Exposing the cognitive algorithm used in probabilistic reasoning through eye movement analysis

Authors: Manuele Reani (a)*, Niels Peek (b), Caroline Jay (a)

(a) School of Computer Science, the University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK
(b) Division of Informatics, Imaging & Data Sciences (L5)Division of Informatics, Imaging & Data Sciences, The Health eResearch Centre Vaughan House Portsmouth Street, Manchester M13 9GB
∗ Corresponding author: m.reani@manchester.ac.uk

ABSTRACT

Humans find probabilistic reasoning difficult. Whilst previous studies have proposed explanations for aspects of faulty reasoning, we currently lack a cognitive model that elucidates the full reasoning process. Through analysis of the eye movements of people engaged in probabilistic reasoning, we expose the visual behaviour that is associated with accurate and inaccurate responses to problems. Using the data from visual fixations and transitions, we present an evidence-based cognitive model covering both accurate and inaccurate reasoning, demonstrating the process by which people reach a correct response, and illuminating where mistakes occur prior to an erroneous response. The results provide the first behavioural evidence for the way humans deal with information during probabilistic reasoning, with applications to a wide range of fields where complex decision making its required.

Keywords: Bayesian reasoning; eye-tracking; information visualization; decision making; transition analysis; cognitive modelling

1. Introduction

Inaccurate reasoning about uncertainty leads to poor judgment and faulty decision-making (Kahneman & Tversky, 1973). Humans find many calculations requiring an understanding of probability hard to perform, and this is particularly true of Bayesian reasoning problems which ask for a probability prediction given existing knowledge about a situation (Eddy, 1982). An example of this, adapted from Brase (2014), asks people to predict the weather, given historical data showing the relationship between barometric pressure and the likelihood of rain:

In Gotham city, it is either rainy or sunny. On average, out of 1000 days, 200 are rainy days. Of these rainy days, 180 are associated with low barometric pressure. 300 of the remaining sunny days are also associated with low barometric pressure. What is the probability that today will be rainy, given that it is a low pressure day?

Arriving at the correct answer – which is 0.38 – requires the use of the Bayes' theorem, shown in Equation 1, where P(H|E) is the probability of the hypothesis (it is a rainy day) given the evidence (it is a low pressure day), P(E|H) is the probability of the evidence (it is a low pressure day) given the hypothesis (it is rainy day), P(H) is the probability of the hypothesis (it is a rainy day), P(E|¬H) is the

probability of the evidence (it is a low pressure day) given the opposite hypothesis (it is a sunny day) and P(¬H) is the probability of the opposite hypothesis (it is a sunny day).

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E|H) * P(H) + P(E|\neg H) * P(\neg H)} \quad (1)$$

Over the past four decades, a wealth of research has examined how information format and visual representations of problems affect Bayesian reasoning (e.g. Krynski & Tenenbaum. 2007; Khan et al., 2015; Cosmides & Tooby, 1996; Brase, Cosmides & Tooby, 1998; Sedlmeier, 1999; Ayton & Wright, 1994; Evans, Handley, Perham, Over, & Thompson, 2000; Girotto & Gonzalez, 2001; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Kahneman & Tversky, 1982; Mellers & McGraw, 1999; Sloman et al., 2003; Yamagishi, 2003; Bauer & Johnson-Laird, 1993, Brase, 2009, Micallef et al., 2012, Kellen et al 2013; Sirota et al., 2014; Garcia-Retamero & Hoffrage, 2013; Binder et al., 2015; Brase, Cosmides & Tooby, 1998; Sedlmeier, 1999, Sloman et al., 2003; Yamagishi, 2003; Bauer & Johnson-Laird, 1993; Brase, 2009; Micallef et al., 2012; Kellen et al 2013; Sirota et al., 2014; Binder et al., 2015; Khan et al., 201; Brase, 2014; Garcia-Retamero et al., 2015; Binder et al., 2015; Ottley et al. 2016; Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Ayton & Wright, 1994; Evans, Handley, Perham, Over, & Thompson, 2000; Girotto & Gonzalez, 2001; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Kahneman & Tversky, 1982; Mellers & McGraw, 1999, Sloman et al., 2003; Reani et al., 2018). There is evidence that people are better able to make correct inferences when a problem is framed using natural frequencies rather than percentages or probabilities (Gigerenzer & Hoffrage, 1995); that people may fail to consider the base rate information – i.e. P(H) in the example above – a phenomenon known as "base rate neglect" (Bar-Hillel, 1980) and; that reasoning is affected by the magnitude of the base rate information provided in the problem (Reani et al., 2018). There is also evidence of an association between Bayesian reasoning and individual differences such as different levels of numeracy and spatial ability (Garcia-Retamero et al., 2015; Ottley et al., 2016). Nevertheless, details of the cognitive mechanisms underpinning the full Bayesian reasoning remain elusive.

Here, we examine how humans perform Bayesian reasoning, taking the cognitive algorithm reported by Gigerenzer and Hoffrage (1995) as a starting point, and directly observing whether there is evidence for this in people's gaze behaviour. The algorithm states that a person arriving at a correct Bayesian inference follows four steps:

Step 1: Base rate cut. The reasoner extracts the base rate information (rainy days) from the total – i.e. P(H) in Equation 1.

Step 2: Hit rate cut. The reasoner separates the true positive (rainy days that have low pressure) from the base rate information – i.e. P(E|H).

Step 3: False alarm cut. From the data that remain from the base rate cut, the reasoner extracts the false positive (rainy days with high pressure) – i.e. P(E|¬H).

Step 4: Comparison step. Finally the reasoner compares the ratio of the true positive with the sum of the true positive and the false positive, which in the above example results in the positive predictive value – i.e. P(H|E).

To determine whether humans do indeed follow these steps when making a correct inference, and to understand how people arrive at an incorrect inference, we recorded the eye movements of people trying to solve the weather problem described above. The information was presented in either tree or Venn diagrams, which are simple, familiar ways of demonstrating relationships between sets used extensively in previous research (e.g. Binder et al., 2015; Khan et al., 2015; Reani et al., 2018), and which demonstrate the relationship between items of information in different ways. Tree diagrams support reasoning through directionality – there is an order in which the information is read – whilst Venn diagrams emphasise the nesting between sets (Reani et al., 2018). These representations also clearly separate in space the items of information used in reasoning, which is important within eye tracking studies. If a problem is in textual form, the items of information are not clearly separated, making it difficult to map gaze onto the components of the problem.

