Behavioral/Systems/Cognitive

Human Insula Activation Reflects Risk Prediction Errors As Well As Risk

Kerstin Preuschoff,[1,4] Steven R. Quartz,[1,2] and Peter Bossaerts[1,2,3]

[1]Computation and Neural Systems Program and [2]Division of Humanities and Social Sciences, Caltech, Pasadena, California 91125, [3]Ecole Polytechnique Fédérale Lausanne, CH-1015 Lausanne, Switzerland, and [4]Institute for Empirical Research in Economics, University of Zurich, CH-8006 Zurich, Switzerland

Understanding how organisms deal with probabilistic stimulus-reward associations has been advanced by a convergence between reinforcement learning models and primate physiology, which demonstrated that the brain encodes a reward prediction error signal. However, organisms must also predict the level of risk associated with reward forecasts, monitor the errors in those risk predictions, and update these in light of new information. Risk prediction serves a dual purpose: (1) to guide choice in risk-sensitive organisms and (2) to modulate learning of uncertain rewards. To date, it is not known whether or how the brain accomplishes risk prediction. Using functional imaging during a simple gambling task in which we constantly changed risk, we show that an early-onset activation in the human insula correlates significantly with risk prediction error and that its time course is consistent with a role in rapid updating. Additionally, we show that activation previously associated with general uncertainty emerges with a delay consistent with a role in risk prediction. The activations correlating with risk prediction and risk prediction errors are the analogy for risk of activations correlating with reward prediction and reward prediction errors for reward expectation. As such, our findings indicate that our understanding of the neural basis of reward anticipation under uncertainty needs to be expanded to include risk prediction.

Key words: **risk prediction; insula; risk; uncertainty; reinforcement learning; reward prediction**

Introduction

In the context of uncertain rewards, our understanding of expected reward (average reward anticipated) in the brain has been advanced considerably by quantitative models of reward processing (Sutton, 1988; Montague et al., 1996). Reinforcement learning models suggested that reward-related processing required two signals: a reward prediction signal and a reward prediction error signal. The resulting quantitative framework has proven crucial for understanding reward processing in the dopaminergic system, both in terms of action selection and for learning uncertain reward distributions.

One can envisage an analogous quantitative framework for estimating risk, where errors in risk prediction are used to update future estimates of risk. Here, we use this framework to investigate risk processing in the brain. Specifically, the framework predicts the existence of two risk-related signals, risk prediction and risk prediction error. Informally, risk prediction is the risk that is associated with an uncertain outcome and is measured as reward variance (or its square root, SD). If the risk prediction is misjudged, errors arise, referred to as risk prediction errors, which may be used to improve future estimates of risk prediction.

There exist many reward learning approaches that accommodate risk indirectly. Examples include nonlinear transformation of rewards [the expected utility approach (Koenig and Simmons, 1994)], max/min policies [robust control (Heger, 1994)], and sign-based weighting of prediction errors [effectively inducing loss aversion (Mihatsch and Neuneier, 2002)]. Here, we consider direct tracking of the risk of predicting rewards, as in Kalman filtering, but allowing risk to change stochastically, or to be unknown [as in the study by Stroud and Bengtsson (2006)]. The generalized autoregressive conditional heteroscedasticity model of Engle (1982, 2002) is a canonical example. There, changes in risk are driven by squared (reward) prediction errors. This model has proven very successful for tracking risk in financial economics.

A framework with direct risk prediction has not been applied widely yet. However, direct risk prediction has been shown to be reflected in the bodily states of professional financial traders (Lo and Repin, 2002). In addition, recent behavioral evidence shows that human subjects adjust their learning rate to changing risk, implying that they must somehow be tracking risk (Behrens et al., 2007; Preuschoff and Bossaerts, 2007). But the neurobiological foundations of such risk processing remain unknown.

We hypothesized a system in the brain that encodes both risk prediction and risk prediction errors, in analogy with the dopamine system, which encodes both reward prediction and reward prediction errors. We predicted that this system may be insula, as

Received Sept. 19, 2007; revised Jan. 11, 2008; accepted Jan. 14, 2008.

This work was supported by National Science Foundation Grant 0093757, the David and Lucile Packard Foundation, the Gordon and Betty Moore Foundation, and the Swiss Finance Institute. We thank members of the Social Cognitive Neuroscience Laboratory for helpful comments. We thank Steven Flaherty for technical assistance.

Correspondence should be addressed to Kerstin Preuschoff, Institute for Empirical Research in Economics, Bluemlisalpstrasse 10, CH-8006 Zurich, Switzerland.

DOI:10.1523/JNEUROSCI.4286-07.2008

 0270-6474/08/282745-08$15.00/0

insula activation correlates with a broad range of risk-related characteristics of gambles involving probabilistic rewards, such as complexity (Huettel et al., 2005; Grinband et al., 2006), ambiguity (Hsu et al., 2005; Huettel et al., 2006), and uncertainty (Elliott et al., 2000; Critchley et al., 2001; Ernst et al., 2002; Paulus et al., 2003). Despite its general involvement in encoding uncertainty, the precise role of the insula remains unresolved because of the lack of a quantitative framework for investigating risk signals in the brain. We provide such a framework.

Nineteen subjects played a simple card game (see Fig. 1*A*) while their brain activity was recorded using functional magnetic resonance imaging (fMRI) (Preuschoff et al., 2006). Throughout the experiment, we manipulated risk and, hence, induced risk prediction errors. We found activations in insula that correlated with risk prediction and, separately, with risk prediction errors.

