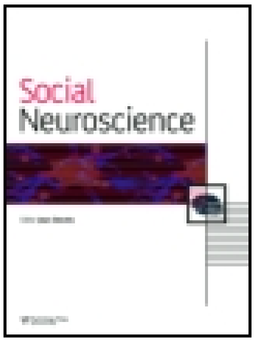

Social Neuroscience

ISSN: 1747-0919 (Print) 1747-0927 (Online) Journal homepage: http://www.tandfonline.com/loi/psns20

An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents

Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande & Frank Krueger

SOCIAL NEUROSCIENCE, 2016
http://dx.doi.org/10.1080/17470919.2016.1205131

Routledge
Taylor & Francis Group

An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents

Kimberly Goodyear[a], Raja Parasuraman[b], Sergey Chernyak[a], Ewart de Visser[b,c], Poornima Madhavan[d], Gopikrishna Deshpande[e,f,g] and Frank Krueger[b]

[a]Molecular Neuroscience Department, George Mason University, Fairfax, VA, USA; [b]Department of Psychology, George Mason University, Fairfax, VA, USA; [c]Human Factors and UX Research, Perceptronics Solutions, Inc., Falls Church, VA, USA; [d]Board on Human-Systems Integration, National Academies of Sciences, Engineering and Medicine, Washington, DC, USA; [e]Auburn University MRI Research Center, Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA; [f]Department of Psychology, Auburn University, Auburn, AL, USA; [g]Alabama Advanced Imaging Consortium, Auburn University and University of Alabama, Birmingham, AL, USA

ABSTRACT

As society becomes more reliant on machines and automation, understanding how people utilize advice is a necessary endeavor. Our objective was to reveal the underlying neural associations during advice utilization from expert human and machine agents with fMRI and multivariate Granger causality analysis. During an X-ray luggage-screening task, participants accepted or rejected good or bad advice from either the human or machine agent framed as experts with manipulated reliability (high miss rate). We showed that the machine-agent group decreased their advice utilization compared to the human-agent group and these differences in behaviors during advice utilization could be accounted for by high expectations of reliable advice and changes in attention allocation due to miss errors. Brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention, were recruited during the task and the advice utilization network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase. Our findings expand on the existing literature by showing that misses degrade advice utilization, which is represented in a neural network involving salience detection and self-processing with perceptual integration.

ARTICLE HISTORY
Received 24 February 2016
Revised 16 June 2016
Published online 15 July 2016

KEYWORDS
Expert advice; functional magnetic resonance imaging (fMRI); effective connectivity; Granger causality; errors

Introduction

People are often given numerous options regarding the type and source of advice they can receive. For example, when individuals travel to a new country, they can ask a native citizen or use a smartphone with a Global Positioning System (GPS) for directions. Given the different options available, it is becoming a necessity to understand how individuals utilize or discount advice from different sources. Factors such as source credibility (expert and novice) (Madhavan & Wiegmann, 2007; Van Swol & Sniezek, 2005) and initial expectations of reliable advice (Dzindolet, Pierce, Beck, & Dawe, 2002) can influence how someone responds to advice. Dzindolet et al. (2002) proposed that individuals may possess a "perfect automation schema," which is an expectation that automation performs near perfectly and can ultimately cause a person to disuse the advice given to them when errors occur. Initial expectations of reliable advice can be impacted, however, when disconfirmation evidence of misleading advice is encountered.

To fully understand the influence of bad advice on decision-making behaviors requires an examination of error types: false alarms and misses. The type of error is of particular interest because, while a false alarm error is misleading, it is not necessarily harmful. In contrast, a miss error can lead to disastrous results such as a luggage-screener failing to detect a bomb in a suitcase. Previous evidence has shown that false alarms can cause a "cry wolf effect," in which an individual may tend to ignore true alerts (Breznitz, 2013) and misses may affect monitoring strategies leading to an adaptation of attention allocation (Onnasch, Ruff, & Manzey, 2014). False alarms have been shown to decrease trust and decrease reliance and compliance, while misses have been shown to only decrease reliance (Dixon, Wickens, & McCarley, 2007; Rice & McCarley, 2011). Furthermore, studies comparing humans and machines

CONTACT Frank Krueger FKrueger@gmu.edu Department of Psychology, George Mason University, 4400 University Drive, Mail Stop 2A1, Fairfax, VA 22030, USA

Research Conducted: George Mason University & Auburn University.
Supplemental data for this article can be accessed here.

have shown that expert humans were trusted more than expert machines due to differences in dispositional credibility (Madhavan & Wiegmann, 2007), and allocation of tasks to humans compared to automation can be affected by trust in automation (Lewandowsky, Mundy, & Tan, 2000). To expand on the existing literature on humans and machines, we previously investigated the impact of false alarms on decision-making behaviors (Goodyear et al., 2016), and to demonstrate that there are behavioral and neural differences between false alarms and misses, the current study examined misses.

The neural processes involved with advice taking have been recently investigated with functional magnetic resonance imaging (fMRI) advice-taking paradigms examining expert advice (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Meshi, Biele, Korn, & Heekeren, 2012) and during adaptive learning (Biele, Rieskamp, Krugel, & Heekeren, 2011). Furthermore, neuroimaging studies examining interactions between humans and robots during perspective taking (Krach et al., 2008) and during social observations (Wang & Quadflieg, 2015) have also been investigated. The default-mode network (cognitive functions such as self-processing) involving areas like the temporoparietal junction, precuneus and the salience network (detection of internal and external events) with areas such as the dorsal anterior cingulate cortex, insulae have been additionally implicated in other advice-taking tasks (Engelmann, Capra, Noussair, & Berns, 2009), as well as during robot-human interaction paradigms (Chaminade et al., 2012).

However, in spite of the existing literature on advice taking, the neural basis and underlying brain networks associated with miss errors from expert human and machines remains to be elucidated.

We implemented an X-ray luggage-screening task with fMRI combined with multivariate Granger causality analysis (GCA) to investigate the impact of misses on decision-making behaviors and to reveal the underlying brain network associated with advice utilization from unreliable agents framed as experts. GCA provides a useful tool to infer causal relationships by investigating directional interactions between activated brain regions. Based upon previous studies investigating misses and false alarms (Dzindolet et al., 2002; McBride, Rogers, & Fisk, 2014), we expected advice utilization to decrease due to the catastrophic consequences of a miss error and also due to disconfirmation evidence about the agents' expertise (i.e., high miss rate) provided by feedback. We further expected that the reevaluation of the agents' perceived expertise/credibility to cause a mismatch of perceptions due to bad advice, which would ultimately cause an adjustment in attention allocation strategies during advice utilization. In addition, based upon previous work investigating advice acceptance and trust between expert human and machine agents (Madhavan & Wiegmann, 2007), we expected participants interacting with the machine agent to have a greater depreciation of advice utilization compared to the human agent due to perceptions involved with the perfect automation schema and varying degrees of perceived dispositional credibility. Lastly, brain regions involved with self-processing (e.g., precuneus) and error monitoring and salience detection (e.g., anterior cingulate cortex) would be recruited when comparing the human agent and the machine agent due to deviations in expectations (agents framed as experts), resulting from a change in attention strategies from a high miss rate.