To understand the reasoning process, we considered both *which* items of information participants fixate on, and, crucially, how they *compare* these items, which can be ascertained from how they transition between them. Transition frequencies were represented as probability distributions, and a permutation test was used to assess the difference, calculated using a distance metric (Kübler et al., 2014; Davies et al., 2016), between those making correct inferences and those making incorrect inferences.

We tested the hypothesis that people who reason correctly about Bayesian problems exhibit a different gaze strategy to people who do not, and found that there are indeed significant differences in terms of both fixation and transition behaviour between the two groups. Using the four-step cognitive algorithm reported above as a foundation, we proposed a comprehensive model of reasoning, using the eye movement data as a proxy for cognition.

2. Material and Methods

The experiment was a within-subject design with only one factor, Information Format. This had two levels, Tree and Venn. Correctness, the metric measuring participants' performance, had two values, correct and incorrect.

2.1 Participants

Forty-nine participants (age range 16-36 years; 37 male and 12 female), were recruited from the University of Manchester to take part in a laboratory experiment conducted in university facilities. They were presented with an adaptation of the weather problem on a computer screen while their eye movements were recorded.

2.2 Stimuli and Procedure

Participants sat in front of a computer and, after informed consent was taken, they received the instructions on a computer screen explaining the nature of the study. They were also asked to complete a questionnaire to assess their numeracy, measured using the Subjective Numeracy Scale - ability subscale (Fagerlin et al., 2007). They were then presented with the following version of the weather problem:

In Gotham city on average, out of 1000 days, some are rainy days. Of these rainy days, some are associated with low barometric pressure. Some of the remaining sunny days are also associated with low barometric pressure.

Subsequently, participants were presented with a set of data in either a tree or Venn diagram format (the data for each were different - see Figure 1), and asked verbally to fill in the blanks of the statement:

Out of a total of 1000 days, Gotham city has ____ days of low barometric pressure, ____ of which will be associated with it actually raining.

A Tobii eye-tracker with Tobii Studio 3.2 software was used to record eye movements. The problem was presented first without data; then a second screen showed either a Tree or a Venn representation of the problem with the data. Participants were asked to answer the question below the graph verbally, providing two figures: the number of days with low pressure (the sum of 'rainy and low pressure days' and 'sunny and low pressure days'). The second figure was 'number of rainy days with low pressure'. The presentation of the two stimuli was counterbalanced; thus, half of the participants were shown the Tree first and then the Venn representation, and vice versa for the other half. Responses were audio recorded.

Bayesian problems generally ask for a prediction conveyed by a single number or a fraction. Success rates for such problems tend to be very low, even when participants are highly educated (Eddie, 1982) and visual representations of the problem are provided (Khan et al., 2015; Reani et al., 2018). If few people answer correctly, there is insufficient data to compare participants' reasoning strategies. The version of the problem used here thus asked participants to provide two quantities, which, as demonstrated in research, is likely to increase the number of correct answers (Girotto & Gonzalez, 2001; Brase, 2008; Brase, 2014).

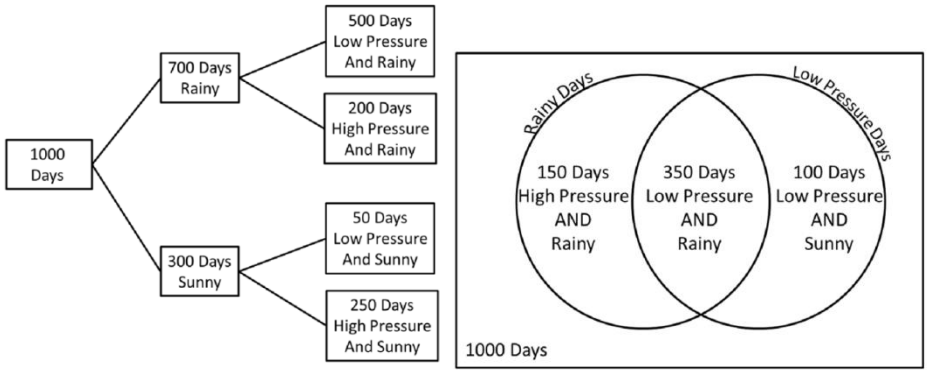

Figure 1 – Tree Diagram (left) and Venn Diagram (right)

2.3 Analysis

A Mann–Whitney U test was used to investigate whether Numeracy had any effect on Correctness. A multilevel logistic regression was used to determine whether there was a relationship between the predictor Format and the response variable Correctness, controlling for the covariate Numeracy.

Descriptive gaze data statistics were calculated for dwell time and fixation frequency for given AOIs. Transitions were defined as sub-sequences of 2 characters in specific orders (e.g. R-Rl is the

transition from ‘total rainy days’ to ‘rainy days with low pressure’). An analysis of longer subsequences was not performed as the possible number of unique transition patterns increases dramatically with the pattern length, and each tends to occur with a much lower frequency, leading to sparse data that is difficult to interpret (Kübler et al. 2017). To identify significant differences in the distribution of transitions between the Correct and Incorrect groups, 2 permutation tests, obe for each condition, with 10,000 permutations were performed using the Hellinger distance as defined by Equation 2.

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{x\in X}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2} \qquad (2)$$

In Equation 2, “p” and “q” are the two discrete distributions to be compared over a vector of values “x”, which here represent a vector of transitions with their relative frequencies (Hellinger, 1909). This method was used as the assumptions for standard statistical tests were violated, it is robust against Type 1 error (Wilcox, 2010) and it is suitable to compare unevenly-sized groups that often occur when participants are assigned to these on a post hoc basis.