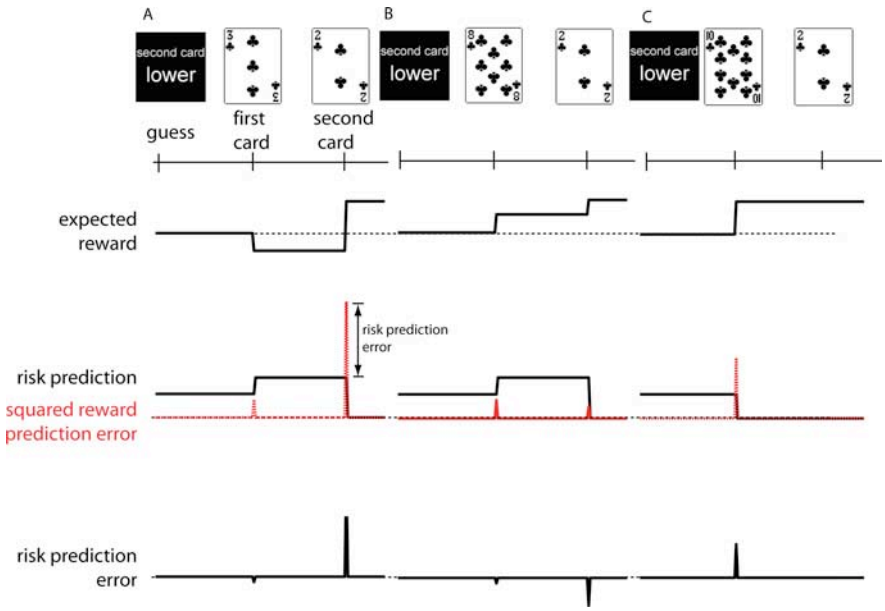

Figure 1. ***A***, Time line for a single trial. Within a trial, predictions and prediction errors arise twice: before and when the first card is revealed and before and when the second card is revealed. In this trial, the subject guesses that the second card will be lower. The first card drawn is a 3. The second card is a 2. Hence, the subject wins \$1.00. The solid black line in the second panel tracks the risk prediction, which is measured here as expected variance (its square root is the SD). The actual size-squared of the reward prediction error (red dashed line) is a little smaller than expected; therefore, the risk prediction error (third panel) is negative. At the second card, the reward prediction error is much larger than expected. ***B***, The subject guesses that the second card will be lower. The first card drawn is an 8. The second card is a 2. Hence, the subject wins \$1.00. The size-squared of the reward prediction error at the first card is smaller than expected; therefore, the risk prediction error is negative. At the second card, the reward prediction error is smaller than expected. Therefore, the risk prediction error is negative. ***C***, The subject guesses that the second card will be lower. The first card drawn is a 10. The second card is a 2; hence, the subject wins \$1.00. The size-squared of the reward prediction error at the first card is much larger than expected; therefore, the risk prediction error is positive and large. Because no uncertainty remains about the outcome after card 1, the risk prediction is zero, and there are no prediction errors at the time of card 2. Note that the risk prediction associated with the first card is the same across all trials because no information is available about the first card. The risk prediction of the second card is a function of the first card and therefore varies across trials.

Materials and Methods

A total of 19 subjects participated in the study (10 male, 9 female; 18–30 years of age; mean age, 21.4 years). All participants gave full informed consent. The study was approved by the California Institute of Technology Institutional Review Board.

Experimental paradigm. In each trial, two cards were drawn consecutively from a randomly shuffled deck of 10. Before seeing either card, players guessed whether the second card would be higher or lower than the first. Subjects made one dollar if they were right; they lost one dollar otherwise. We then displayed the first card followed ~7 s later by the second card. To ensure that subjects paid attention, we then asked subjects to confirm whether they won or lost.

Within each trial, predictions occur twice: once before the first card, and again before the second card. Both these predictions generate corresponding prediction errors, once when the first card is revealed, and again when the second card is revealed. This is illustrated for three exemplary trials in Figure 1. At the beginning of a trial, the player has an estimate of the average number on the first card, and thus of the average expected reward after seeing the first card. Likewise, the player has an estimate of the variability of the number on the first card and, hence, a prediction of the risk of forecasting the reward to be expected based on the number on the first card. Once the first card is revealed, the average expected reward and the risk prediction are compared with the actual values (the actual expected reward after display of the first card, and the deviation of actual expected reward from its a priori average). These comparisons result in a reward prediction error and in a risk prediction error. The player then estimates the reward revealed through the second card and the corresponding prediction risk, which again results in errors once the second card is revealed. See Figure 2, Appendix, and supplemental material (available at www.jneurosci.org as supplemental material) for formal definitions and qualitative illustration of all variables as a function of the probability of winning after seeing card 1.

Subjects were given written instructions for the game and completed a brief training session outside the magnet. Before each session, subjects were given an initial endowment of \$25.00. One dollar was at stake in each trial. Failure to place a bet resulted in an automatic loss. Subjects also lost \$0.25 if they failed to report or incorrectly reported the outcome of their bet at the end of each trial. Accumulated gains were shown only at the end of each session. A total of three sessions with 30 trials per session were played by each subject. During scanning, trials were randomly ordered. At the end of the experiment, subjects selected one of the three sessions at random, which determined their final payoff.

fMRI acquisition. Each scanning session started with a localizer scan and T1-weighted anatomical scans (256 × 256 matrix; 176 1-mm sagittal slices). While subjects performed the gambling task, functional images were acquired using a Siemens TRIO 3.0T full-body MRI scanner using T2*-weighted PACE EPI (repetition time, 2000 ms; echo time, 30 ms; 64 × 64; 3.28125 × 3.28125 mm^2; 32 3.0-mm slices; no gap; field of view, 210). For each subject, three functional runs were collected (392–400 scans each).