Methods

Participants

A normative rating study was conducted at George Mason University (GMU) and an fMRI study was conducted at Auburn University (AU). All studies were conducted according to the ethical guidelines and principles of the Declaration of Helsinki. For the normative rating study, 23 male students (age ($M \pm SD$) = 24.0 ± 2.6) participated to standardize the X-ray luggage images for the experimental studies based on clutter, general difficulty and confidence in finding the target. For the fMRI study, 24 healthy right-handed volunteers (14 males, 10 females; age = 22.3 ± 2.4) participated in the X-ray luggage-screening task while receiving advice. Participants gave written consent approved by the Institutional Review Boards at GMU and AU and they received financial compensation for their participation (see Goodyear et al., 2016, for details on methods).

X-ray luggage-screening task

Participants partook in an X-ray luggage-screening task and were asked to search for the presence or absence of a knife (Madhavan & Gonzalez, 2006) (Figure S1(b)). Participants were randomly assigned to either the human-agent group or the machine-agent group with 60% reliability (40% misses) and they received good (advice-congruent: hits, correct rejections) and bad (advice-incongruent: misses) advice (Figure S1(a)). The reliability for the agents was chosen based on earlier work demonstrating differences between false alarms and misses (Dixon et al., 2007).

The jitter times were generated by an fMRI simulator software (http://www.mccauslandcenter.sc.edu/CRNL/tools/fmrisim) and consisted of a minimum of 1 second and an average of 4 seconds to optimize timing. Participants responded by using fiber optic response pads (Current Designs, http://www.curdes.com/); they were given an initial endowment of $40 and each incorrect answer resulted in a deduction of $0.30 from the remaining total. Advice utilization, response times and monetary deductions were collected during the experiment. The stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Inc.).

Procedure

Pre-experimental phase

Participants completed self-report questionnaires as control measures to investigate individual differences approximately 1–2 weeks before the fMRI experiment. The control measures included: Interpersonal Reactivity Index (IRI) (Davis, 1983), Complacency-Potential Rating Scale (CPS) (Singh, Molloy, & Parasuraman, 1997), National Readiness Technology Scale (NTRS) (Parasuraman, 2000), NEO Five-Factor Inventory (NEO-FFI) (Costa & McCrae, 1992) and Propensity to Trust (PTT) (Merritt, Heimbaugh, LaChapell, & Lee, 2013).

Experimental phase

Participants completed a practice run where they read descriptions about the human or machine agent (reliability was not disclosed), rated their trust in and reliability of the human or machine agent on a 10-point Likert scale (0 = very low, 10 = very high), familiarized themselves with the five possible knives that could be present in the bags and then completed four practice trials of the task. The participants then completed two experimental runs of the task while in the scanner and rated reliability and trust afterwards.

Post-experimental session

After completion of the fMRI experiment, participants were asked to rate their confidence in finding the target (i.e., knife) in each of the images presented during the experiment on a 10-point Likert scale (1 = very low, 10 = very high).

Neuroimaging acquisition

Imaging data were acquired on a 7 T actively shielded whole-body scanner (Siemens Magnetom) with a 32-channel head coil (Nova Medical) at AU MRI Research Center, Auburn, Alabama. The anatomical imaging data were based on a 3D T1-weighted MPRAGE sequence with TR = 2020 ms, TE = 2.7 ms, flip angle = 7°, slice thickness = 1.2 mm, voxel dimension = 1.1 mm × 1.1 mm × 1.2 mm and number of slices = 240. The functional imaging data were based on a 2D gradient-echo multiband EPI sequence with TR = 1000 ms, TE = 20 ms, flip angle = 70°, slice thickness = 2 mm, voxel dimensions = 2.1 mm × 2.1 mm × 2.0 mm, number of slices = 45 per volume in an axial orientation parallel to the anterior–posterior commissure and a multiband factor of 2. The first two volumes were discarded to allow for T1 equilibrium effects and a total of 660 volumes were taken for each run.

Behavioral data analysis

Behavioral data were analyzed with the statistical package for the Social Sciences 20.0 (SPSS 20.0, IBM Corp.) and the alpha was set to $p < .05$ (two-tailed). Data were normally distributed (Kolmogorov–Smirnov test) and the assumption of homogeneity of variance (Bartlett's test) was not violated. Mixed 2 × 2 × 2 repeated-measures ANOVAs with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor were employed to examine advice utilization, response times and monetary deductions. In addition, we investigated reliability, trust and confidence ratings with mixed 2 × 2 repeated-measures ANOVAs with Agent (human, machine) as the between-subjects factor. The within-subjects factor for the reliability/trust ratings were Time (pre, post) while for confidence ratings, it was Target (yes, no).

Neuroimaging data analysis

The fMRI data were analyzed through NeuroElf software (http://neuroelf.net) and BrainVoyager QX 2.8 (Brain Innovation). The functional imaging data were preprocessed using Statistical Parametric Mapping (SPM, Wellcome Department of Cognitive Neurology) functions batched via NeuroElf, including three-dimensional motion correction (six parameters), slice-scan time correction (temporal interpolation). A mean functional image was computed for each participant across all runs and was then co-registered with the anatomical images using a joint-histogram for the different contrast types. Preprocessing procedures for the anatomical images included segmenting images with a unified segmentation procedure (Ashburner & Friston, 2005) and the functional images had spatial warping applied to them to normalize the data to a standard Montreal Neurological Institute (MNI) brain template. To account for any residual differences across participants, spatial smoothing (Gaussian filter of 6 mm FWHM) was applied to the images.

A general linear model (GLM) that was corrected for first-order serial correlations fit to the data (Friston, Harrison, & Penny, 2003), which consisted of 37 regressors based on advice utilization (accept, reject), advice type (good, bad), time (run 1, run 2) for each of the five phases (fixation, advice, bag, decision, feedback) and seven parametric regressors of no interest for the global signal and 3D motion correction (translations in X, Y, Z directions, rotations around X, Y, Z axes). The regressor time courses were adjusted for the hemodynamic response delay by convolution with a dual-gamma canonical hemodynamic response function (Büchel, Holmes, Rees, & Friston, 1998). Random-effect analyses were performed at the multi-subject level to explore brain activations associated with the decision and feedback phases during advice utilization.