3. Results

All participants completed both problems. Of the 49 participants, 30 answered correctly with the tree format, and 29 with the Venn format. A Mann–Whitney U test indicated that Numeracy was higher for the Correct group (Mdn = 4.5) than for the Incorrect group (Mdn = 3.75), U = 737.5, p-value = 0.003. A multilevel regression analysis showed that information format (tree or Venn diagram) did not affect correctness (odds-ratio = 0.813, 95% CI [0.26, 2.54]).

To perform the eye movement analysis, areas of interest (AOIs) were defined over each stimulus (see the supplementary material in the appendix), and coded using representative letters, such that gaze metrics could be calculated (see Table 1).

Recoding of the AOIs

Tree Diagram		Venn Diagram	
Code	**AOIs**	**Code**	**AOIs**
B	Background	B	Background
T	Total number of days	T	Total number of days
Q	Question	Q	Question
Rh	Rainy and high pressure days	Rh	Rainy and high pressure days
Rl	Rainy and low pressure days	Rl	Rainy and low pressure days
Sl	Sunny and low pressure days	Sl	Sunny and low pressure days
Sh	Sunny and high pressure days		
R	Rainy days		
S	Sunny days		

Table 1 – AOIs re-Coding for Eye-tracking Analysis, using letters

Tree diagrams display all the information found in Venn diagrams plus the marginal values (total rainy and total sunny days), and one additional intersection (Sunny Days ∩ High Pressure Days), although it should be noted these pieces of information are not necessary to reach a correct

inference. To answer the given question correctly, the reasoner must report the total number of days with low pressure (computed by summing the number of sunny days with low barometric pressure and the number of rainy days with low barometric pressure) and the number of rainy days with low barometric pressure. The items of information required to compute these values are found in both graphs, and correspond to "Rl" and "Sl" in Table 1. AOI "B" (in the table) refers to the background of the two graphs. Gaze data from the background is excluded from the analysis, as it did not display any useful information.

3.1 Dwell time

Dwell time, measured as the cumulative time spent fixating in each AOI, varied between the Correct and Incorrect groups (see Table 2). When using the Tree format, people in the Correct group spent, on average, more time fixating in locations "R", "S", "Rl", "Rh", "Sl" and "Sh" than those in the Incorrect group. When information was displayed in the Venn format, people in the Correct group spent on average more time fixating in locations "Sl" and "Q", and less time fixating in locations "Rh", "Rl" and "T" than those in the Incorrect group.

TREE					VENN				
	Correct		Incorrect			Correct		Incorrect	
AOI	Median	IQI	Median	IQI	AOI	Median	IQI	Median	IQI
T	134	359	133	541	T	133	367	217	637
Q	900	1974	817	1341	Q	1919	3131	600	1775
Rh	284	695	166	342	Rh	783	1651	1225	1639
Rl	317	701	167	692	Rl	583	2017	1175	1534
Sl	642	862	249	692	Sl	1716	3050	908	1086
Sh	859	970	384	942					
R	250	450	0	232					
S	259	591	100	600					

Table 2 – Median dwell time across AOIs, for Tree (left) and Venn (right)

3.2 Fixation frequency

When viewing information using the Tree format, the average number of fixations per participant was higher in the Correct group (m = 50.3, sd = 30.83), than the Incorrect group (m = 40.53, sd = 19.8). This was also true for the Venn format, where for the Correct group, m = 55.93, sd = 30.29, and for the Incorrect group, m = 50.58, sd = 23.07. For analysis purposes, relative frequencies were calculated by dividing the frequency for a specific AOI by the total number of fixations across all the AOIs in that format. Table 3 shows the relative frequencies for each AOI according to format (tree/Venn) and group (Correct and Incorrect). The table also reports the odds-ratios measured as OR = (p/1-p)/(q/1-q), where "p" and "q" are the distributions of fixations across the AOIs of the two groups (Correct Vs. Incorrect).

TREE				VENN			
AOI	Correct	Incorrect	Odds	AOI	Correct	Incorrect	Odds
T	0.05	0.08	0.62	T	0.06	0.08	0.69
Q	0.26	0.32	0.76	Q	0.32	0.23	1.58
Rh	0.08	0.09	0.92	Rh	0.2	0.19	1.04

Rl	0.1	0.08	1.32	Rl	0.15	0.27	0.46
Sl	0.13	0.14	0.92	Sl	0.28	0.23	1.33
Sh	0.19	0.18	1.06				
R	0.09	0.04	2.60				
S	0.1	0.08	1.19				

Table 3 – Relative Frequencies of Fixations across AOIs, for Tree (left) and Venn (right)

The odds-ratios show that when viewing information in the tree format, the Correct group fixated more frequently in locations "R" and "Rl" than locations "T" and "Q" compared with the Incorrect group. Similarly, with the Venn format, the Correct group fixated more frequently in locations "Sl" and "Q" and less frequently in locations "Rl" and "T" compared with the Incorrect group.

3.3 Fixation Duration

For both groups and both formats the distribution of the duration of fixations was heavily skewed, with shorter fixations occurring more frequently than longer ones. When using the Tree format, for the Correct group the median fixation duration Mdn = 167ms, Q1 = 117ms, Q3 = 217ms (IQI = 100). For the Incorrect group, the median fixation duration Mdn = 150 ms, Q1 = 102 ms, Q3 = 217ms (IQI = 115). Some fixations extended over the boundary of 367ms for the Correct group, and 389.5ms for the Incorrect group (where the boundary is computed as B = Q3 + 1.5 IQR). A similar result is found for the Venn format, where for the Correct group the median Mdn = 167ms, Q1 = 133ms, Q3 = 233ms; for the Incorrect group the median Mdn = 167 ms, Q1 = 117 ms, Q3 = 217ms.

Longer fixations indicate that participants are engaged in greater information processing, or are experiencing higher cognitive load (Debue & van de Leemput, 2014). As 217ms is the 3rd quartile for all but the Venn, Correct group, this is used as the threshold for a "long fixation". Table 4 shows the relative frequencies of long fixations and their related odds ratios across both groups and formats.

