Sulcus	6.9831	-7	-47	4	441
Cerebellum	6.3404	0	-38	-22	441
Left Insula	6.9554	-30	16	-11	95
MPFC	6.6398	10	70	5	112
Right Insula	6.2643	37	14	-13	94
Medial Obitofrontal Cortex	5.862	-1	57	-14	109
Left Superior Frontal Gyrus	5.7438	-29	63	8	146
Left Suprior Frontal Gyrus	5.7269	-23	46	40	100
Left Middle Frontal Gyrus	5.2549	-45	18	42	160
Right Superior Frontal Gyrus	5.2287	22	38	47	138
Right TPJ	5.1702	57	-55	26	141
Cerebellum	5.0045	0	-48	-38	115
Supplementary Motor Area	4.4452	-9	28	58	198
Right Middle Temporal Gyrus	4.222	65	-35	4	61
How > Why, All Targets:					
Left Precentral Gyrus	7.1005	-43	2	27	325
Left Supramarginal Gyrus	6.5289	-58	-27	38	466
Right Superior Parietal Lobule	5.7678	32	-53	68	121
Right Postcentral Gyrus	5.7523	56	-25	48	622
Left Inferior Temporal Gyrus	5.72	-48	-49	-10	140
Right Supramarginal Gyrus	5.6481	59	-35	29	69
Left Inferior Frontal Gyrus, pars triangularis	4.6565	-46	39	14	75
Left Precentral Gyrus	7.1005	-43	2	27	325
Why > How, High-knowledge target:					
Posterior Cingulate Cortex	9.6572	2	-51	32	1394

Right Middle Temporal Gyrus	8.537	63	-17	-8	1231
	6.1301	56	5	-21	1231
	5.8581	69	-37	7	1231
Left TPJ	8.2227	-52	-62	41	838
Left Middle Temporal Gyrus	6.8394	-60	-25	-11	1074
	6.3456	-47	17	-28	1074
Right Superior Frontal Gyrus	4.6847	10	42	58	201
	5.9794	20	49	38	201
Right TPJ	5.9063	56	-59	38	200
Left Gyrus Rectus	5.8263	-1	21	-23	89
DMPFC	5.6756	-7	52	22	143
Cerebellum	5.3979	26	-82	-28	76
Left Middle Frontal Gyrus	4.8367	-37	20	47	136
MPFC	4.6824	-8	58	-8	73
Cerebellum	4.2827	-23	-85	-32	139
Why > How, No-Knowledge target:					
Left Calcarine Gyrus	6.5931	-10	-68	17	1225
	6.0892	14	-66	16	1225
	5.4176	13	-31	-22	1225
	4.322	-1	-73	35	1225
	10.6293	4	-45	1	1225
Left Calcarine Gyrus	8.269	7	-96	-8	63
Middle Temporal Gyrus	8.1974	-53	-12	-13	113
Cerebellum	7.1379	21	-78	-23	74
Cerebellum	6.3891	48	-68	-36	67
Precuneus	6.0966	-5	-47	51	206
MPFC	6.0678	-4	68	0	142
	3.557	-25	65	-8	142
Right Lingual Gyrus	5.9794	5	-73	-7	134
Right Temporale Pole	5.3338	52	13	-30	84
Precuneus	5.3131	-1	-73	56	64
Left TPJ	5.2088	-50	-67	28	126
Left Middle Frontal Gyrus	4.8689	-23	23	46	109
Frontal_Sup_R	4.5836	26	34	51	93
DMPFC	4.4729	-1	36	54	155
	3.8212	-2	54	42	155

Table 3:

Summary of whole-brain analysis: Trait Judgment Task

Test Effect/Anatomical Region	t	x	y	z	k
High-knowledge > Other targets:					
Precuneus	16.0494	-2	-54	35	710
	8.5008	0	-54	15	710
MPFC/DMPFC	7.6116	-14	67	3	1254
	10.582	-11	45	44	1254
	8.3952	-14	64	28	1254
	9.6238	6	59	27	1254
VMPFC	15.3255	-7	50	-16	1254
	10.2805	0	31	-24	1254
Calcarine Sulcus	13.2276	3	-85	-14	212
	7.3486	-16	-99	-14	212
Cerebellum	12.7464	24	-78	-21	235
	9.5086	40	-64	-21	235
Left Inferior Frontal Gyrus, pars orbitalis	11.9656	-43	28	-17	685
	4.1877	-54	36	10	685
Left TPJ	11.1694	-50	-60	30	101
Cerebellum	10.6827	-26	-79	-31	147
Left Middle Temporal Gyrus	10.6348	-58	-5	-12	251
Left Inferior Frontal Gyrus, pars triagularis	10.3362	-55	22	18	202
Frontal_Sup_Medial_R	9.645	14	70	9	90
Right Temporal Pole	8.738	41	15	-19	145
Precuneus	7.9646	12	-41	3	61
Left Fusiform Gyrus	7.8389	-17	-39	-13	81
Right Superior Temporal Gyrus	7.5486	56	3	-13	210
Right Inferior Frontal Gyrus	6.7793	55	37	0	63
Calcarine Gyrus	5.8595	7	-74	15	63
Confidence Parametric Modulator:					
Middle Temporal Gyrus	8.1349	-62	-2	-12	79
Lingual Gyrus	7.6344	17	-60	0	232
Subgenual ACC	7.1492	3	24	-6	70
Left SFG/DMPFC	7.033	-16	53	39	129
Left Inferior Occipital Cortex	6.0525	-28	-97	-12	122
MPFC	5.9661	-7	68	14	70
Left Middle Frontal Gyrus	4.4951	-37	39	40	62

Left Postcentral Gyrus	4.4369	-64	-17	33	63
Left Supramarginal Gyrus	4.4364	-58	-25	18	118
Confidence x Idiosyncrasy Parametric Modulator:					
Cerebellum	9.228	41	-65	-35	69
Left Middle Temporal Gyrus	8.2073	-62	-13	-12	101
Right Inferior Parietal Lobule	7.8676	50	-49	47	138
Left TPJ	7.6851	-49	-61	27	91
Left Inferior Frontal Gyrus	7.596	-45	29	-18	67
Left superior Frontal Gyrus	7.3682	-22	30	48	61
MPFC	7.3536	3	59	21	105
Right Superior Frontal Gyrus	6.8694	31	58	17	124
Left Temporal Pole	6.438	-39	6	-36	130
Right Postcentral Gyrus	6.1765	56	-4	31	151
Left Middle Temporal Gyrus	5.6166	-63	-45	-6	269
Right Superior Frontal Gyrus	5.4919	25	62	1	62
DMPFC	5.3221	0	40	55	186
	5.2529	4	58	35	186
Left Superior Frontal Gyrus	4.5498	-31	64	6	85

Figure 1: A) Why > How contrast, collapsing over all targets. B) How > Why contrast, also collapsing across all targets. All results FDR $p<0.05$.