Data processing and analysis. Data were processed and analyzed using BrainVoyager v1.26. Preprocessing included motion correction (six-parameter rigid body transformation), slice timing correction, linear drift removal, high-pass filtering, normalization to Talairach space, and spatial smoothing with a full width at half maximum Gaussian kernel of 8 mm. For each subject, a separate linear model was constructed that included regressors for reward prediction, risk prediction, and their respective errors as described below as well as for wins and losses and visual and motor activation. Each regressor modeled the blood oxygen level-dependent response to the specified events by applying a convolution kernel to a boxcar function. Temporal autocorrelations were corrected using a first-order autoregression. For each subject, contrasts were calculated at every voxel in the brain. In a random-effects analysis, a one-sample *t* test determined where the average contrast value for the group as a whole ($n = 19$ subjects) differed significantly from zero. A signifi-

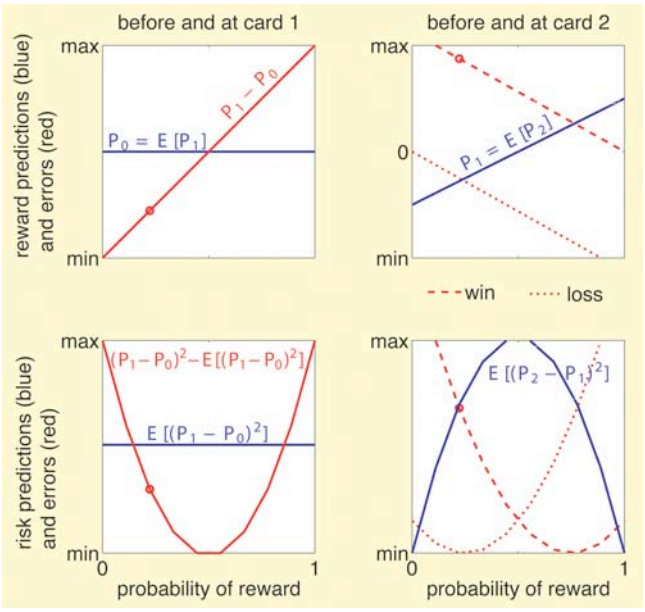

Figure 2. Reward predictions, risk predictions, and corresponding errors as a function of the probability of winning after display of the first card. Top row, Reward predictions (blue) and reward prediction errors (red) before and at the first (left) and second (right) card. Bottom row, Risk prediction (blue) and risk prediction errors (red) before and at the first (left) and second (right) card. Left column, Before the first card is seen, both reward prediction (top) and risk prediction (bottom) are constant across all trials independent of the probability of winning after the first card. The error at the first card is a function of the first card or probability of winning. The reward prediction error is linear and the risk prediction error is quadratic (U-shaped) in the probability of winning. Right column, Before the second card is seen, the reward prediction is linear in the probability of winning, whereas the risk prediction is quadratic (inversely U-shaped). Both errors are a function of the probability of winning as well as the outcome (dashed line, win; dotted line, loss). The red point in each plot depicts the error in the numerical example of Figure 1 *A*. See also Appendix.

cance ($p < 0.0005$) and cluster size (≥5 voxels) threshold was applied to all statistical maps.

To identify the regions of interests (ROIs), we proceeded as follows. We divided the period between display of cards 1 and 2 into two consecutive epochs, a short epoch (1 s from card 1) followed by a long epoch (6 s) modeling the remainder of the period. Both epochs were modeled with three predictors: a constant, a linear function of the probability of reward, and a quadratic function of the probability of reward; these predictors are referred to as 0th-, first-, and second-order predictors. The first- and second-order predictors were meant to capture expected reward (which increases linearly in probability of reward) (see Fig. 2 and Appendix) and risk prediction or risk prediction errors (which change nonlinearly with probability of reward) (see Fig. 2 and Appendix). Note that, in theory, the first- and second-order predictors are orthogonal (reward probability and squared reward probability are orthogonal), but minor correlation was present in our setting because of the finite number of samples and hemodynamic response function smoothing. However, this correlation will be fully accounted for when we present time courses (see below). Activation after display of card 2 was modeled using a constant term, terms indicating a win and a loss, respectively, and, to capture activation potentially related to risk prediction errors, a term that changed quadratically in the (previous) probability of winning (see Fig. 2 and Appendix). All variables after display of card 2 are block variables with duration of 1 s. All models (including the ROI models described below) included regressors to account for possible confounds including visual and motor activation, instruction screen, final score, initial endowment screen, and responses given at the answer screen.

For the second step of our analysis, we used as regions of interest the clusters in bilateral anterior insula, the activity of which after card 1 correlated significantly with the second-order term of the above model (i.e., with the square of reward probability). Two pairs of (bilateral) regions were obtained, one for each epoch (short, 1 s epoch and longer epoch covering the remainder of the period between cards 1 and 2). We emphasize that the regions of interest were obtained only from activations after card 1 and before card 2. However, had we also obtained regions of interest for the short epoch after card 2, they would not have been much different from those in the corresponding epoch after card 1) (Fig. S2, available at www.jneurosci.org as supplemental material).

We obtained new β values for the regions activating in the short epoch after card 1 by taking the original general linear model and replacing the 0th-, first-, and second-order predictors by one regressor for each level of probability of winning. We could thus verify that the activations encoded risk prediction errors by comparing the activation patterns (estimated βs as a function of probability of reward) with the expected patterns (see Fig. 2 and Appendix). If these regions indeed encode risk prediction errors, they should also activate after display of card 2, when another risk prediction error emerges (Fig. 1). To verify this, we likewise replaced the 0th-, first-, and second-order predictors in the 1 s epoch after display of card 2 with regressors for each level of probability and checked whether the activation patterns as a function of probability were as predicted and whether their magnitudes were comparable with those from the 1 s epoch after card 1. This is a true out-of-sample test, because we verify the patterns and magnitudes after card 2 in regions of interest that were obtained only from activations after card 1. We summarize the findings with respect to risk prediction error by plotting the β estimates against the risk prediction error (see Fig. 3*B*). β Values for probabilities of winning that correspond to the same risk prediction error are pooled. We also obtained new β estimates for the regions of interest in insula for the second, longer epoch in the interval between cards 1 and 2. One new β estimate was computed for each level of reward probability. The resulting patterns suggested that the activations encoded risk prediction and, hence, corresponded to the activations reported elsewhere in the literature, as we explain below.