TREE				VENN			
AOI	Correct	Incorrect	Odds	AOI	Correct	Incorrect	Odds
T	0.05	0.07	0.71	T	0.03	0.07	0.45
Q	0.24	0.34	0.63	Q	0.28	0.23	1.31
Rh	0.07	0.06	1.27	Rh	0.22	0.17	1.42
Rl	0.1	0.13	0.71	Rl	0.14	0.25	0.48
Sl	0.11	0.12	0.92	Sl	0.32	0.28	1.24
Sh	0.21	0.17	1.23				
R	0.13	0.06	2.22				
S	0.09	0.05	2.06				

Table 4 – Relative Frequencies of Long Fixations across AOIs, for Tree (left) and Venn (right)

When using the Tree format, the Correct group had a higher relative frequency of long fixations in locations "R" and "S", and a lower relative frequency of long fixations in locations "T", "Rl" and "Q" compared with the Incorrect group. For the Venn format, the Correct group had a higher relative frequency of long fixations in locations "Rh", "Sl" and "Q", and a lower relative frequency of long fixations in locations "Rl", and "T", compared with the Incorrect group.

3.4 Transition Analysis

The Tree format had 8 AOIs, excluding the background, yielding 56 potential transitions. The Venn format had 5 AOIs, yielding 20 potential transitions. In this analysis we excluded transitions from an AOI to itself (e.g. transition QQ) as this was interpreted as prolonged attention on a single AOI. A permutation test compared the Hellinger distance between the Correct and Incorrect groups against the distance between two groups created at random. For the Tree format Hd = 0.315 (p-value = 0.022), and for the Venn format Hd = 0.196 (p-value = 0.161). These results indicate that for the Tree format in particular, there are transitions that discriminate between groups. The optimal transitions (i.e. the ones that could best discriminate gaze behaviour between the Correct and Incorrect group) were identified based on the differences in relative transition frequencies between the two groups.

Table 5 shows the 6 transitions (out of 56) that distinguished best between the groups in terms of frequencies, using an odds-ratio scale (following transformation using Laplace smoothing). The scale is represented here as OR = (p/1-p)/(q/1-q) where "p" (for Correct) and "q"(for Incorrect) are the distributions of transitions of each of the groups. The top three are transitions with the largest odds-ratio values, the bottom three are transitions with the smallest odds-ratio values. For the tree format, Transition R-Rl occurred more frequently in the Correct group, followed by transitions R-Sh and Rl-R; transition Sl-T occurred more frequently in the Incorrect group, followed by transitions Rh-T and S-T.

TREE diagram

Trans	Correct		Incorrect		OR [CI]
	Freq	Prob	Freq	Prob	
R-Rl	12	0.03	0	0	7.49 [0.98, 57.57]
R-Sh	10	0.02	0	0	6.31 [0.81, 49.16]
Rl-R	9	0.02	0	0	5.73 [0.73, 44.98]
S-T	3	0.01	7	0.03	0.28 [0.08, 0.93]
Rh-T	0	0	2	0.01	0.19 [0.02, 1.8]
Sl-T	0	0	6	0.02	0.08 [0.01, 0.64]

VENN diagram

Trans	Correct		Incorrect		OR [CI]
	Freq	Prob	Freq	Prob	
Sl-Q	33	0.1	6	0.02	4.33 [1.79, 10.51]
Q-Sl	36	0.1	11	0.04	2.55 [1.27, 5.11]
Q-Rh	17	0.05	6	0.02	2.12 [0.82, 5.46]
Sl-T	6	0.02	11	0.04	0.39 [0.14, 1.06]
T-Sl	3	0.01	6	0.02	0.36 [0.09, 1.45]
Rl-T	3	0.01	7	0.03	0.31 [0.08, 1.2]

Table 5 – Top 6 transitions by absolute odds-ratio for Tree (top) and for Venn (bottom), with confidence intervals. Odds-ratios obtained using Laplace smoothing.

For the Venn format, transitions Sl-Q and Rl-T had the largest weight in relation to their discriminative power. Transition Sl-Q was more frequent in the Correct group, followed by transitions Q-Sl and Q-Rh; transition Rl-T was more frequent in the Incorrect group, followed by transitions T-Sl and Sl-T. To understand the fixation and transition data better, we map it on to the AOIs of the visual stimuli (Figure 2 shows the tree format, and Figure 3 the Venn format), to show

the most “discriminative pathways” of a typical reasoner in the Correct (left) and Incorrect (right) groups. An arrow with a thick line signifies a transition that occurs with higher frequency and an AOI in bold represents a longer dwell time and/or a higher fixation count. AOIs in grey-shaded colour indicate significantly higher frequency of long fixations. Dashed thick arrows (which are shown for the Tree format only) represent discriminative transitions that do not follow the natural directionality of the diagram.

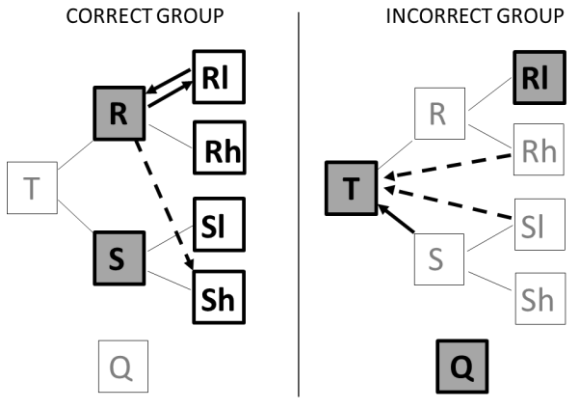

Figure 2 – Discriminative eye-movements in the tree format, for the Correct (left) and Incorrect groups.

Figure 2 shows that participants in the Incorrect group focused more on location “T” (the total number of days) compared with participants in the Correct group. The highlighted pathways Rh-T, Sl-T and S-T also demonstrate a tendency for participants in the Incorrect group to glance back repetitively to the total number of days. Participants in the Correct group focused more on locations “R”, “S”, “Rh”, “Sl” and “Sh”. Three discriminative paths found in Correct group were R-Rl, Rl-R and R-Sh. The first two are the paths associated with the second step of the cognitive algorithm of Bayesian reasoning presented in the introduction (the ‘hit rate’ cut). The box containing location “Rl” has a thick contour for both groups. For the Correct group, this is due to a larger number of total fixations and higher dwell time; for the Incorrect group, this is due to a larger number of long fixations in that specific location (see also Table 2, 3 and 4). This last AOI is shaded to indicate higher cognitive load represented by long fixations. Participants in the Incorrect group also spent more time on location “Q” – the question, where a higher frequency of long fixations was found.

