A more detailed model of belief formation based on Bayes’ theorem.

Introduction

There are many philosophical explanations for the formation of beliefs, such as the Spinozan theory and the Cartesian theory. These theories have both been supported and questioned and there are many experiments provided evidence for these theories. This article will introduce these theoretical models and their experimental evidence. In addition, this essay will propose a more detailed belief formation model by combining one of the experiments with Bayes’ theorem.

In the first part of this essay, I will introduce two opposing theoretical models of belief formation in philosophy: the the Spinozan theory and the Cartesian theory. In the second part, I'll introduce Gilbert's experiment. The results of this experiment provide some evidence for supported the Spniozan theory. In the third part, I will introduce Vorm experiments which is a revision of Gilbert's experiment and provides evidence to cast doubt on the Spinozan theory. In the fourth part, I will introduce Bayes' theorem. In the fifth part, I will use Bayes’ theorem to interpret the result of Vorm's experiment and put forward a neural model hypothesis.

Section1 Theory of belief formation

In philosophy and cognitive science, two distinct theories of belief formation are prevalent: the Spinozan theory and the Cartesian theory, which propose different mechanisms of how humans form beliefs.

The Spinozian theory assumes that the process of understanding and believing are the same. When we understand a proposition, we automatically accept it as true. Doubt or disbelief requires a subsequent cognitive process. Therefore, it takes more time and cognitive effort to question a proposition than to believe it.

The Cartesian theory holds that understanding a proposition is separate from believing it. According to this model, when we encounter new proposition, we understand it and analyze it, so that we can questioning the proposition or accepting the proposition. Therefore, questioning a proposition takes the same amount of time and cognitive effort as believing it.

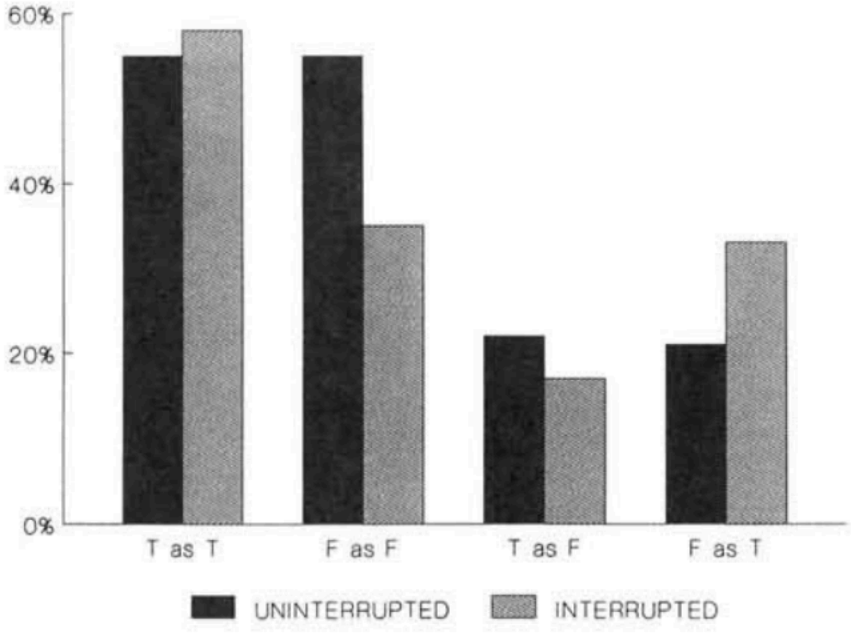
Section 2 Gilbert’s Experiment

A classic experiment by Gilbert supported the Spinozan theory on belief formation and against the Cartesian theory. Specifically, in the learning phase, the subjects were asked to sit in front of a screen, then it will show a proposition. The subjects will be told whether the previous proposition was true or false through the screen. In some of these experiments, the processing of propositions was interrupted an unrelated task quickly like press a button in response to a tone. Finally, in the test phase the subjects were shown the original propositions and asked to judge whether they were true or false.

This experiment chose the Hopi language as the experimental material to ensure the exclusion of subjects' prior knowledge. Specifically, the designers wanted to exclude the influence of the subjects' prior knowledge on the propositions by using a completely new language. In this way, the subjects' assessment of the propositions was controlled. The propositions of the experiment were presented in the form: An X is Y, where X is a Hopi word and Y is its corresponding English word. This process was like learning a language in that each Hopi word was meaningless until it was learnt.