Mixed $2 \times 2 \times 2$ ANOVAs on parameter estimates were applied with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor. Brain activations for the decision and feedback phases were reported after correcting for multiple comparisons using a cluster-level statistical threshold (cluster-level statistical threshold estimator plugin in BrainVoyager QX). The thresholded map ($p < .005$) was used for a whole-brain correction criterion, which is based off an estimate of the map's spatial smoothness and on a Monte Carlo simulation (1000 iterations). The minimum cluster size at a specified confidence level ($\alpha = 0.05$) was then calculated (Forman et al., 1995; Goebel, Esposito, & Formisano, 2006). The significant activation clusters were displayed in MNI space on an anatomical brain template reversed left to right (i.e., radiological convention).

Effective connectivity analysis

Granger causality is based on a concept of causality that can be used to predict directional influences among chosen brain regions through mulitvariate effective connectivity modeling of ROI (region of interest) time courses (Deshpande, LaConte, James, Peltier, & Hu, 2009; Friston et al., 2003; Granger, 1969; Preusse, van der Meer, Deshpande, Krueger, & Wartenburger, 2011). The model examines the relationship of variables in time, such that given two variables, *a* and *b*, if past values of *a* better predict the present value of *b*, then as a function of earlier time points, causality between the variables can be inferred (Goodyear et al., 2016; Hampstead et al., 2011; Krueger, Landgraf, van der Meer, Deshpande, & Hu, 2011; Roebroeck, Formisano, & Goebel, 2005). Granger causality analysis is a data-driven approach and thus is advantageous for application of effective connectivity since there is no requirement for pre-specified connectivity models like dynamic causal modeling (Deshpande & Hu, 2012; Deshpande et al., 2009; Deshpande, Sathian, & Hu, 2010).

Effective (or directional) connectivity data were analyzed using a code developed in-house using MATLAB (www.mathworks.com) (Grant et al., 2014; Lacey, Stilla, Sreenivasan, Deshpande, & Sathian, 2014). The effective connectivity in the network of activated regions was performed through multivariate Granger causality analysis (GCA) and only regions that survived the fMRI analysis threshold for the main effect of Agent (human, machine) for the decision and feedback phases were selected as ROIs. Time series of the blood-oxygen-level-dependent (BOLD) signal for the selected ROIs were extracted around peak activation maxima (sphere of $6 \times 6 \times 6$ mm^3), averaged across voxels and normalized across participants, per run. Blind hemodynamic deconvolution of the mean ROI BOLD time series was performed using a Cubature Kalman filter and smoother (Havlicek, Friston, Jan, Brazdil, & Calhoun, 2011) and the resulting latent neural signals were entered into a first order dynamic multivariate autoregressive (dMVAR) model to assess directed interactions of multiple nodes as a function of time (Feng et al., 2016; Grant, Wood, Sreenivasan, Wheelock, & White, 2015; Hampstead, Khoshnoodi, Yan, Deshpande, & Sathian, 2016; Hutcheson et al., 2015; Wheelock et al., 2014).

Granger connectivity path weights for the condition of interest (advice utilization) for each agent (human, machine) were extracted, populated into two samples, and independent samples *t*-tests were employed (q (FDR) $< .05$) (Benjamini & Hochberg, 1995) to reveal significantly different effective connectivity paths between the agent groups (Figure S2). Effective connectivity of brain regions (i.e., nodes, edges) was displayed on a brain surface using BrainNet Viewer, a graphical interface visualization tool (Xia, Wang, & He, 2013).

Results

Behavioral results

For *advice utilization*, the mixed ANOVAs demonstrated a significant main effects of Agent ($F(1, 22) = 5.24$, $p = .032$), Advice ($F(1, 22) = 140.72$, $p < .0001$) and Time ($F(1, 22) = 22.36$, $p < .0001$). These results indicate that participants accepted advice more from the human agent compared to the machine agent. Furthermore, good advice was accepted more than bad advice and

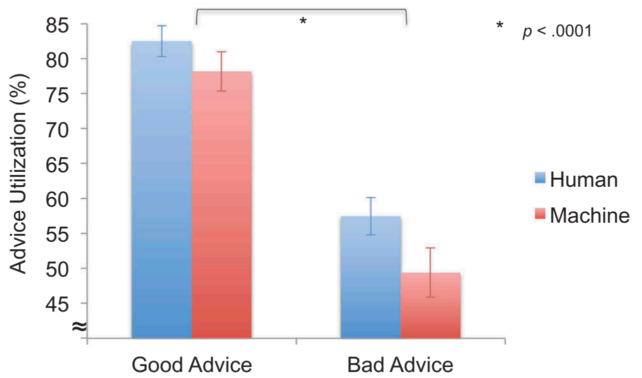

Figure 1. Behavioral results for decision phase ($M \pm SEM$). Advice utilization was significantly lower for bad advice compared to good advice and was also significantly lower for the machine-agent group compared to the human-agent group.

advice utilization decreased overtime (Figure 1). In addition, a significant two-way interaction of Advice × Time was identified ($F(1, 22) = 10.17$, $p = .004$), but no significant two-way interaction effects of Advice × Agent ($F(1, 22) = 0.69$, $p = .415$), Time × Agent ($F(1, 22) = 0.46$, $p = .505$), or three-way interaction of Advice × Time × Agent ($F(1, 22) = 1.40$, $p = .249$) were found.

In addition, we looked at pre- and post-reliability/trust ratings. One participant's data were not used due to lack of understanding, which was indicated by the high values for all pre/post scales. The *reliability ratings* showed no significant main effect of Agent ($F(1, 21) = 0.76$, $p = .394$), but a significant main effect of Time ($F(1, 21) = 5.43$, $p = .030$), showing that reliability ratings decreased from pre- to post-experiment (Figure 2(a)). No significant interaction effect of Time × Agent ($F(1, 21) = 0.00$, $p = .960$) was found. Furthermore, one-sample *t*-tests on perceived versus actual reliability (60%) of the agent revealed that pre-reliability ratings were significantly higher than the actual reliability for the human agent ($t(11) = 4.53$ $p = .001$) and the machine agent ($t(10) = 3.55$ $p = .005$). For *trust ratings*, no significant main effect of Agent ($F(1, 21) = 0.01$, $p = .905$) was found, but a significant main effect of Time ($F(1, 21) = 8.18$, $p = .009$) was observed, showing that trust ratings significantly decreased from pre- to post-experiment (Figure 2(b)). No significant interaction effect of Time × Agent ($F(1, 21) = 0.00$, $p = .960$) was demonstrated.

Lastly, we analyzed differences in control measures (e.g., demographic measures and questionnaires) with independent samples *t*-tests. No significant group differences were identified for any of the control measures (Table S1). Note that statistical analyses on response times, monetary deductions and confidence ratings are provided in the Supplementary Material section.