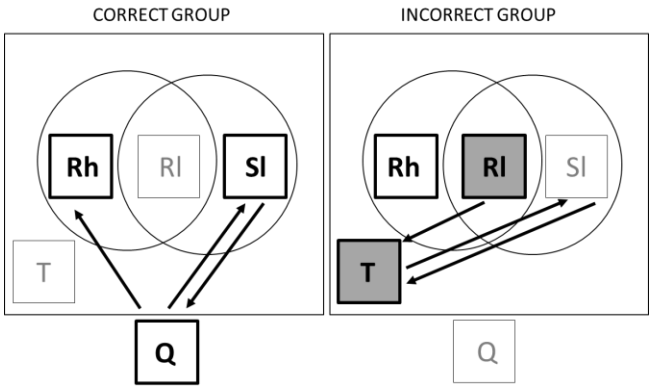

Figure 3 – Discriminative eye-movements in the Venn format, for the Correct (left) and Incorrect groups.

When using the Venn format, participants in the Incorrect group also focused more on location “T” compared with participants in the Correct group. Transitions Rl-T, Sl-T and T-Sl occur more frequently in the Incorrect Group, indicating, as with the Tree format, that participants who answer incorrectly repetitively revert to the total number of days. By contrast, participants in the Correct group directed their attention more often on location “Q”, the question, and exhibited an increased tendency to perform the transitions Rh-Q, Sl-Q and Q-Sl, thus repetitively revisiting the question at various points. Using these results, we construct a cognitive model of Bayesian reasoning, presented in the discussion.

3.5 Cognitive Algorithm Analysis

The transition analysis identified pathways that discriminate between the visual behaviour of the Correct and Incorrect groups (see Figures 2 and 3). To construct our model of reasoning, we begin by cross-referencing these with the original cognitive algorithm developed by Gigerenzer and Hoffrage (1995). The steps of this algorithm are here mapped to the AOIs in Tree the format, where a greater difference in visual transition behaviour was found between the groups. When using the Venn format, one cannot follow all the steps in the algorithm because some pieces of information are missing from this representation. Nevertheless, a comparison with participant behaviour with the Venn diagram is also reported to obtain a better understanding of the algorithm efficacy.

In the original algorithm, which we will henceforward call the GH-algorithm for brevity, the reasoner follows 4 steps:

1. Base rate cut. This would manifest itself as either transition T-R or transition R-T, depending on the order in which the reasoner visits these locations.
2. Hit-rate cut. This is represented by transition R-Rl or Rl-R.
3. False-alarm cut. This is represented by transition S-Sl or Sl-S.
4. Comparison step. This is represented by transition Rl-Sl or Sl-Rl.

We report here the odds-ratios with 95% confidence intervals to compare these transitions between the two groups. As a reminder, values considerably larger or smaller than 1 indicate larger differences between groups. To perform the first step, one needs to obtain information from location “T” (total) and location “R” (rainy days), in order to perform the base-rate cut. The odds-ratio for T-R was 1.01, 95% CI [0.34, 3.06]; for R-T, OR = 1.99, 95% CI [0.41, 9.62], indicating a higher frequency for the Correct group. For the hit-rate cut, the transition of interest is either R-Rl or Rl-R. For R-Rl, OR = 7.49, 95% CI [0.98, 57.57]; for Rl-R, OR = 5.73, 95% CI [0.73, 44.98]. For this second step, the difference between groups was large – i.e. there was a much higher frequency of relevant transitions in the Correct group. Indeed, R-Rl was one of the transitions with the largest relative difference between groups (see Table 5). For performing the false-alarm cut, the transitions of interest were either S-Sl, OR = 0.91, 95% CI [0.37, 2.23], or Sl-S, OR =1.59, 95% CI [0.57, 4.47]. Finally, for the comparison step, the transition of interest was either Rl-Sl, OR = 0.73, 95% CI [0.31, 1.68], or Sl-Rl, OR = 0.43, 95% CI [0.16, 1.17]. When using the Venn format, for the Rl-Sl transition OR = 0.82, 95% CI [0.5,1.35] and for Sl-Rl OR = 0.89, 95% CI [0.53, 1.5]. For these transitions, the relative frequency was higher for the Incorrect group in both formats. In our simplified version of the problem, this last step represented the computation of the total number of low-pressure days by simply adding two quantities together (Rl + Sl). The fact that the participants in the Incorrect group

transited more often between these two location may indicate that they found some difficulties in understanding that they had to add these two quantities together.

4. Discussion

Figures 2 and 3 show significant transitions that were not predicted by the GH-algorithm but were actually responsible for much of the difference between the two groups. For the Tree format, these were Rh-T, Sl-T, S-T and R-Sh. We identify two new steps associated with these transitions, one that relates to a correct answer, and the other which relates to an incorrect answer. Firstly, the higher frequency of transitions Rh-T, Sl-T and S-T in the incorrect group represents a tendency to regularly revisit the total number of days. We call this tendency "Reversion to Total". As shown in Figures 2 and 3, this behaviour is consistent across stimuli, suggesting that it is associated with poor reasoning independent of the type of representation used. This behaviour is consistent with the fact that almost all of the people who answered incorrectly reported the total number of days, rather than the subset of the population of interest. This is also consistent with previous research where people were asked to make an inference in the form of a proportion, and about 50% of the participants who correctly identified the numerator failed to select the correct denominator, choosing the entire population instead (Khan et al., 2015, Reani et al., 2018).

Transition R-Sh represents what we term "Checking Irrelevance", which involves switching focus to the group of days (sunny) and the indicator (high pressure) that are not required to answer the question. This step, which was found more frequently in the Correct group when using the Tree format, may be used to "cancel out" from visual memory irrelevant information shown in the graph and focus only on the useful data, or check that no information has been missed.