A unrelated task was used during the experiment to interrupt the participants' processing of the propositions. Specifically, after the proposition appeared, the participant would hear a sharp ring. After this ring, the participant had to quickly press a button in response to the ring. After the proposition appeared, the screen would display three screens. There are three types of results displayed, true, false, or blank During the testing phase, participants were asked to judge the displayed propositions, which consisted of four options, true, false, no information and never seen. During the test phase, participants were shown the propositions what they had learned before. At this time, the screen will appear four options, true, false, no information and never seen. Participants were asked to choose the correct answer. In the experiment, correct recognition means that the participant correctly judged the proposition to be true or false during the test phase. For example, the true proposition in the experimental material was correctly judged to be true. Reversal means that the participant misidentifies the true value of the proposition in the test phase. For example, the true proposition in the material was misidentified to be false. In addition to this, if participants chose no information, they forgot whether the proposition was true or false. If participants chose never seen, they completely forgot they had seen the proposition.

According to The Spinozan theory, false propositions require extra steps to form while true propositions do not. Then the interruption in the experiment should have no effect on the true propositions and an effect on the false propositions. This means that a higher percentage of false propositions will be reversed than true propositions in the experimental result.
According to the Cartesian theory, both true and false propositions require extra steps to be formed. An interruption in the experiment then affects both true and false propositions. This means that the results of the experiment will have the same proportion of true propositions and false propositions being reversed.



(Gilbert et al., n.d.)

The results of the experiment are as figure show above. Regarding the correct identification rate of propositions (T as T and F as F): the percentage of true propositions correctly identified after the interruption was 58%, while the percentage of true propositions correctly identified without the interruption was 55%. The difference between these two is not significant, which could indicate that interruptions had no effect on the correct identification of true propositions. It also denied the possibility that interruptions may provide participants with time for thinking and therefore increased the proportion of correct identifications.

On the other hand, the proportion of correct identifications of false propositions after interruption was 35% (false propositions identified as false propositions), whereas the proportion of correct identifications of false propositions without interruption was 55%. It can be seen that the interruption task has a significant impact on the correct identification of false propositions, from 55% to 35%.

The results of the experiments regarding the reversal of propositions (misidentification) are as follows: the percentage of true propositions reversed after interruption is even lower than the percentage without interruption, which 17% and 22% respectively. On the other hand the proportion of false propositions reversed after interruption was 33%, while the proportion of false propositions reversed without interruption was 21%. Thus, it is seen that interruptions make the proportion of false propositions reversed increase by 12%. Indicating that interrupting the task increased the misidentification of false propositions, while it had no effect on the misidentification of true propositions.

Furthermore, during the learning phase, participants' responses to the interruption task appeared to be influenced by the truth or false of the propositions. The reaction time to perform the interruption task (hearing the ringing sound and pressing the button) after a false proposition was 577 ms on average, whereas the same data was 537 ms for true propositions. This mean that when the participant is processing the false proposition,it takes more time than the true proposition, therefore increasing the time to perform the interrupt task. This seems to confirm The Spinozan theory that processing with false propositions requires an extra step which slows down the reaction.

Also, this experiment rules out some other explanations. For example, the forgetting hypothesis, which suggests that interruptions make participants totally forget those propositions that they have learned. And in this case, participants tended to guess the unfamiliar proposition as true. However, the results showed that when participants were presented with propositions which only appeared in the test phase, 89% of them chose never seen, 8% chose false and 3% chose true. This means that participants are more likely to guess a proposition they've never seen before is false than true. This disproves the forgetting theory.

Another interpretation is the uncertainty hypothesis. This interpretation suggests that false propositions are not completely forgotten after an interruption, instead they are just uncertain of the truth value of propositions. In this case, participants tend to guess that the proposition was true. To corroborate this guessing process, the experiment considers the amount of time that participants might use to guess (hesitation). The result of the experiment was that the interrupted false propositions were evaluated to be true in 3405 ms, whereas the uninterrupted false propositions were evaluated to be true in 3985 ms. This proves that the interruptions did not increase the time used to evaluate the false propositions to be true, but decreased it, which suggest that there is no guessing process.