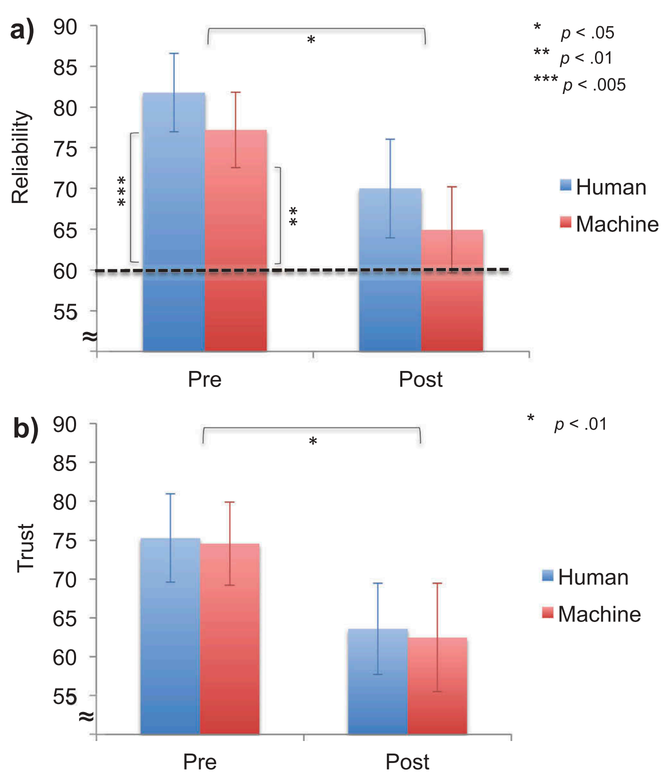

Figure 2. Results for ratings ($M \pm SEM$). (a) Pre- and post-reliability. For both groups, the perceived pre-reliability was significantly higher than the actual reliability of the agent (60%) and post-reliability ratings significantly decreased. (b) Pre- and post-trust. Post-trust was significantly lower than pre-trust for both groups.

Neuroimaging results

We investigated brain activations during the decision and feedback phases with mixed ANOVAs. For the decision phase, a significant main effect of Agent ($\alpha < .05$, $k = 11$) was found in the right (R) lingual gyrus (LG) (BA 18), R anterior cingulate cortex (ACC) (BA 24), left (L) anterior precuneus (aPreC) (superior parietal lobule; BA 7) and L cuneus (CUN) (BA 18) (Figure 3, Table 1). Note that to avoid circularity, or double dipping, no further statistical analyses were performed for the decision and feedback phases (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). A main effect of Advice ($\alpha < .05$, $k = 11$) was found in the R middle frontal gyrus (BA 8), R medial frontal gyrus (BA 8), R rostrolateral prefrontal cortex (rlPFC) (superior frontal gyrus; BA 10), R primary visual cortex (V1) (BA 17), R pre-supplementary motor area (pre-SMA) (superior frontal gyrus; BA 6), L cerebellar culmen, L inferior occipital gyrus (IOG) (BA 18).

For the feedback phase, a main effect of Agent ($\alpha < .05$, $k = 10$) was found in the R precentral gyrus

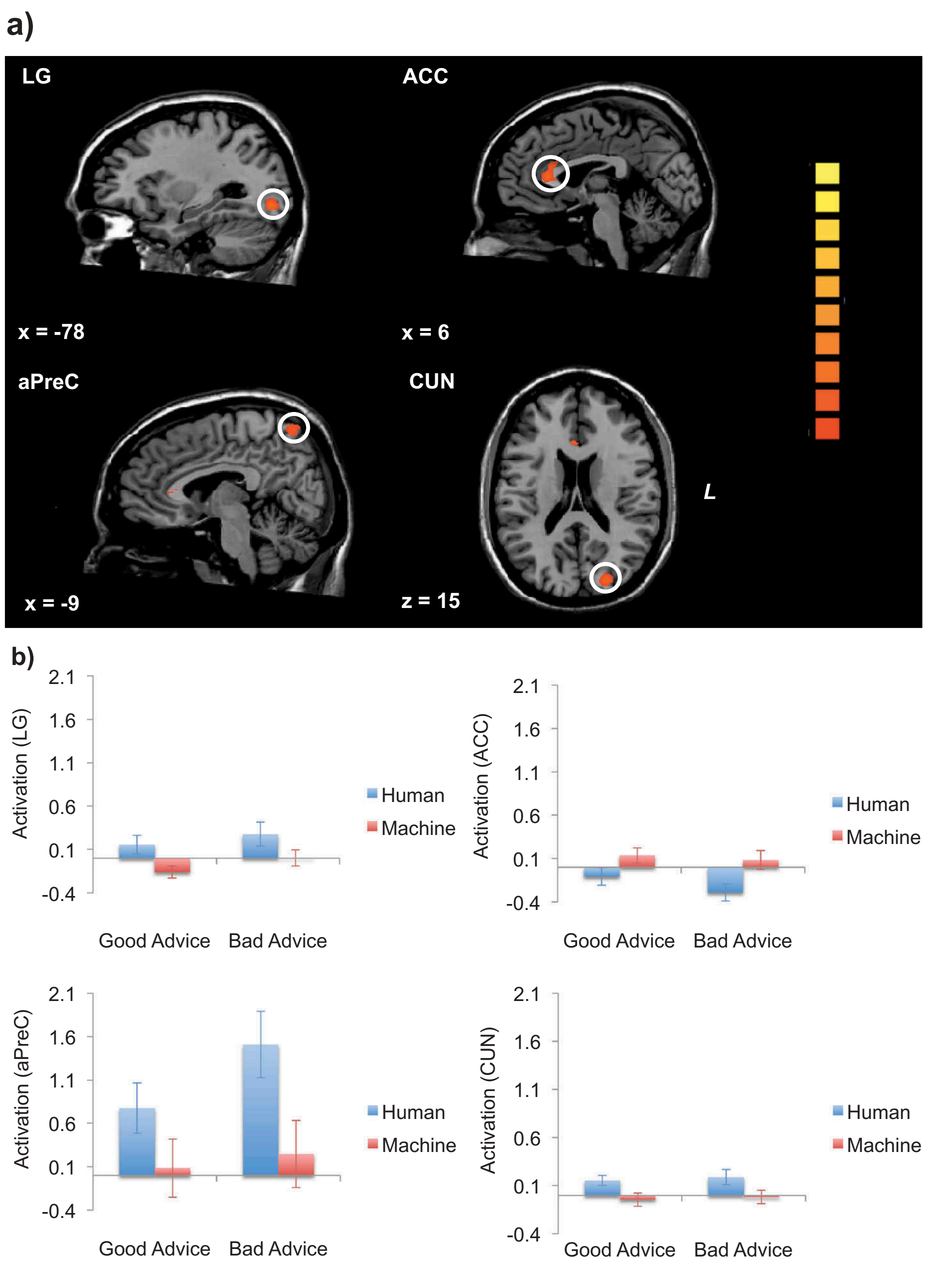

Figure 3. Brain activations during decision phase ($\alpha < .05$, $k = 11$). (a) The main effect of Agent during the decision phase significantly activated the right lingual gyrus (LG), right anterior cingulate cortex (ACC), left anterior precuneus (aPreC) and left cuneus (CUN). (b) The activation pattern indicates higher activation for the human-agent compared with machine-agent group for all regions except the ACC. The bar plots shown are for visualization purposes.