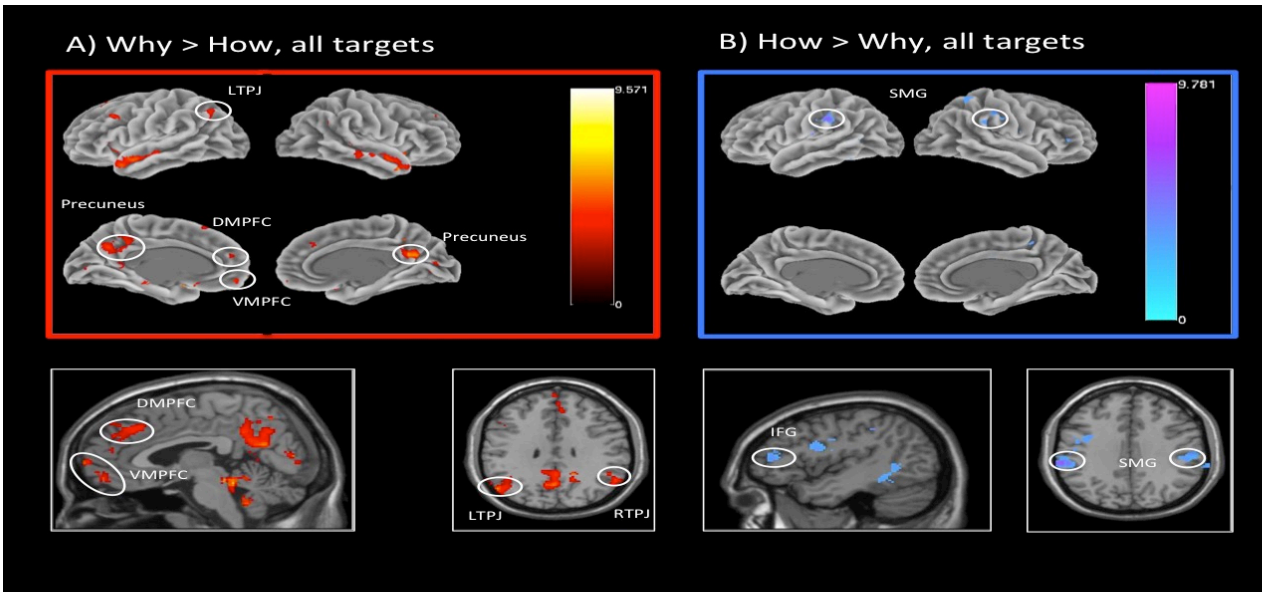


Figure 2: A) Why > How contrast depicted for the high-knowledge target. Parameter estimates for C) are extracted from the VMPFC cluster circled in red. B) Why > How contrast for no-knowledge target. Parameter estimates for D) are extracted from the DMPFC cluster circled in light blue.