Section 3 Vorms' Experiment

The other experiment was operated by Vorms in 2022, and the whole experiment was modelled on Gilbert's but with slight modifications. This experiment questioned Gilbert's results to some extent and refuted the Spinozan theory. Gilbert's experiment seems to directly support the Spinozan's theory. However, the material of this experiment is a language that the participants never seen before -Hopi language. Therefore, it is difficult to generalize the conclusion to a larger scope, such as the information they processed in daily life. In 2022, Vorms designed a revised experiment based on Gilbert's experiment. This experiment applied description of fact as material, which is closer to information in daily life. Moreover, Vorm's experiments came to a completely different conclusion than Gilbert's and disproved the Spinozan's theory

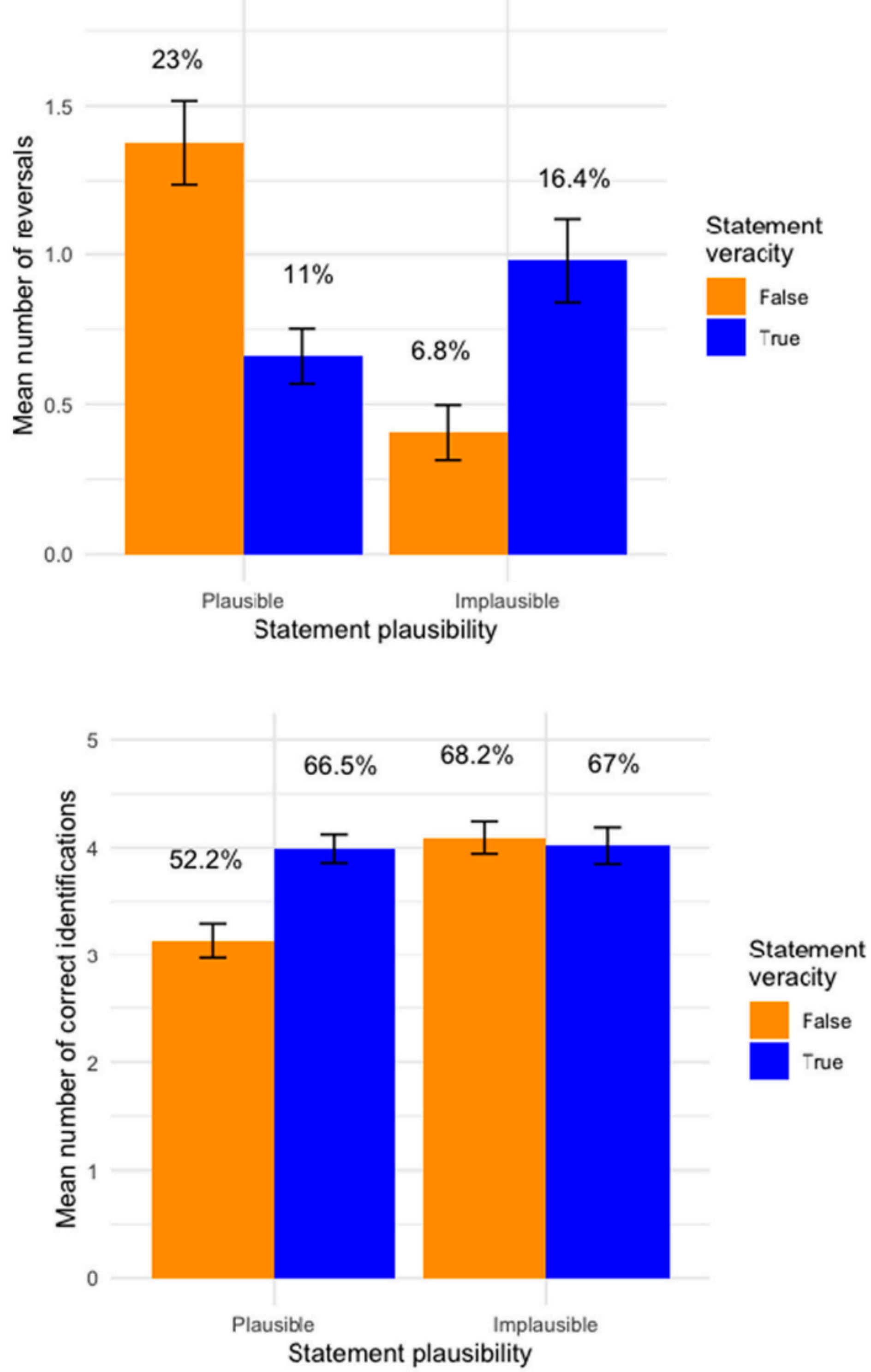
In general, during the learning phase, participants were presented with propositions on a screen. After the proposition disappears, the screen gives a sequence of numbers in voice. Then the screen will display the truth value of previous propositions. Then the participants were asked to type in a series of numbers they had just heard. When this is done, the next proposition appears. In the test phase, it was the same as Gilbert experiment. Participants saw propositions on a screen and then chose one of four options from true, false, no information and never seen.