Table 1. Brain regions associated with the agent and advice main effects. Brain regions showing significant activation clusters associated during the decision phase: Agent (minimum cluster of 11) and Advice (minimum cluster of 11); and feedback phase: Agent (minimum cluster of 10) and Advice (minimum cluster of 9) (α < .05, cluster-level threshold corrected, MNI space).

	F value	Cluster size (mm^3)	x	y	z
Decision phase					
Agent					
Right lingual gyrus	18.44	629	27	−78	−6
Right anterior cingulate cortex	24.35	1246	6	27	12
Left anterior precuneus	19.84	727	−9	−63	57
Left cuneus	19.95	758	−21	−84	15
Advice					
Right middle frontal gyrus	24.75	822	42	18	42
Right medial frontal gyrus	21.30	3182	21	27	33
Right rostrolateral prefrontal cortex	28.51	560	24	54	6
Right primary visual cortex	19.72	1722	15	−96	−3
Right pre-supplementary motor area	19.86	665	6	9	56
Left cerebellar culmen	17.93	601	−12	−36	−24
Left inferior occipital gyrus	16.37	1936	−24	−90	−6
Feedback phase					
Agent					
Right precentral gyrus	16.66	456	51	−6	6
Right inferior parietal lobule	15.00	398	48	−26	24
Right cuneus	15.37	422	24	−84	15
Left putamen	19.30	1445	−27	−15	6
Left fusiform gyrus	19.58	990	−42	−47	−21
Advice					
Right postcentral gyrus	19.33	960	42	−18	27
Right middle frontal gyrus	16.78	631	33	21	39
Right hippocampus	18.19	1347	29	−39	3
Right extra-nuclear	18.66	468	24	21	15
Right orbitofronal cortex	25.94	892	21	45	−3
Right posterior cingulate cortex	31.47	1049	12	−63	23
Right anterior precuneus	23.25	1865	6	−69	47
Left cerebellar culmen	23.43	2945	−6	−42	−21
Left pons	16.91	373	3	21	51
Left pre-supplementary motor area	18.27	644	−18	−24	−30
Left parahippocampal gyrus	29.02	1102	−24	−42	0
Left postcentral gyrus	31.04	1300	−42	−21	27

(PrG) (BA 6), R inferior parietal lobule (IPL) (BA 40), R CUN (BA 17), L putamen (Pu) and L fusiform gyrus (FG) (BA 37) (Figure 4, Table 1). Lastly, a significant main effect of Advice (α < .05, k = 9) during the feedback phase was found in the R postcentral gyrus (PoG) (BA 3), R middle frontal gyrus (BA 8), R hippocampus, R extranuclear, R orbitofrontal cortex (OFC) (BA 10/11), R posterior cingulate cortex (PCC) (BA 31), R aPreC (BA 7), L cerebellar culmen, L pre-SMA (BA 6/8), L pons, L parahippocampal gyrus (BA 19) and L PoG (BA 2).

Effective connectivity results

To identify effective connectivity among brain regions when comparing the human to the machine agents during the decision and feedback phases, we implemented multivariate GCA based upon our results from the fMRI analysis (q(FDR) < .05). The LG was identified as the source ROI for the advice utilization network for the decision phase, that sent output connections to all target ROIs (ACC, aPreC, CUN) and the FG was the source ROI for the feedback phase sending an output connection to the IPL (Figure 5, Table 2).

Discussion

The objective of this research was to expand on our earlier work investigating the behavioral and neural signatures of advice utilization differences between expert human and machine agents during good and bad advice (Goodyear et al., 2016). We manipulated agent reliability with a high miss rate to reveal the underlying neural basis (in terms of both activated brain regions and the directional interactions between them) involved with advice utilization. We revealed that advice utilization decreased more for the machine-agent group compared to the human-agent group, coinciding with another study investigating the effects of source credibility with varying reliability from humans and machines (Madhavan & Wiegmann, 2007).

As hypothesized, our results demonstrated that advice utilization decreased more for the machine-agent group compared to the human-agent group. The degradation of advice utilization occurred regardless of the type of the advice (good, bad) given, showing that disconfirmation experience during bad advice had an effect on all decision-making behaviors. In our earlier work, we showed that false alarms caused a degradation of advice utilization during bad advice (Goodyear et al., 2016), but for our current study, we expected that misses would cause an overall adjustment in attention allocation due to previous evidence showing that more critical types of events (misses) lead to an adaptation in monitoring strategies (Onnasch et al., 2014). Our results indicated that advice utilization decreased for both groups, which provides evidence that participants made changes in their decision-making behaviors to compensate for the unreliable advice that they received.

In addition, we compared the pre-reliability ratings with the actual reliability of each agent to uncover any preconceived notions that participants had about the human and machine agents. We demonstrated that for both groups the pre-reliability ratings were significantly higher than the actual reliability, which could indicate that participants had higher initial expectations of reliable advice since the agents were framed as experts. In addition, reliability ratings decreased overall from pre- to post-experiment, showing that participants were