Subsequently, adjusted time courses for the regions of interest were computed (see Fig. 4 and Figs. S1, S2, available at www.jneurosci.org as supplemental material). These are time courses corrected for the effects (confounds) of all predictors except the second-order predictors. They were obtained as errors of a reduced model that included all regressors except for second-order predictors. Any effect not included in the reduced model shows up in the error term, although any effect included in the reduced model will not show up. For activations related to risk prediction, time courses were grouped into high (corresponding to probabilities of reward p between 0.3 and 0.7), medium ($0 < p < 0.3$ or $0.7 < p < 1$) or low (or no risk: $p = 0$ or $p = 1$) anticipated risk, and time-locked to card 1. For the activations reflecting risk prediction errors at display of card 1, time courses were grouped into high (corresponding to $p = 0$ or $p = 1$), medium ($0 < p < 0.3$ or $0.7 < p < 1$), and low ($0.3 < p < 0.7$) risk prediction errors, and time-locked to card 1. Adjusted time courses of activations at display of card 2 and correlating with risk prediction errors were grouped into low (<0), medium (between 0 and 1), and high (>1).

Results

Activation in anterior insula correlates with risk prediction

Between placing the bet and seeing the first card, risk prediction is constant across all trials and subjects. This reflects the fact that the information about the outcome of the gamble at the time of bet does not change across trials. However, between the first and second card, risk prediction depends on the first card and therefore varies across trials, which allows us to identify a neural risk prediction signal. Risk prediction indeed modulates bilateral anterior insula activation (see Table S3, available at www.jneurosci.org as supplemental material). This finding is consistent with previous studies that have identified uncertainty-related activation in insula (Elliott et al., 2000; Critchley et al., 2001; Ernst et al., 2002; Paulus et al.,

2003; Hsu et al., 2005; Huettel et al., 2005, 2006; Grinband et al., 2006).

Activation in anterior insula correlates with risk prediction errors

The risk prediction before the first card is followed by a risk prediction error when the first card is revealed. The subsequent estimate of risk prediction, before the second card is displayed, is followed by a risk prediction error when the second card is revealed. Both risk prediction errors correlated significantly with activity in bilateral anterior insula [Fig. 3*A* and Fig. S2 (available at www.jneurosci.org as supplemental material)]. Increasing risk prediction errors were reflected in increasing activation (right insula, Fig. 2*B*) ($p < 0.001$; $r^2 = 0.65$). Importantly, Figure 3*B* also shows that the activation levels for risk prediction errors recorded after the first card are equivalent to those recorded after the second card. That is, the relative magnitudes of the activations after the first and second card correspond to the relative magnitudes of the risk prediction errors. The average activation at zero risk prediction error constitutes an outlier. In one important sense, it should be, as it is the average activation across all the trials where there is no risk after seeing the first card: the subject knows for certain that she will win or lose. That is, there is no risk prediction error, because there is no risk. In contrast, in all other trials, there is risk (and the risk prediction error is always nonzero).

Figure 4 shows how the revelation of the first card is immediately followed by activation that reflects the risk prediction error. Activation correlating with the risk prediction after seeing the first card (and referring to prediction of the outcome revealed through the second card) emerges later (~5 s delay), in an area in insula that is slightly more superior and anterior. The time courses in previous studies suggest that the onset of uncertainty-related activation is delayed with respect to the risk onset (Huettel et al., 2005). In accordance with these results, we find here that the risk prediction signals in bilateral insula area show a late onset. Figure 5 shows the time courses for activation in anterior insula related to risk prediction error and time-locked at display of card 2. Trials with zero risk after display of card 1 and, hence, zero risk prediction error at card 2, are not included (they constitute the outlier in Fig. 3*B*). At the peak of the response, activations stratify significantly by level of risk prediction error. Modulo the usual delay in hemodynamic response, the activation effects appear to be immediate.

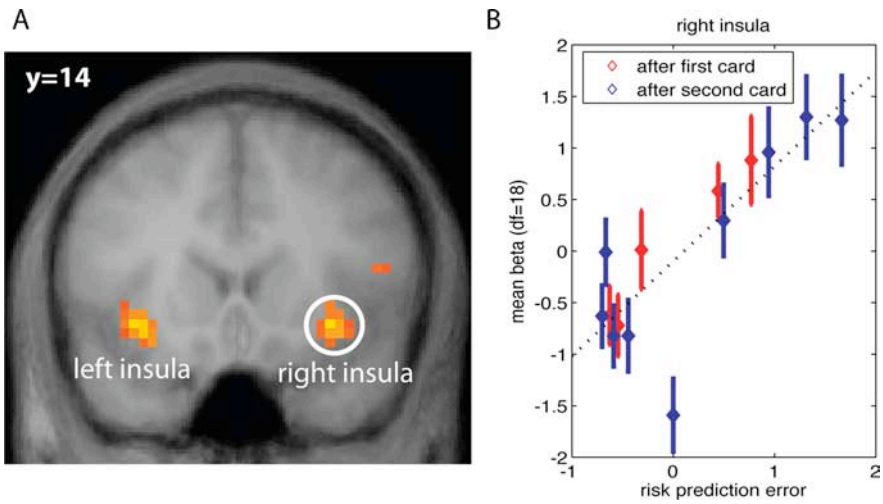

Figure 3. ***A***, Activation in bilateral insula correlates positively with risk prediction error as of display of both cards 1 and 2 (random effects, df = 18; $p < 0.0005$). ***B***, Activation levels in right insula show a significant linear relationship with the level of risk prediction error at the time of the first card (blue) as well as the second card (red). Furthermore, the functional relationships are comparable at the first and second card. The variable on the horizontal axis is: sign(RPE)*sqrt(|RPE|) (RPE, risk prediction error; sqrt, square root function; sign, sign function). This transformation makes the magnitudes of the RPE comparable with that of reward prediction errors.