The analysis for the Venn condition (figure 3) also highlights an important phenomenon that was not observed in the Tree condition. It appears that people who answered correctly frequently check what pieces of information the question refers to. By doing so, they may gain a better understanding of the problem, and this may partially explain why this group answered correctly. This behaviour was found only for Venn diagrams, demonstrating that the visual presentation of the material can influence the way in which people perform Bayesian inference. In the Tree format, the Incorrect group looked at the question for longer, but without frequently transitioning between this and other pieces of data in the graph. As tree diagrams convey more information than Venn diagrams, they may be more self-explanatory. As such, when using tree diagrams, frequently looking at the question is less useful to the correct reasoner as he/she can easily extract information from the graph itself. On the other hand, the poor reasoner may fail to understand the problem and this struggling may prompt him/her to look at the question for longer but not necessarily more often.

The resulting cognitive model is presented in Figure 4. The central line represents the 4 original steps of the GH-algorithm, which is the normative model of Bayesian reasoning. The left and the right columns represent the deviations from the GH-algorithm, by the Correct and Incorrect groups respectively. They form a descriptive model of Bayesian reasoning.

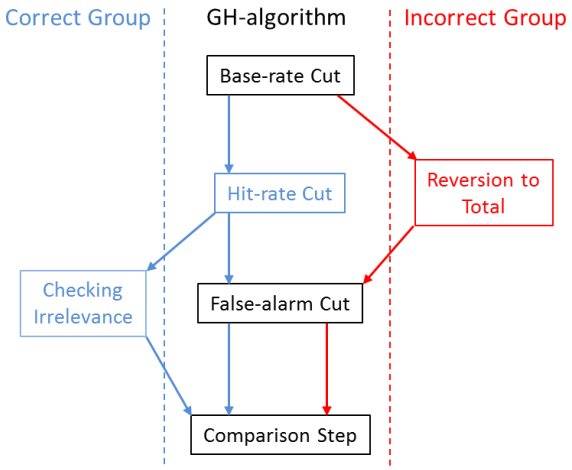

Figure 4 – Cognitive Model of Bayesian Reasoning. At the centre is the normative model. On the left is the deviation for Correct reasoning behaviour, on the right the deviation for Incorrect reasoning behaviour, according to the transitions shown in Table 5.

Along with the new steps presented above, an interesting feature of this model is the second step, the hit-rate cut, which is part of the original GH-algorithm. This appears to be performed only by participants who answered correctly. As reported by Gigerenzer and Hoffrage (1995), participants may skip this step if there is an excellent hit rate, in which case P(E|H) would be practically the same size as P(H). This was not the case in our problem where low-pressure was not a very good indicator of rain (i.e. it had low diagnosticity). Thus, participants who answered incorrectly may have overestimated the reliability of the measure of low-pressure in predicting the chance of rain.

The model may be useful for developing interventions to support reasoning, and for educational purposes. For instance, the reason why people revert to the total when solving a Bayesian problem may be related to a misunderstanding of the question. As reported in previous research, people may fail to identify the correct subset of days, replacing it with the total, because they do not understand what is asked for (Khan et al., 2015). Prior work has shown that presenting the problem statements separately, emphasising the nested-sets relationships, has a positive effect on people's ability to reason (Ottley et al., 2016). A similar approach may be used for presenting the question, where the subset of interest could be highlighted separately, and where the total in the graph could be de-emphasized, using graphical techniques to reduce reversion.

A limitation of our approach is that whilst eye-tracking analysis is able to expose cognitive mechanisms relating to overt attention, we cannot assume this maps directly to a mental model of reasoning. Nevertheless, this provides a novel, objective form of evidence that is less subject to bias than think aloud or stimulated recall techniques, which can interfere with the reasoning process (Blondon et al., 2015).

As we increasingly rely on computation, it remains essential that people who use the results of complex analyses are able to understand the process by which they are reached. We present a method for observing probabilistic reasoning via eye movement analysis, which provides the first behavioural evidence of how people perform Bayesian reasoning tasks. These results have important implications for understanding how to represent complex numerical information to humans.

Acknowledgements

This original research article was funded by the Engineering and physical Science Research Council (EPSRC). The authors are affiliated with The University of Manchester.

Author contributions

MR, the main author, developed the study concept, conducted the experiments, was responsible for study design and data collection, performed the data analysis and interpretation under the supervision of NP and CJ, and wrote the main draft of the manuscript. NP and CJ provided critical revisions. NP helped with the statistical analysis and the realization of the manuscript. CJ supervised the research process and contributed to the realization of the manuscript. All authors approved the final version of the manuscript for submission.

Declarations of interest: none.

Appendix

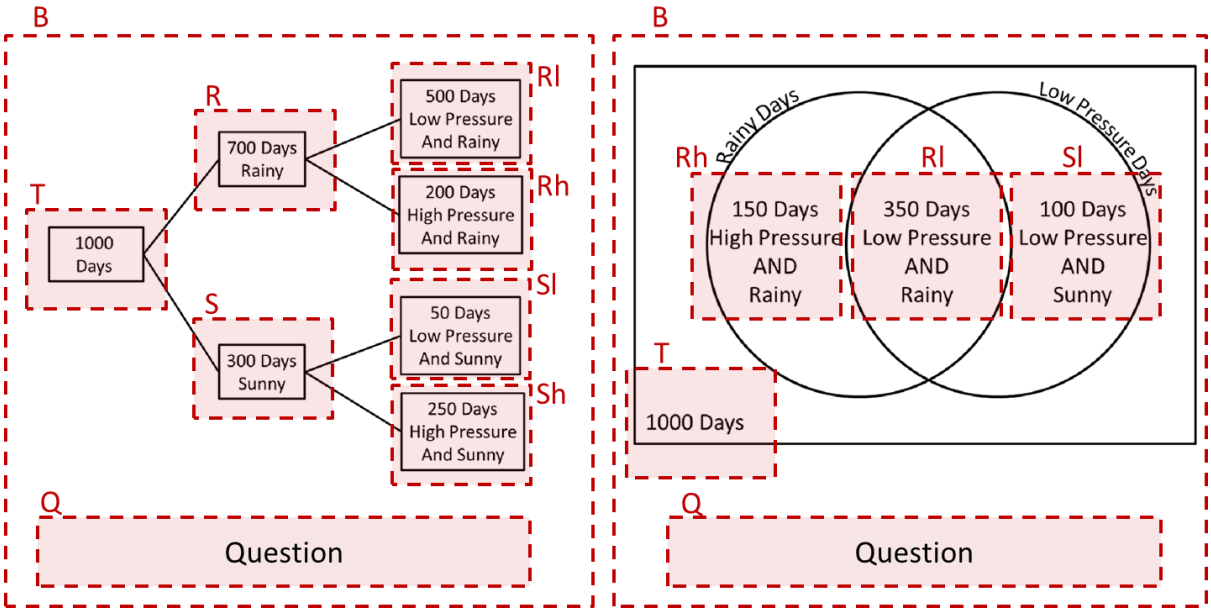

Fig. A – Schematic representation of the AOIs for the two stimuli, Tree on the left and Venn diagram on the right.