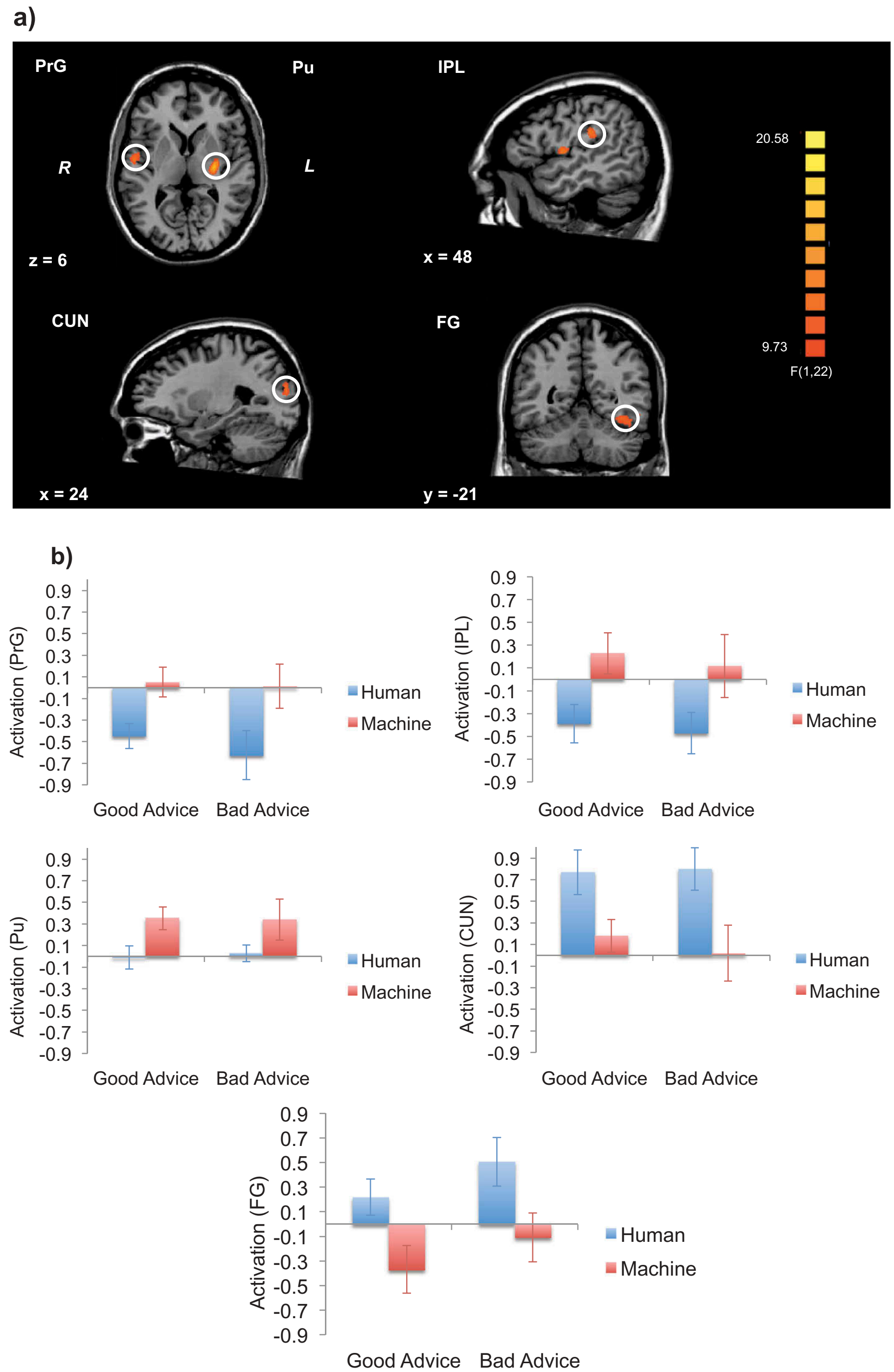

Figure 4. Brain activations during feedback phase ($\alpha < .05$, $k = 10$). (a) The main effect of Agent during the feedback phase significantly activated the right precentral gyrus (PrG), right inferior parietal lobule (IPL), R cuneus (CUN), left putamen (Pu) and left fusiform gyrus (FG). (b) The activation pattern shows higher activation for the machine-agent group compared to the human-agent group for all regions except for FG and CUN. The bar plots shown are for visualization purposes.

able to decipher the performance of the agents, while also recalibrating their expectations due to bad advice. Furthermore, we revealed that trust decreased overall from pre- to post-experiment, revealing that misses degraded trust, which has previously been reported for false alarms (Dixon et al., 2007; Goodyear et al., 2016; Rice & McCarley, 2011). Although the reliability and trust ratings did not significantly decrease more for the machine-agent group, the ratings were still lower compared to the human-agent group, which demonstrates that as trust and reliability decrease, advice utilization may degrade as well. Lastly, since we showed

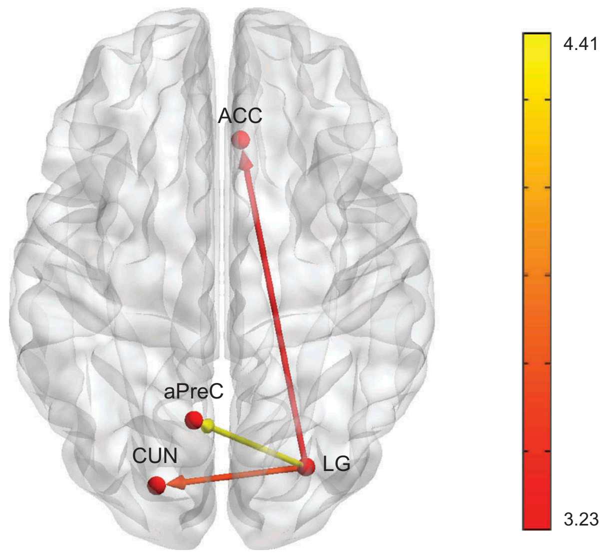

Figure 5. Results for multivariate granger causality analysis. The effective connectivity network for advice utilization during the decision phase when comparing the human with machine agent showed that the LG (lingual gyrus) was the driver of the network and source ROI, sending outputs to all target ROIs (ACC (anterior cingulate cortex), aPreC (anterior precuneus), and CUN (cuneus)) (all connections survived q(FDR) < .05). The color bar represents the t-value of the comparisons shown in Table 2.

Table 2. Path weights for granger causality analysis. The path weights displayed show significant effective connectivity paths that are stronger in the human-agent group compared to the machine-agent group during advice utilization (q(FDR) < .05). The directionality of the connectivity is shown in the first two columns, with the source column showing the ROIs that predict activation in the target column ROIs. The strength of connectivity is given by the mean path weights in the third column. LG, lingual gyrus; ACC, anterior cingulate cortex; aPreC, anterior precuneus; CUN, cuneus; FG, fusiform gyrus; IPL, inferior parietal lobule.

Source	Target	Path weight Human machine	t value	p value
Decision phase				
LG	ACC	0.087–0.003	3.23	6.18×10^{-6}
	aPreC	0.115–0.009	4.41	5.23×10^{-8}
	CUN	0.094–0.006	3.49	2.43×10^{-8}
Feedback phase				
FG	IPL	0.087–0.156	3.03	1.20×10^{-4}

no differences for control measures or confidence ratings between the agent groups, our results cannot be explained by those findings.

We next identified the neural associations and the underlying directional brain network differentially involved with advice utilization between humans and machines. For the decision phase, our effective connectivity network revealed the LG as the driver, or source ROI, of the network, sending outputs to the ACC, aPreC and CUN. Furthermore, the strength of the paths emanating from LG was significantly higher for human advice compared to machine advice. The results may indicate that the LG modulated attention during advice utilization through the bottom-up sensory processing of task-relevant information, but future studies are needed to confirm this hypothesis. It has been postulated that sensory processing involves a large-scale integration of networks with attention modulation to form a behavioral outcome, or a cognition (Mesulam, 1998). For example, it has been shown that detection of stimulus information initially starts in primary sensory areas, and is then conveyed to regions such as the ACC, showing the interaction between bottom-up and top-down processing during attentional control (Crottaz-Herbette & Menon, 2006). Furthermore, a study investigating advisor competence showed increased activity in the visual cortex during advice integration from incompetent advisors (Schilbach, Eickhoff, Schultze, Mojzisch, & Vogeley, 2013). The authors conclude that the activity in the visual cortices may relate to "perceptually based strategies" during reassessment of one's own judgments, which could support our findings about the influence of visual regions on upstream structures such as PreC and ACC during advice utilization with unreliable human advisors. Moreover, the involvement of the visual areas during the decision phase could be attributed to the fact that participants had to revisualize the X-ray images in order to compare what they saw to the advice they received.