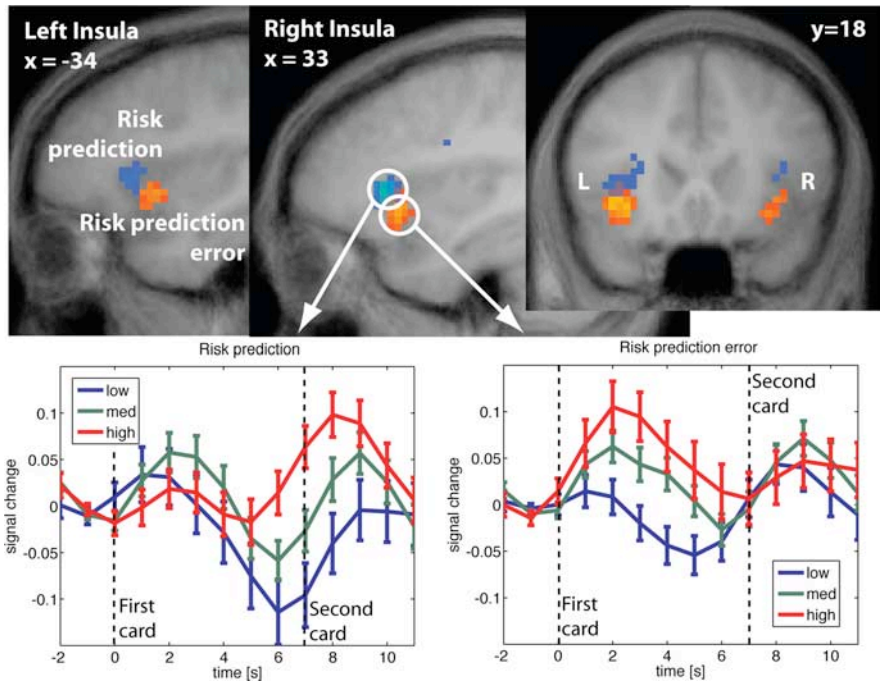

Figure 4. Top, Activation in bilateral insula correlates with both risk prediction (blue) and risk prediction error (red). Risk prediction is reflected in an area slightly more superior and anterior than risk prediction error. Note, that both the red and blue clusters reflect positive correlations (random effects, df = 18; $p < 0.0005$). Different colors were chosen for better visualization. Bottom, Adjusted time courses in right insula at the first card. Before the first card, risk prediction is constant across all trials (Fig. 1). The risk prediction error at the first card is a function of the subject's bet and the first card. It is reflected in the time course immediately after the first card is shown (bottom right panel). Preceding the second card, a second estimate of risk prediction arises, which is reflected in the time course after the first card but only after a short delay (bottom left panel).

Pattern of activation in anterior insula when ignoring risk prediction errors

Previous studies (cited above) of insula activation in the context of uncertainty only reported signals that increased in the level of uncertainty. These studies do not report a risk prediction error, not only because it was not hypothesized, but also because of its peculiar timing. To determine whether our activations were con-

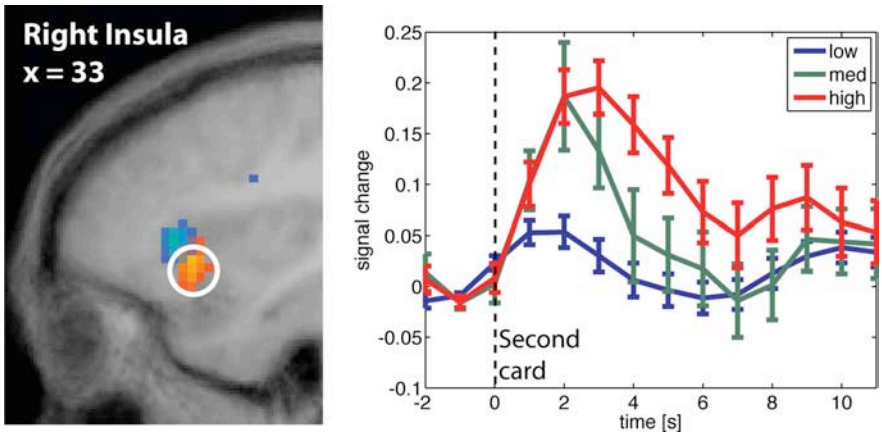

Figure 5. Adjusted time course in right anterior insula around display of the second card, time locked to the presentation of card 2. Location of activation is displayed to the left (Fig. 3). The risk prediction error is reflected in the time course immediately after the second card is shown. Trials in which all risk is resolved at the first card (i.e., first card = 1 or 10) are excluded, because there is no risk prediction and therefore no risk prediction error.

sistent with the earlier accounts of uncertainty-related activations in insula, we replaced our model with one similar to previous studies. Specifically, we included only risk prediction at the two stages of the trial as predictors, leaving out risk prediction errors. We used a block regressor (a "boxcar" function over the length of the interval between card 1 and card 2 with height modulated by the square of the reward probability) equivalent to those of previous studies to try to pick up uncertainty activations without regard to their exact timing. Insula activations in response to this regressor were positive and therefore consistent with earlier accounts in the literature, albeit less significant ($p < 0.05$, uncorrected). The lower significance is attributable to the fact that there is no risk prediction error at stimulus presentation in previous studies and, hence, no potential confounding factor. The only exception is the study by Critchley et al. (2001), who also reports activation at lower significance levels.

Discussion

We found two signals in bilateral anterior insula: one reflecting risk prediction and one reflecting risk prediction error. The two signals were not only spatially separated but also temporally (Fig. 4). Different levels of risk prediction were best dissociated with a short delay after the risk cue (first card), whereas different levels of risk prediction error were well discriminated immediately after either card was shown.

These findings support our hypothesis that there are two signals in anterior insula, a late-onset risk prediction signal followed by a fast-onset risk prediction error signal at the time of the outcome. Although previous studies have documented risk-related activations in insula (Elliott et al., 2000; Critchley et al., 2001; Ernst et al., 2002; Paulus et al., 2003; Hsu et al., 2005; Huettel et al., 2005, 2006; Grinband et al., 2006), the neural responses to risk prediction errors have not been reported, nor has their occurrence been differentiated from that of the risk prediction signals.