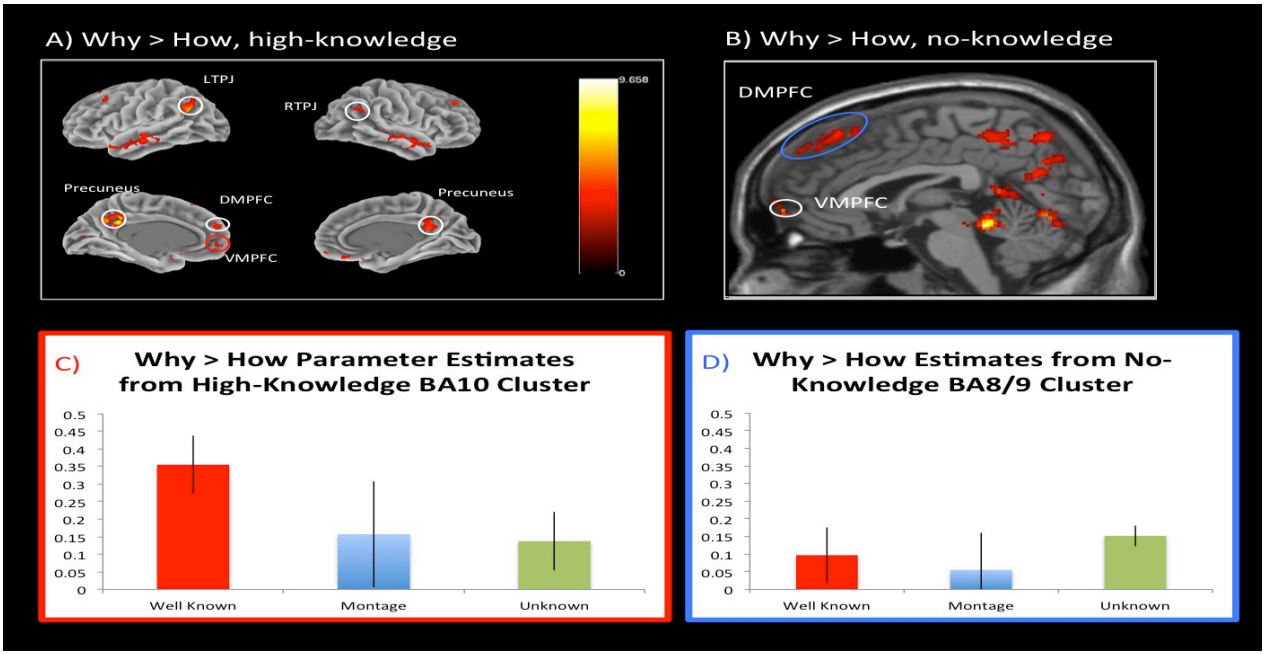
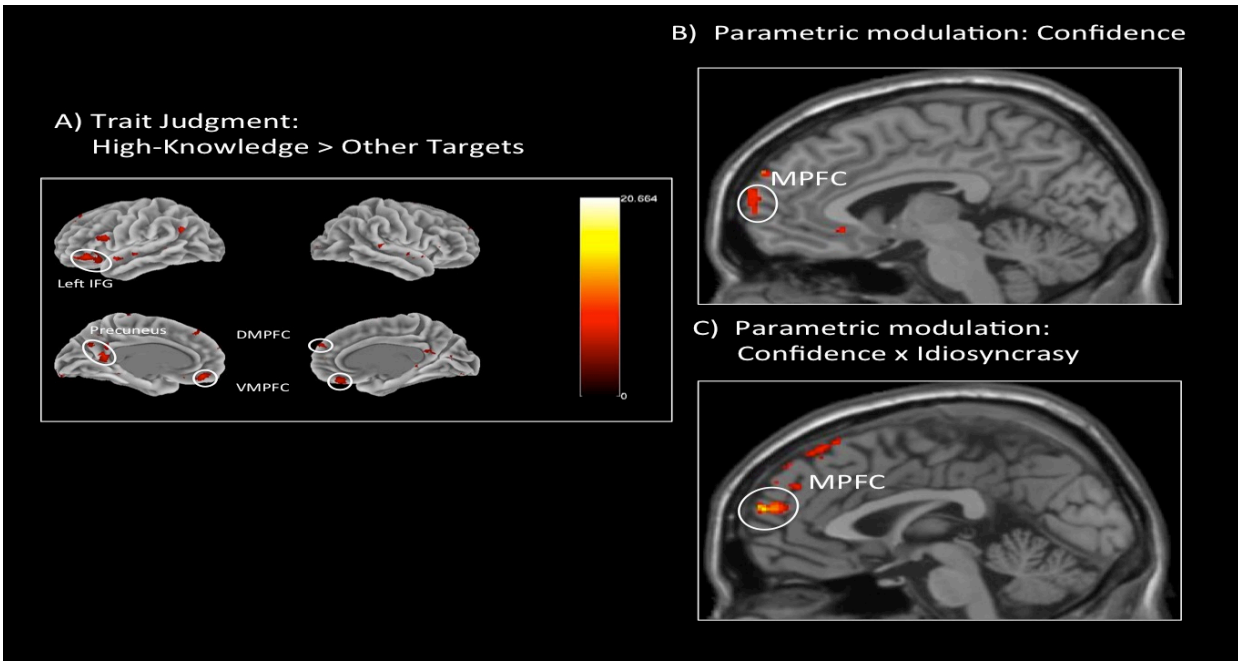


Figure 3: Principal results from the person-specific trait-judgment task. A) Comparison of high-knowledge trait-judgment trials to intermediate-knowledge and no-knowledge trials. B) Parametric modulation analysis with confidence parametric modulator. C) Parametric modulation analysis with Confidence x Idiosyncrasy cross-product parametric modulator, controlling for separate mean-centered confidence and idiosyncrasy parametric modulators.