Furthermore, our neuroimaging results for the decision phase revealed brain regions associated with attentional control and salience detection (ACC), self-processing (aPreC) and sensory information processing (LG and CUN). LG activation has been associated with comparing advice versus no advice in expert and peer groups (Suen, Brown, Morck, & Silverstone, 2014) and activity in the LG and CUN has been implicated during decisions under risk when comparing a message to accept or reject advice with no message (Engelmann et al., 2009) and during decisions correlated with value or saliency (Litt, Plassmann, Shiv, & Rangel, 2011). ACC activation has been shown to be involved with conflict monitoring during decision-making (Botvinick, 2007) and error detection and prediction error (Beckmann, Johansen-Berg, & Rushworth, 2009), while the PreC has been identified to play a role in integrations of one's mental state (Terasawa, Fukushima, & Umeda, 2013). Our neuroimaging results demonstrated that all areas except for the ACC had higher activations for the human-agent group compared to the machine-agent group, indicating that participants in the human-agent group may have had a greater increase in perceptual processing and perceivably less monitoring of errors.

Conversely, participants in the machine-agent group were more attuned to the advice errors, which were also indicated behaviorally, which could explain the ACC activation differences.

In addition to the decision phase, we expected a behavioral adjustment in advice utilization due to feedback. For the feedback phase, our effective connectivity network showed that the FG was the driver of the network that sent an output to the IPL. The FG has been associated with receipt of monetary rewards and penalties during an outcome phase (Dillon et al., 2008), while the IPL has been identified to play a role during advice evaluation when interacting with competent and incompetent advisors (Schilbach et al., 2013) and during decision uncertainty when given trial-by-trial feedback (Vickery & Jiang, 2009). Furthermore, the neuroimaging results for the feedback phase revealed activity in the PrG, CUN and Pu. Activity in the PrG has been implicated during comparisons of humans and computers during rock-paper-scissors games (Chaminade et al., 2012) and CUN activity has been shown to be related to inferential errors during a feedback phase (Cooper, Kreps, Wiebe, Pirkl, & Knutson, 2010). Lastly, we revealed activity in the dorsal striatum (Pu), which has been implicated in stimulus-response learning (Packard & Knowlton, 2002) and during responses to affective feedback in regards to valence and magnitude (Delgado, Locke, Stenger, & Fiez, 2003). Our results for the feedback phase illustrate that, for all regions except for CUN and FG, activations were higher for the machine-agent group compared to the human-agent group. This pattern of activation may indicate that as participants in the machine-agent group became more aware of the errors in advice, brain regions such as the Pu and IPL were recruited during feedback of the advice given by the agent.

There are a couple of limitations that need to be considered with the interpretation of our results. First, we looked at differences between good and bad advice with only misses as the type of error. However, our previous research on false alarms (Goodyear et al., 2016) provided substantiation for expanding on the effects of advice utilization with different error types and future studies could include both types of errors to compare the two directly. In addition, participants received advice before they made their decisions in order to prevent cognitive anchoring, or the tendency to rely on the first piece of information acquired. Future studies could investigate the effects of cognitive anchoring by implementing a task where participants receive advice after they make their decisions.

In conclusion, our results have shown that advice utilization differs between humans and machines and those distinctions are contingent on miss errors. Our findings expand on the existing literature by showing that misses degrade advice utilization, which is represented in a neural network involving salience detection and self-processing with perceptual integration. As our society progresses in technological terms, having a greater conceptualization of how decision-making processes differ during interactions with humans and machines can provide pertinent information about the involvement of cognitive processes. A better understanding of the behavioral and neural mechanisms involved with those interactions can ultimately allow for development of safety measures to prevent any mishaps that can occur during advice taking.

Acknowledgments

We would like to thank Paul Anglin for assisting with data collection. We would like to acknowledge the memory of our dear colleague and friend, Raja Parasuraman, who was a valuable contributor to this research.

Disclosure statement

K.G., R.P., S.C., E.D., P.M., G.D. and F.K. declared that they had no conflicts of interest with respect to their authorship or the publication of this article and no financial interest or benefit arising from the direct applications of this research.

Funding

This work was supported by the Air Force Office of Scientific Research (AFOSR) [Grant Number: 202857].

References

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839–851. doi:10.1016/j.neuroimage.2005.02.018

Beckmann, M., Johansen-Berg, H., & Rushworth, M. F. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. *Journal of Neuroscience*, *29*(4), 1175–1190. doi:10.1523/JNEUROSCI.3328-08.2009

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. *PLoS Biology*, *9*(6), e1001089. doi:10.1371/journal.pbio.1001089

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–1571. doi:10.1016/j.neuron.2013.10.024

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366. doi:10.3758/CABN.7.4.356

Breznitz, S. (2013). *Cry wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Büchel, C., Holmes, A. P., Rees, G., & Friston, K. J. (1998). Characterizing stimulus–response functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage*, *8*, 140–148. doi:10.1006/nimg.1998.0351

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*. doi:10.3389/fnhum.2012.00103

Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T., & Knutson, B. (2010). When giving is good: Ventromedial prefrontal cortex activation for others' intentions. *Neuron*, *67*(3), 511–521. doi:10.1016/j.neuron.2010.06.030

Costa, P., & McCrae, R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Crottaz-Herbette, S., & Menon, V. (2006). Where and when the anterior cingulate cortex modulates attentional response: Combined fMRI and ERP evidence. *Journal of Cognitive Neuroscience*, *18*(5), 766–780. doi:10.1162/jocn.2006.18.5.766

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. doi:10.1037/0022-3514.44.1.113

Delgado, M., Locke, H., Stenger, V., & Fiez, J. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(1), 27–38. doi:10.3758/CABN.3.1.27

Deshpande, G., & Hu, X. (2012). Investigating effective brain connectivity from fMRI data: Past findings and current issues with reference to Granger causality analysis. *Brain Connectivity*, *2*(5), 235–245. doi:10.1089/brain.2012.0091

Deshpande, G., LaConte, S., James, G. A., Peltier, S., & Hu, X. (2009). Multivariate Granger causality analysis of fMRI data. *Human Brain Mapping*, *30*(4), 1361–1373. doi:10.1002/hbm.v30:4