The time courses of the two signals support the hypothesis that they play a distinct role in risk processing. Risk prediction may act as an anticipatory signal before risk is realized; correspondingly, we found that the neural response to risk prediction was delayed after the risk cue and remained active at the time of the outcome. In contrast, the risk prediction error may mediate learning. Speed is important, to enable quick adaptation to rapidly changing uncertain environments. Confirming this view, the response to risk prediction errors emerged immediately after risk was realized and remained active only briefly.

The risk anticipation signal, in contrast, was delayed. With peak activation at ~8 s after display of card 1, and accounting for the usual 4 s lag until maximal hemodynamic response, activation related to risk prediction appears to emerge with a delay of about 4 s. One cannot exclude the possibility, however, that the activation is time-locked to the resolution of uncertainty (i.e., display of card 2). To discriminate between the hypothesis of activation with a fixed delay, and one time-locked to the outcome, one would need an experimental design where the time between cards 1 and 2 is varied substantially.

The relationship between risk processing and emotions in insula

The insula has been implicated as a critical structure in linking affective processing with motivation, decision making, and behavior. Specifically, insula activation has been reported for a wide variety of negatively and positively valenced affective processes, including integration of body states and emotions (interoception) (Mesulam, 1998), hunger (Craig, 2002), craving and addiction (Naqvi et al., 2007), empathy for pain (Singer et al., 2004), and social rejection (Eisenberger et al., 2003). Anatomically, the insula is well positioned to play an integrative role in linking affective value with adaptive behavior, because it possesses bidirectional connections with numerous structures implicated in reward and decision making, including orbitofrontal cortex, amygdala, anterior cingulate, and nucleus accumbens (Reynolds and Zahm, 2005).

Altogether, this suggests that the insula plays a crucial role in integrating bodily states and affective value and forms part of a functionally specialized network for reward-related adaptive behavior (Critchley et al., 2001; Craig, 2002; Bechara and Damasio, 2005). Consistent with this view, and with a previous report that changes in bodily states of professional financial traders reflect risk (Lo and Repin, 2002), we found neural signals that correlated with risk prediction errors. Our evidence suggests that anterior insula is not just a crude relay of the information carried by bodily states, but that it transmits this information in a precise, quantitative manner. Defects in the functioning of insula would hinder this transmission and, therefore, may lead to anomalous attitudes in a context of uncertainty, consistent with recent evidence (Bechara and Damasio, 2005). This view is further corroborated by studies that have suggested the involvement of insula in risk aversion (Kuhnen and Knutson, 2005; Rolls et al., 2008) and anxiety (Stein et al., 2007). Our evidence that at least one brain structure resolves risk beyond a simple high/low classification is predicted by modern decision theory, according to which risk needs to be precisely measured to determine the correct value of an uncertain outcome (Markowitz, 1991) and to correctly update estimates of expected reward (Preuschoff and Bossaerts, 2007).

Learning

Although forms of reward prediction learning have been found for a diversity of species, including bees (Real, 1991) and nonhu-

man primates (Schultz, 2004), the extent of risk prediction learning across species is not known. In humans, risk prediction learning must somehow take place, because bodily states (Lo and Repin, 2002) and learning rates (Behrens et al., 2007) have been shown to correlate with risk prediction.

Estimation of risk can play a dual role. The first is to guide choice for all risk-sensitive agents. Such risk sensitivity is seen in the behavior of many organisms (Caraco, 1982; Barnard and Brown, 1985), including most humans (Weber et al., 2004). The second role for estimates of risk prediction is to modulate learning of expected rewards, even for risk-neutral agents. Reward learning is typically associated with the subcortical dopaminoceptive structures (McClure et al., 2003; O'Doherty et al., 2003). It is not clear to what extent dopamine modulates risk learning. Dopamine is usually associated with positive aspects of learning, whereas risk could be considered aversive. Still, a role for dopamine in aversive learning has been suggested (Nader et al., 1997), and the view that risk is aversive is not entirely correct. Risk may generate positive value through discovery of new opportunities. This view is made explicit in modern option valuation theory (Black and Scholes, 1973). More importantly, other neurotransmitters, specifically serotonin, noradrenaline, and acetylcholine, have been suggested to be implicated in learning (Doya, 2002, Yu and Dayan, 2003) and are known to modulate activation during learning in insula (Berman et al., 2000).

In summary, we hypothesized and showed that insula activations in the context of a monetary gamble reflected both risk prediction and risk prediction errors. These are crucial inputs for the assessment of risk in a rapidly changing, uncertain world. Our results suggest that the previous understanding that insula is involved in crude, uncertainty-related phenomena such as complexity, ambiguity, and risk needs to be expanded to allow for the possibility that insula encodes precise quantitative information about risk prediction emanating from changes in bodily states. Most significantly, our findings indicate that the role of insula is not limited to assessing uncertainty: the activations that correlate with risk prediction errors suggest insular involvement in risk prediction learning. Reward anticipation in the dopaminergic system developed in an analogous way: it started with the idea of encoding expected rewards, but it later needed to accommodate reward prediction errors. The idea of a reward prediction error has led to new insights into addiction, mental illnesses, and pathological decision making (Montague et al., 2004). Analogously, it is to be expected that the notion of risk prediction errors and possible disruptions in risk prediction learning may also have significant clinical implications.

Appendix

The mathematics of prediction errors, risk predictions, and risk prediction errors

Definitions

In our gamble, let P_1 denote the expected reward conditional on the number on card 1, and P_2 the actual reward, revealed on display of card 2. Before display of card 1, the task is to predict P_1; after display of card 1 and before display of card 2, the task is to predict P_2.