References

Coco, M. I. 'The Statistical Challenge of Scan-Path Analysis'. In 2009 2nd Conference on Human System Interactions, 372–75, 2009. doi:10.1109/HSI.2009.5091008.

Davies, Alan, Gavin Brown, Markel Vigo, Simon Harper, Laura Horseman, Bruno Splendiani, Elspeth Hill, and Caroline Jay. 'Exploring the Relationship Between Eye Movements and Electrocardiogram Interpretation Accuracy'. Scientific Reports 6 (5 December 2016): 38227. doi:10.1038/srep38227.

Holmqvist, Kenneth, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. Eye Tracking: A Comprehensive Guide to Methods and Measures. Reprint edition. Oxford: Oxford University Press, 2015.

Wilcox, Rand R. Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. 2 edition. New York, NY: Springer, 2010.

Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. (eyeLCS4) In Proceedings of the Symposium on Eye Tracking. Kubler, T.C., Kasneci, E., & Rosenstiel, W. (2014b). Research and Applications, (pp. 319–322).

Kübler, Thomas C., Enkelejda Kasneci, and Wolfgang Rosenstiel. 'SubsMatch: Scanpath Similarity in Dynamic Scenes Based on Subsequence Frequencies'. In Proceedings of the Symposium on Eye Tracking Research and Applications, 319–322. ETRA '14. New York, NY, USA: ACM, 2014. doi:10.1145/2578153.2578206.

Kübler, Thomas C., Colleen Rothe, Ulrich Schiefer, Wolfgang Rosenstiel, and Enkelejda Kasneci. 'SubsMatch 2.0: Scanpath Comparison and Classification Based on Subsequence Frequencies'. Behavior Research Methods 49, no. 3 (June 2017): 1048–64. doi:10.3758/s13428-016-0765-6.

Špakov, O., and D. Miniotas. 'Visualization of Eye Gaze Data Using Heat Maps'. Elektronika Ir Elektrotechnika 74, no. 2 (16 March 2015): 55–58. doi:10.5755/j01.eee.74.2.10372.

Davies, Alan, Gavin Brown, Markel Vigo, Simon Harper, Laura Horseman, Bruno Splendiani, Elspeth Hill, and Caroline Jay. 'Exploring the Relationship Between Eye Movements and Electrocardiogram Interpretation Accuracy'. Scientific Reports 6 (5 December 2016): 38227. doi:10.1038/srep38227.

Engbert, R., & Kliegl, R. (2001). Mathematical models of eye movements in reading: A possible role for autonomous saccades. Biological Cybernetics, 85(2), 77–87.

Kanan, C., Ray, N.A., Bseiso, D.N., Hsiao, J.H., & Cottrell, G.W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In Proceedings of the symposium on eye tracking research and applications (pp. 287–290).

Räihä Kari-Jouko, Anne Aula, Päivi Majaranta, Harri Rantala, and Kimmo Koivunen. 2005. Static Visualization of Temporal Eye-Tracking Data. Springer Berlin Heidelberg, Berlin, Heidelberg, 946-949. DOI: http://dx.doi.org/10.1007/11555261_76

Keller, Carmen, Christina Kreuzmair, Rebecca Leins-Hess, and Michael Siegrist. 'Numeric and Graphic Risk Information Processing of High and Low Numerates in the Intuitive and Deliberative Decision Modes: An Eye-Tracker Study'. Judgment and Decision Making 9, no. 5 (2014): 420–32.

Glöckner, A., & Herbold, A. K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. Journal of Behavioral Decision Making, 24 (1), 71–98. http://dx.doi.org/10.1002/bdm.684.

Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. Judgment and Decision Making, 4, 335–354.

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. Judgment and Decision Making, 6 (8), 733-739.

Brase, Gary L. 'Frequency Interpretation of Ambiguous Statistical Information Facilitates Bayesian Reasoning'. Psychonomic Bulletin & Review 15, no. 2 (1 April 2008): 284–89. doi:10.3758/PBR.15.2.284.

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. Cognition, 78, 247-276.

Brase Gary L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. Journal of Cognitive Psychology, 26:1, 81-97, DOI: 10.1080/20445911.2013.861840.

Fagerlin, A., Zikmund-Fisher, B.J., Ubel, P.A., Jankovic, A., Derry, H.A., & Smith, D.M. Measuring numeracy without a math test: Development of the Subjective Numeracy Scale (SNS). Medical Decision Making, 2007: 27: 672-680.

Lipkus IM, Samsa G., Rimer BK General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001;21(1):37—44.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Ottley, A., E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. J. Han, and R. Chang. 'Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability'. IEEE Transactions on Visualization and Computer Graphics 22, no. 1 (January 2016): 529–38. doi:10.1109/TVCG.2015.2467758.

Gigerenzer, Gerd, and Ulrich Hoffrage. 'How to Improve Bayesian Reasoning without Instruction: Frequency Formats'. Psychological Review 102 (1995): 684–704.

Krynski, Tevye R., and Joshua B. Tenenbaum. "The Role of Causality in Judgment under Uncertainty." Journal of Experimental Psychology. General 136, no. 3 (August 2007): 430–50. doi:10.1037/0096-3445.136.3.430.