Deshpande, G., Sathian, K., & Hu, X. (2010). Effect of hemodynamic variability on Granger causality analysis of fMRI. *Neuroimage*, *52*(3), 884–896. doi:10.1016/j.neuroimage.2009.11.060

Dillon, D. G., Holmes, A. J., Jahn, A. L., Bogdan, R., Wald, L. L., & Pizzagalli, D. A. (2008). Dissociation of neural regions associated with anticipatory versus consummatory phases of incentive processing. *Psychophysiology*, *45*(1), 36–49.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 564–572. doi:10.1518/001872007X215656

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94. doi:10.1518/0018720024494856

Engelmann, J. B., Capra, C. M., Noussair, C., & Berns, G. S. (2009). Expert financial advice neurobiologically "offloads" financial decision-making under risk. *PLoS One*, *4*(3), e4957. doi:10.1371/journal.pone.0004957

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y.-J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647. doi:10.1002/(ISSN)1522-2594

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273–1302. doi:10.1016/S1053-8119(03)00202-7

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, *27*(5), 392–401. doi:10.1002/(ISSN)1097-0193

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). *Advice taking from humans and machines: An fMRI and effective connectivity study*. Manuscript submitted for publication.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, *37*, 424–438. doi:10.2307/1912791

Grant, M. M., White, D., Hadley, J., Hutcheson, N., Shelton, R., Sreenivasan, K., & Deshpande, G. (2014). Early life trauma and directional brain connectivity within major depression. *Human Brain Mapping*, *35*(9), 4815–4826. doi:10.1002/hbm.22514

Grant, M. M., Wood, K., Sreenivasan, K., Wheelock, M., & White, D. (2015). Influence of early life stress on intra- and extra-amygdaloid causal connectivity. *Neuropsychopharmacology*, *40*(7), 1782–1793.

Hampstead, B., Khoshnoodi, M., Yan, W., Deshpande, G., & Sathian, K. (2016). Patterns of effective connectivity during memory encoding and retrieval differ between patients with mild cognitive impairment and healthy older adults. *Neuroimage*, *124*, 997–1008. doi:10.1016/j.neuroimage.2015.10.002

Hampstead, B. M., Stringer, A. Y., Stilla, R. F., Deshpande, G., Hu, X., Moore, A. B., & Sathian, K. (2011). Activation and effective connectivity changes following explicit-memory training for face-name pairs in patients with mild cognitive impairment: A pilot study. *Neurorehabilitation and Neural Repair*, *25*(3), 210–222. doi:10.1177/1545968310382424

Havlicek, M., Friston, K. J., Jan, J., Brazdil, M., & Calhoun, V. D. (2011). Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *Neuroimage*, *56*(4), 2109–2128. doi:10.1016/j.neuroimage.2011.03.005

Hutcheson, N. L., Sreenivasan, K. R., Deshpande, G., Reid, M. A., Hadley, J., White, D. M., … Lahti, A. C. (2015). Effective connectivity during episodic memory retrieval in schizophrenia participants before and after antipsychotic medication. *Human Brain Mapping*, *36*(4), 1442–1457. doi:10.1002/hbm.v36.4

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS One*, *3*(7), e2597. doi:10.1371/journal.pone.0002597

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. doi:10.1038/nn.2303

Krueger, F., Landgraf, S., van der Meer, E., Deshpande, G., & Hu, X. (2011). Effective connectivity of the multiplication network: A functional MRI and multivariate Granger causality mapping study. *Human Brain Mapping*, *32*(9), 1419–1431. doi:10.1002/hbm.v32.9

Lacey, S., Stilla, R., Sreenivasan, K., Deshpande, G., & Sathian, K. (2014). Spatial imagery in haptic shape perception. *Neuropsychologia*, *60*, 144–158. doi:10.1016/j.neuropsychologia.2014.05.008

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*(2), 104.

Litt, A., Plassmann, H., Shiv, B., & Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. *Cerebral Cortex*, *21*(1), 95–102. doi:10.1093/cercor/bhq065

Madhavan, P., & Gonzalez, C. (2006). *Effects of sensitivity, criterion shifts, and subjective confidence on the development of automaticity in airline luggage screening*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting. doi:10.1177/154193120605000326

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 773–785. doi:10.1518/001872007X230154

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science*, *15*(6), 545–577. doi:10.1080/1463922X.2013.817625

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: effects of implicit attitudes toward automation on trust in an automated system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(3), 520–534. doi:10.1177/0018720812465081

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PLoS One*, *7*(11), e49748. doi:10.1371/journal.pone.0049748

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*(6), 1013–1052. doi:10.1093/brain/121.6.1013

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies*, *72*(10–11), 772–782. doi:10.1016/j.ijhcs.2014.05.001

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, *25*(1), 563–593. doi:10.1146/annurev.neuro.25.112701.142937

Parasuraman, A. (2000). Technology readiness index (Tri): A multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, *2*(4), 307–320. doi:10.1177/109467050024001

Preusse, F., van der Meer, E., Deshpande, G., Krueger, F., & Wartenburger, I. (2011). Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. *Frontiers in Human Neuroscience*, *5*, 22. doi:10.3389/fnhum.2011.00022

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*(4), 320–331.

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, *25*(1), 230–242. doi:10.1016/j.neuroimage.2004.11.017

Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., & Vogeley, K. (2013). To you I am listening: Perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Social Neuroscience*, *8*(3), 189–202. doi:10.1080/17470919.2013.775967

Singh, I. L., Molloy, R., & Parasuraman, R. (1997). Automation-induced "complacency": Developement of the complancency-potential rating scale. *The International Journal of Aviation Psychology*, *3*(2), 111–122. doi:10.1207/s15327108ijap0302_2

Suen, V. Y. M., Brown, M. R. G., Morck, R. K., & Silverstone, P. H. (2014). Regional brain changes occurring during disobedience to "experts" in financial decision-making. *PLoS One*, *9*(1), e87321. doi:10.1371/journal.pone.0087321

Terasawa, Y., Fukushima, H., & Umeda, S. (2013). How does interoceptive awareness interact with the subjective experience of emotion? An fMRI study. *Human Brain Mapping*, *34*(3), 598–612.

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*(3), 443–461. doi:10.1348/014466604X17092

Vickery, T. J., & Jiang, Y. V. (2009). Inferior parietal lobule supports decision making under uncertainty in humans. *Cerebral Cortex*, *19*(4), 916–925. doi:10.1093/cercor/bhn140

Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Social Cognitive and Affective Neuroscience*, *10*, 1515–1524. doi:10.1093/scan/nsv043

Wheelock, M. D., Sreenivasan, K. R., Wood, K. H., Ver Hoef, L. W., Deshpande, G., & Knight, D. C. (2014). Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage*, *102*(Pt 2), 904–912. doi:10.1016/j.neuroimage.2014.08.005

Xia, M., Wang, J., & He, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS One*, *8*(7), e68910. doi:10.1371/journal.pone.0068910