Let P_0 be the prediction of P_1, i.e., $P_0 = E[P_1]$. The prediction error (as of display of card 1) equals $P_1 - P_0$; the risk prediction (before display of card 1) is the expected size-squared of this prediction error, namely, the variance $E[(P_1 - P_0)^2]$. The risk prediction error is the actual minus the expected size-squared: $(P_1 - P_0)^2 - E[(P_1 - P_0)^2]$.

Analogously, after display of card 1, P_1 is the prediction of P_2: $P_1 = E[P_2]$. The prediction error at card 2 equals $P_2 - P_1$. The risk prediction (before display of card 2) is the expected size-squared of this prediction error, namely, $E[(P_2 - P_1)^2]$. The risk prediction error is the actual minus the expected size-squared: $(P_2 - P_1)^2 - E[(P_2 - P_1)^2]$.

Examples

In the three exemplary trials in Figure 1, the subject bets that the second card is lower.

Before display of the card 1, the odds that the subject wins or loses ($1.00) are 50–50. This yields a reward prediction $P_0 = 0$ for all trials (Fig. 1, top graph). To compute the prediction risk, first consider all possible prediction errors that could obtain when card 1 is displayed. If card 1 equals 1, then the subject loses for sure and $P_1 = -1$; if card 1 equals 2, then the subject wins only if card 2 is 1, which occurs with 1/9 probability; otherwise she loses; so, $P_1 = \frac{1}{9}1 + \frac{8}{9}(-1) = -\frac{7}{9}$. Analogously, if card 1 equals 3, 4 . . . 10 then $P_1 = -\frac{5}{9}, -\frac{3}{9}, -\frac{1}{9}, \frac{1}{9}, \frac{3}{9}, \frac{5}{9}, \frac{7}{9}, 1$. Because $P_0 = 0$, the prediction errors are $P_1 - P_0 = P_1$, and the possible size-squared of these prediction errors are $(P_1 - P_0)^2 = \left(\frac{1}{9}\right)^2, \left(\frac{3}{9}\right)^2, \left(\frac{5}{9}\right)^2, \left(\frac{7}{9}\right)^2, 1$. Because each value has equal likelihood, 1/5, the risk prediction (i.e., the expected size-squared of the prediction error) before card 1 is $E[(P_1 - P_0)^2] = \frac{1}{5}\left\{\left(\frac{1}{9}\right)^2 + \left(\frac{3}{9}\right)^2 + \left(\frac{5}{9}\right)^2 + \left(\frac{7}{9}\right)^2 + 1\right\} = 0.41$. This value is the same for all trials; it is depicted by the black horizontal segment of the middle graph in Figure 1 before display of card 1.

At display of card 1. If card 1 equals 3 (Fig. 1*A*), the actual size-squared of the prediction error is $(P_1 - P_0)^2 = \left(\frac{5}{9}\right)^2$; this is indicated by the first red spike in the middle graph of Figure 1*A* after display of card 1. When card 1 equals 8 (Fig. 1*B*), the size-squared of the prediction error is the same (as when card 1 equals 3). See the red spike at card 1 in Figure 1*B*. When the first card equals 10, the size-squared of the prediction error is maximal, namely 1 (red spike at card 1 in Fig. 1*C*). The risk prediction error after display of card 1 is the difference between the actual size-squared of the prediction error (indicated by the red spikes) and the preceding risk prediction (black segment of middle graph before card 1); the risk prediction errors are displayed as spikes in the bottom graphs of Figure 1. For instance, in Figure 1*A*, the size-squared of the prediction error equals $\left(\frac{5}{9}\right)^2$, whereas the prediction risk was 0.41, so the risk prediction error equals −0.01. Because there are five different values of $(P_1 - P_0)^2$, there are five different values of risk prediction errors; activations corresponding to each value are indicated in red in Figure 3*B*; notice that the risk prediction error at display of card 1 is never zero.

Before display of card 2. Analogous computations can be made for risk prediction before card 2 and risk prediction errors at card 2. The outcome P_2 is either −1 (subject lost bet) or +1 (subject won). The reward prediction before card 2 is simply $P_1 = E(P_2)$ and is indicated by the line segments between card 1 and card 2 in the top graphs of Figure 1*A–C*. The risk prediction can be obtained by taking the prediction at card 1, e.g., $P_1 = -\frac{5}{9}$ (Fig.1*A*), and weighing the two possible squared prediction

errors, $\left(1 + \frac{5}{9}\right)^2$ and $\left(-1 + \frac{5}{9}\right)^2$, with the respective probabilities, $\frac{2}{9}$ and $\frac{7}{9}$, to obtain: $E[(P_2 - P_1)^2] = \frac{2}{9}\left(1 + \frac{5}{9}\right)^2 + \frac{7}{9}\left(-1 + \frac{5}{9}\right)^2 = 0.69$. Additional calculations reveal that the prediction risk when card 1 equals 3 (Fig. 1*A*) is the same as when card 1 equals 8 (Fig. 1*B*). The prediction risk is indicated with the horizontal segments between card 1 and card 2 of the middle graph of Figure 1*A–C*.

At display of card 2

The red spikes at display of card 2 denote the squared prediction errors (i.e., the realized risks). The risk prediction errors at card 2 are obtained by comparing the squared prediction error at card 2 with the risk predicted before card 2 (but after card 1). They are indicated with the spikes at card 2 in the bottom graph of Figure 1*A–C*.

For example, in Figure 1*A*, the squared prediction error is $\left(1 + \frac{5}{9}\right)^2 (= 2.42)$, and the risk prediction is 0.69, so the risk prediction error is 2.42–0.69 = 1.73.

Additional comments

There are many more possible risk prediction errors at card 2 than at card 1. Importantly, the risk prediction error may equal 0; this happens when the prediction at card 1 is perfect (as is the case in Fig. 1*C*), and hence there is no prediction risk to start with. Otherwise, the risk prediction errors are nonzero. That is, in our setting, whenever there is risk, there will be a nontrivial risk prediction error. Except when there is no prediction risk, at card 2, there are two possible squared prediction errors for each level of prediction risk, so there are in total nine possible levels of risk prediction errors (two for each of the four levels of nontrivial prediction risk and one when there is no risk). The activations corresponding to these nine possible levels are indicated in blue in Figure 3*B*.