References:

Brett, M., Anton, J-L., Valabregue, R., and Poline, J-B. (2002). Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in Neuroimage 16(2).

Corradi-Dell'Acqua, C., Turri, F., Kaufmann, L., Clement, F., Schwartz, S., (in press). How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation. *Cortex*.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162-173.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., Rizzolatti, G., (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3,* e79.

Kang, P., Lee, J., Sul, S., Kim, H., (2013). Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Frontiers in Human Neuroscience, 7,* 1-11.

Krienen, F.M., Tu, P-C., & Buckner, R.L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. J. Neurosci. 30(41), 13906-13915.

Kaul, C., Ratner, K.G., Van Bavel, J.J., (2014). Dynamic representations of race: processing goals shape race decoding in the fusiform gyri. *Social Cognitive and Affective Neuroscience, 9,* 326-332.

Spunt, R.P., Falk, E.B., Lieberman, M.D., (2010). Dissociable neural systems support retrieval of *How* and *Why* Action Knowledge. *Psychological Science, 21*, 1593-1598.

Spunt, R.P., Lieberman, M.D., (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage, 59,* 3050-3059.

General Conclusion:

The two studies presented for in this dissertation work have attempted to explain the explore the neural correlates of person-specific mentalizing. In Study 1 (Welborn & Lieberman 2015), MPFC demonstrated greater activation to high-knowledge than to low-knowledge targets, regardless of perceived similarity, connectedness, closeness or personal overlap. Hemodynamic response in the MPFC was associated both with independent raters assessments of participants' target knowledge, and the idiosyncrasy of participants trait attribution regarding the targets. In Study 2, participants were induced to form a person-specific theory of mind about a novel target, and mentalizing regions including MPFC exhibited great response to this 'high-knowledge' target than to others during both mental state reasoning (in the Why/How task) and evaluation of the target's dispositional traits (in the Trait Judgment task). Moreover, MPFC activity varied positively with participants' confidence in judging targets to be relatively idiosyncratic. Taken together, these two studies provide converging evidence that MPFC supports the deployment of person-specific theories of mind regarding well-known individuals, and provides a foundation for future exploration of the applicability and limitations of person-specific mental models.

References for Abstract, General Introduction, and Conclusion:

Frith, C.D., & Frith, U., (1999). Interacting Minds==A Biological Basis. *Science, 286,* 1692-1695.

Frith, U., and Frith, C.D., (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London, 358,* 459-473.

Gopnik, A., & Wellman, H.M. (1994). The theory theory. In *Mapping the Mind* (Hirschfeld, L. and Gelman, S., eds), pp. 257-293, Cambridge University Press.

Harris, P.L. (1992). From simulation to folk psychology: the case for development. *Mind & Language, 7*, 120-144.

Krienen, F.M., Tu, P-C., & Buckner, R.L., (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *The Journal of Neuroscience, 30, 41,* 13906-13915.

Leslie, A.M. Friedman, O., German, T.P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences, 8,* 528-533.

Lieberman, M.D., (2010). Social cognitive neuroscience. S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds). *Handbook of Social Psychology* (5th ed.) (pp. 143-193). New York, NY: McGraw-Hill.

Mitchell, J.P., Banaji, M.R., & Macrae, C.N., (2005). *The Journal of Cognitive Neuroscience, 17, 8,* 1306-1315.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R., (2006). Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron, 50,* 655-663.

Premack, D., & Woodruff, G., (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4*, 515-526.

Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19,* 65-72.

Van Overwalle, F. Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *NeuroImage, 48,* 564-584.

Van Overwalle, F., Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: dual processes driven by a shared neural network. *Frontiers of Human Neuroscience, 7,* 560.

Welborn, B.L., & Lieberman, M.D., (2015). Person-specific theory of mind in the medial pFC. *The Journal of Cognitive Neuroscience, 27,* 1-12.