The first difference between this experiment and Gilbert's experiment is that the material they used was no longer Hopi, but some novel descriptions of reality in English. In addition to being true or false, these descriptions are also differentiated between plausible and implausible. These propositions are subjected to a pre-test, ensuring that none of the participants know the true value of the propositions. (This is like the function of Hopi language. That make the participants' judgment on the proposition is completely affected by the truth value given by the experiment, rather than by their own knowledge) The pretest also measured the plausibility score of each proposition. They classified propositions that scored below 27 as implausible and those that scored above 56 as plausible. For example, the implausible fact that female kangaroos have three vaginas is an implausible but true proposition.

The experimental material consists of many pairs of propositions that have the same plausibility but different truth values. For example, in Florida, the law prohibiting single women from skydiving on Sundays is an implausible true proposition. At the same time the experiment provides another implausible false proposition which is in South Carolina, the law prohibits single women from skydiving on Sundays. Plausibility equivalence is maintained by this simple change. Finally, these two pairs of propositions are placed in the list of true propositions and the list of false propositions respectively.

Since the experiment was more concerned with the reversal of propositions, all propositions experienced interruptions, which mean there was no control group. The task of the interruptions was also different. After the proposition appears, the screen will provide a sequence of random numbers such as 6-0-7-4-5 in voice. The screen will then display either the truth value of the proposition or a blank screen for one second. Then a message appeared in the screen, which ask participants to enter the sequence of numbers they had just heard within eight seconds. This task therefore interrupts the participant's processing of the proposition.

The experiment predicted that the plausibility of novel propositions would influence people's beliefs formation. Specifically, when the plausible value is consistent with the true value, the rate of correct identification increases and the rate of reversal (misidentification) decreases.



(Vorms et al., 2022)

The figure shown above is the result of the experiment. For reversal (that is, true propositions are misidentified as false propositions, and false propositions are misidentified as true propositions), the experimental results indicated that the reversal rate of plausible true propositions (11%) was lower than that of implausible true propositions (16.4%); and the reversal rate of implausible false propositions (6.8%) was lower than that of plausible false propositions (23%). Therefore, the experimental results were consistent with the predictions.

For correct identification, the experimental results indicated a higher rate of correct identification of implausible false propositions (68.2%) than plausible false propositions (52.2%), which is consistent with the prediction. However, the correct identification rate of plausible true propositions (66.5%) was almost equal to the correct identification rate of implausible true propositions (67%). For this result, the experiment indicated that it is the surprise effect played a role. A Surprise effect means unexpected things are more memorable. For example, the fact that a female kangaroo has three vaginas is unexpected and therefore more memorable. Therefore, in the experiment, participants were always able to correctly identify that the female kangaroo had three vaginas is true at the test stage. Implausible true propositions caused participants to remember the proposition very well as its incredible content. This increased the correct identification rate of implausible true propositions, and eventually reaching the same rate as the correct identification rate of plausible true propositions. This can be also proved by another experimental data that the rate of implausible true propositions identified as no information and never seen before was the smallest (16.7%) (the smallest rate of forgetting). On the other hand, plausible false propositions do not produce a surprise effect. This is because plausible false propositions are more common in everyday life. For example, it would not be surprising if the law prohibiting the sale of drugs in Florida was false.

These experimental results provide some evidence against the Spinozan theory. According to the Spinozan theory, the proportion of implausible true propositions reversed to false propositions should be smaller than the proportion of implausible false propositions reversed to true propositions. However, the experimental results indicated that the proportion of implausible true propositions reversed to false propositions (16.4%) was larger than the proportion of implausible false propositions reversed to true propositions (6.8%).

Section 4 Introduction of Bayes' theorem

For Vorms' experimental results, I applied Bayes' theorem to explain it. Bayes' theorem suggests that we receive a new knowledge not exactly from 0 but update from the old knowledge. And the result of knowledge updating depends on the combination of prior belief and new evident. Bayes' theorem states that we update our old beliefs each time we accept new knowledge. In another word, our new beliefs arise from the interaction of old beliefs and new knowledge. For example, I used to believe that Paul was an honest man (old belief), but recently he cheated on me (new knowledge), therefore my trust in Paul has decreased (new belief). new knowledge in Bayes' theorem is called new evidence.

Believing something does not follow an all-or-nothing rule, but that there is a certain probability of something happening. Here I would like to pose a question. There is a meek and tidy man call Paul. Then how likely is he to be a farmer and how likely is he to be a librarian? Most people intuitively think that a large percentage of librarians are meek and tidy persons, while a small percentage of farmers are meek and tidy persons. Let us assume that the proportion is 40% for librarians and 10% for farmers. At the same time, people will intuitively consider that there are as many librarians as farmers. Therefore, the probability that the person might be a librarian is greater, closer to 4/5, but the probability that this person is a farmer is 1/5. The specific calculation process is as follows figure. And the possibilities 4/5 and 1/5 can be seen as old beliefs.