As a function of probabilities of reward conditional on card 1, expected rewards P_1 increase linearly, whereas risks $E[(P_2 - P_1)^2]$ change nonlinearly. Likewise, risk prediction errors at card 1 and card 2 are nonlinear (quadratic) in these probabilities (Fig. 2). As such, brain regions potentially involved in tracking expected rewards can be identified by verifying that they generate activation that increases linearly in probability of reward; to identify brain regions potentially involved in encoding risks and risk prediction errors, we looked for activation that changed quadratically with reward probability.

See also supplemental material (available at www.jneurosci.org as supplemental material) for general formulas.

References

Barnard CJ, Brown CAJ (1985) Risk-sensitive foraging in common shrews (Sorex-Araneus L). Behav Ecol Sociobiol 16:161–164.

Bechara A, Damasio AR (2005) The somatic marker hypothesis: a neural theory ofeconomic decision. Games Econ Behav 52:336–372.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221.

Berman DE, Hazvi S, Neduva V, Dudai Y (2000) The role of identified neurotransmitter systems in the response of insular cortex to unfamiliar taste: activation of ERK1–2 and formation of a memory trace. J Neurosci 20:7017–7023.

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Politic Econ 81:637–654.

Caraco T (1982) Aspects of risk-aversion in foraging white-crowned sparrows. Anim Behav 30:719–727.

Craig AD (2002) How do you feel? Interoception: the sense of the physiologicalcondition of the body. Nat Rev Neurosci 3:655–666.

Critchley HD, Mathias CJ, Dolan RJ (2001) Neural activity in the human brain relating to uncertainty and arousal during anticipation. Neuron 29:537–545.

Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495–506.

Eisenberger NI, Lieberman MD, Williams KD (2003) Does rejection hurt? An FMRI study of social exclusion. Science 302:290–292.

Elliott R, Friston KJ, Dolan RJ (2000) Dissociable neural responses in human reward systems. J Neurosci 20:6159–6165.

Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. Econometrica 50:987–1008.

Engle RF (2002) New frontiers for ARCH models. J Appl Econometr 17:425–446.

Ernst M, Bolla K, Mouratidis M, Contoreggi C, Matochik JA, Kurian V, Cadet JL, Kimes AS, London ED (2002) Decision-making in a risk-taking task: a PET study. Neuropsychopharmacology 26:682–691.

Grinband J, Hirsch J, Ferrera VP (2006) A neural representation of categorization uncertainty in the human brain. Neuron 49:757.

Heger M (1994) Consideration of risk in reinforcement learning. In: 11th International Conference on Machine Learning, pp 105–111. San Francisco: Morgan Kaufmann.

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision making. Science 310:1680–1683.

Huettel S, Song A, McCarthy G (2005) Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. J Neurosci 25:3304–3311.

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural signatures of economic preferences for risk and ambiguity. Neuron 49:765–775.

Koenig S, Simmons RG (1994) Risk-sensitive planning with probabilistic decision graphs. In: (KR)'94: principles of knowledge representation and reasoning (Doyle J, Sandewall E, Torasso P, eds), pp 363–373. San Francisco: Morgan Kaufmann.

Kuhnen CM, Knutson B (2005) The neural basis of financial risk taking. Neuron 47:763–770.

Lo AW, Repin DV (2002) The psychophysiology of real-time financial risk processing. J Cogn Neurosci 14:323–339.

Markowitz H (1991) Foundations of portfolio theory. J Finance 46:469–477.

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.

Mesulam MM (1998) From sensation to cognition. Brain 121:1013–1052.

Mihatsch O, Neuneier R (2002) Risk-sensitive reinforcement learning. Machine Learn 49:267–290.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalicdopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Montague PR, Hyman SE, Cohen JD (2004) Computational roles for dopamine in behavioural control. Nature 431:760–767.

Nader K, Bechara A, van der Kooy D (1997) Neurobiological constraints onbehavioral models of motivation. Annu Rev Psychol 48:85–114.

Naqvi NH, Rudrauf D, Damasio H, Bechara A (2007) Damage to the insula disrupts addiction to cigarette smoking. Science 315:531–534.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

Paulus MP, Rogalsky C, Simmons A, Feinstein JS, Stein MB (2003) Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. NeuroImage 19:1439–1448.

Preuschoff K, Bossaerts P (2007) Adding prediction risk to the theory of reward learning. Ann NY Acad Sci 1104:135–146.

Preuschoff K, Bossaerts P, Quartz S (2006) Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51:381–390.

Real LA (1991) Animal choice behavior and the evolution of cognitive architecture. Science 253:980–986.

Reynolds SM, Zahm DS (2005) Specificity in the projections of prefrontal and insular cortex to ventral striatopallidum and the extended amygdala. J Neurosci 25:11757–11767.

Rolls ET, McCabe C, Redoute J (2008) Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. Cereb Cortex 18:652–663.

Schultz W (2004) Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. Curr Opin Neurobiol 14:139–147.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD (2004) Empathy for pain involves the affective but not sensory components of pain. Science 303:1157–1162.

Stein MB, Simmons AN, Feinstein JS, Paulus MP (2007) Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. Am J Psychiatry 164:318–327.

Stroud JR, Bengtsson T (2006) Sequential state and variance estimation within the ensemble kalman filter. In: CISES. Chicago: University of Chicago.

Sutton RS (1988) Learning to predict by the methods of temporal differences. Machine Learn 3:9–44.

Weber EU, Shafir S, Blais AR (2004) Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychol Rev 111:430–445.

Yu AJ, Dayan P (2003) Expected and unexpected uncertainty: ACh and NE in the neocortex. In: Advances in neural information processing systems, Vol 15, pp 173–180. Cambridge, MA: MIT.