Khan, Azam, Simon Breslav, Michael Glueck, and Kasper Hornbæk. "Benefits of Visualization in the Mammography Problem." International Journal of Human-Computer Studies 83 (November 2015): 94–113. doi:10.1016/j.ijhcs.2015.07.001.

Binder, Karin, Stefan Krauss, and Georg Bruckmaier. 'Effects of Visualizing Statistical Information – an Empirical Study on Tree Diagrams and 2 × 2 Tables'. Frontiers in Psychology 6 (26 August 2015). doi:10.3389/fpsyg.2015.01186.

Garcia-Retamero, Rocio, Edward T. Cokely, and Ulrich Hoffrage. 'Visual Aids Improve Diagnostic Inferences and Metacognitive Judgment Calibration'. Frontiers in Psychology 6 (16 July 2015). doi:10.3389/fpsyg.2015.00932.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44, 211–233.

Carreras, Giulia, Michela Baccini, Gabriele Accetta, and Annibale Biggeri. 'Bayesian Probabilistic Sensitivity Analysis of Markov Models for Natural History of a Disease: An Application for Cervical Cancer'. Italian Journal of Public Health 9, no. 3 (13 September 2012). doi:10.2427/7537.

Armero C, Garcia-Donato G, Lopez-Quilez A. Bayesian methods in cost-effectiveness studies: objectivity, computation and other relevant aspects. Health Econ 2010; 19: 629-43.

Briggs A, Ades A, Price M. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the Dirichlet distribution in a Bayesian framework. Med Decis Making 2003; 23: 341–50.

Hellinger, Ernst (1909), "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen", Journal für die reine und angewandte Mathematik (in German), 136: 210–271, JFM 40.0393.01, doi:10.1515/crll.1909.136.210

Wilcox, Allen R. (June 1973), "Indices of Qualitative Variation and Political Measurement", The Western Political Quarterly, 26 (2): 325–343, JSTOR 446831, doi:10.2307/446831

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition, 58, 1– 73.

Brase, G., Cosmides, L., & Tooby, J. (1998). Individuation, Counting, and Statistical Inference: The role of frequency and whole object representations in judgment under uncertainty. Journal of Experimental Psychology: General, 127, 3-21.

Sedlmeier, P. (1999). Improving statistical reasoning: Theoretical models and practical implications. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Binder Karin, Stefan Krauss, and Georg Bruckmaier. 2015. Effects of visualizing statistical information – an empirical study on tree diagrams and 2 x 2 tables. Frontiers in Psychology 6, August (2015), 1–9. DOI: http://dx.doi.org/10.3389/fpsyg.2015.01186

Evans, J. St. B. T., Handley, S. H., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. Cognition, 77, 197–213.

Girotto, Vittorio, and Michel Gonzalez. 'Solving Probabilistic and Statistical Problems: A Matter of Information Structure and Question Form'. Cognition 78, no. 3 (March 2001): 247–76. doi:10.1016/S0010-0277(00)00133-5.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, -P. (1999). Naïve probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62 88.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 509–520). Cambridge: Cambridge University Press.

Mellers, B. A., & McGraw, P. A. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage, 1995. Psychological Review, 106(2), 417–424

Sloman, Steven A., David Over, Lila Slovak, and Jeffrey M. Stibel. 'Frequency Illusions and Other Fallacies'. Organizational Behavior and Human Decision Processes 91, no. 2 (July 2003): 296–309. doi:10.1016/S0749-5978(03)00021-9.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? Experimental Psychology, 50, 97–106.

Bauer M. I. and P. N. Johnson-Laird. How Diagrams Can Improve Reasoning. Psychological Science, 4(6):372–378, 1993.

Brase, Gary L. 'Pictorial Representations in Statistical Reasoning'. Applied Cognitive Psychology 23, no. 3 (1 April 2009): 369–81. doi:10.1002/acp.1460.

Micallef L., P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. Visualization and Computer Graphics, IEEE Transactions on, 18(12):2536–2545, 2012.

Kellen, Vince, Susy Chan, and Xiaowen Fang. 'Improving User Performance in Conditional Probability Problems with Computer-Generated Diagrams'. In Human-Computer Interaction. Users and Contexts of Use, 183–92. Springer, Berlin, Heidelberg, 2013. doi:10.1007/978-3 642-39265-8_20.

Sirota, Miroslav, Lenka Kostovičová, and Marie Juanchich. 'The Effect of Iconicity of Visual Displays on Statistical Reasoning: Evidence in Favor of the Null Hypothesis'. Psychonomic Bulletin & Review 21, no. 4 (1 August 2014): 961–68. doi:10.3758/s13423-013-0555-4.

Garcia-Retamero, Rocio, and Ulrich Hoffrage. 'Visual Representation of Statistical Information Improves Diagnostic Inferences in Doctors and Their Patients'. Social Science & Medicine 83 (April 2013): 27–33. doi:10.1016/j.socscimed.2013.01.034.

Ayton, P., & Wright, G. (1994). Subjective probability: What should we believe. In G. Wright, & P. Ayton (Eds.), Subjective probability (pp. 163–183). Chichester, UK: Wiley.

Brase, Gary L., and W. Trey Hill. 'Adding up to Good Bayesian Reasoning: Problem Format Manipulations and Individual Skill Differences'. Journal of Experimental Psychology. General 146, no. 4 (April 2017): 577–91. https://doi.org/10.1037/xge0000280.

Reani, Manuele, Alan Davies, Niels Peek, and Caroline Jay. "How Do People Use Information Presentation to Make Decisions in Bayesian Reasoning Tasks?" International Journal of Human-Computer Studies 111 (March 1, 2018): 62–77. https://doi.org/10.1016/j.ijhcs.2017.11.004.

Debue, Nicolas, and Cécile van de Leemput. "What Does Germane Load Mean? An Empirical Contribution to the Cognitive Load Theory." Frontiers in Psychology 5 (October 1, 2014). https://doi.org/10.3389/fpsyg.2014.01099.

Blondon, Katherine, Rolf Wipfli, and Christian Lovis. "Use of Eye-Tracking Technology in Clinical Reasoning: A Systematic Review." Studies in Health Technology and Informatics 210 (2015): 90–94.