However, when we learnt new evidence, we updated our prior belief. For example, we learn the new evidence that the ratio of librarians to farmers is 1:20. Then the prior belief is updated. The process is shown below, and eventually we get the new belief that there is a 1/6 chance that Paul is a librarian and a 5/6 chance that Paul is a farmer. The specific calculation process is as follows figure.

$$1.\ \mathrm{P}(prior) = \frac{0.4 * 1/2}{0.4 * 1/2 + 0.1 * 1/2} = \frac{4}{5}$$

$$2.\ \mathrm{P}(new) = \frac{0.4 * 1/21}{0.4 * \frac{1}{21} + 0.1 * 20/21} = \frac{1}{6}$$

Section 5 Interpretation of Vorms' experiment result through Bayes' theorem

For Vorms experiment, I think the plausibility of the proposition can be seen as prior beliefs. The reason is plausibility here represents the participant's judgment of the truth and falsity of the proposition based on his own original experience. e.g. a female kangaroo has three vaginas, which most people think is unbelievable, so his plausibility score is very low in pretest which is 27. But then the screen shows that the proposition is true, which corresponds to the new evidence in Bayes' theorem. This new evidence updates the prior belief so that the possibility that the proposition is true rises. Perhaps you may question why this probability is not increased to 100%. However, I think a phenomenon called duplicity occurring here. Specifically, even if the experiment tells the participant that the truth value provided by the screen is exactly right, the participant will still maintain a certain amount of skepticism about the truth value. For example, the screen showed the implausible proposition A to be true, and then the participant was asked to answer whether proposition A was true or false immediately. The participant will definitely answer that proposition A is true because he has just remembered the correct answer is A. However, he will still be tinged with doubt inward. At this point his thought and his answer are inconsistent. And after a period of time, as the memory of the correct answer becomes fuzzy, he will then answer according to what he believe inward, i.e., his true beliefs.

When participants make judgements about propositions in the test phase, the probability of choosing true or false will be consistent with the likelihood of the proposition being true, which is so-called confidence level. This is because participants in this stage were not completely sure whether the proposition was true or false, as can also be seen from the experimental results. For example, the correct identification rate of plausible true propositions is only 66.5%.I believe that participants make choices during the testing phase based on their internal confidence level in the proposition. For example, if proposition A was developed with a confidence level of 60% during the learning phase, then participants were 60% likely to choose the proposition as true and 40% likely to choose it as false. It may be questioned why participants do not all choose true if they know that proposition A is 60% likely to be true. My explanation is that participants do not see that confidence level, because it is an internal neural mechanism that they cannot be aware of, which I will introduce later.

As shown in the table below, the results of Vorms' experiment can be interpreted as the effect of plausibility and truth value on the confidence level of Proposition A. For ease of discussion, I have labelled the effect of increasing confidence level as + and the effect of decreasing confidence as -. The last three columns of the table show Vorms' experimental results.

	Effect from prior belief (plausibility)	Effect from new evidence (true value)	Total Effect On confidence level	Correct identification rate	Reversal rate (misidentification)	Never seen/no information
Plausible false	+	--	+--	52.2%	23%	24.8%
Plausible true	+	++	+++	66.5%	11%	22.5%
Implausible false	-	--	---	68.2%	6.8%	25%
Implausible true	-	++	s++-	67% (surprise effect）	16.4%（surprise effect）	16.7%

We consider that the truth value provided by the experiment will carry more weight than plausibility (because of the default attitude that participants have in the experiment is believe the authority of the experiment). We assume that participants trusted the experimenter more than they trusted their own prior knowledge by a ratio of 2:1. Therefore, the true value would be assigned with double effect, like ++ and --. That is also the reason that the proportion of correct identifications will always be larger than the proportion of reversals. For example, plausible false propositions (+--) have a higher rate of correct identification (52.2) than reversals (23).

Possibility for propositions can be expressed as +--- and -++ when plausible values and true values are not consistent, and as +++ or --- when plausible values and true values are consistent. Thus, when the proposition is +++ or ---, the probability that the proposition is true is extremely high or extremely low, so the percentage of correct identifications will be higher than + --- and -++. Conversely, when the propositions are + -- and -++, the reversal rate will be higher than +++ and ---.

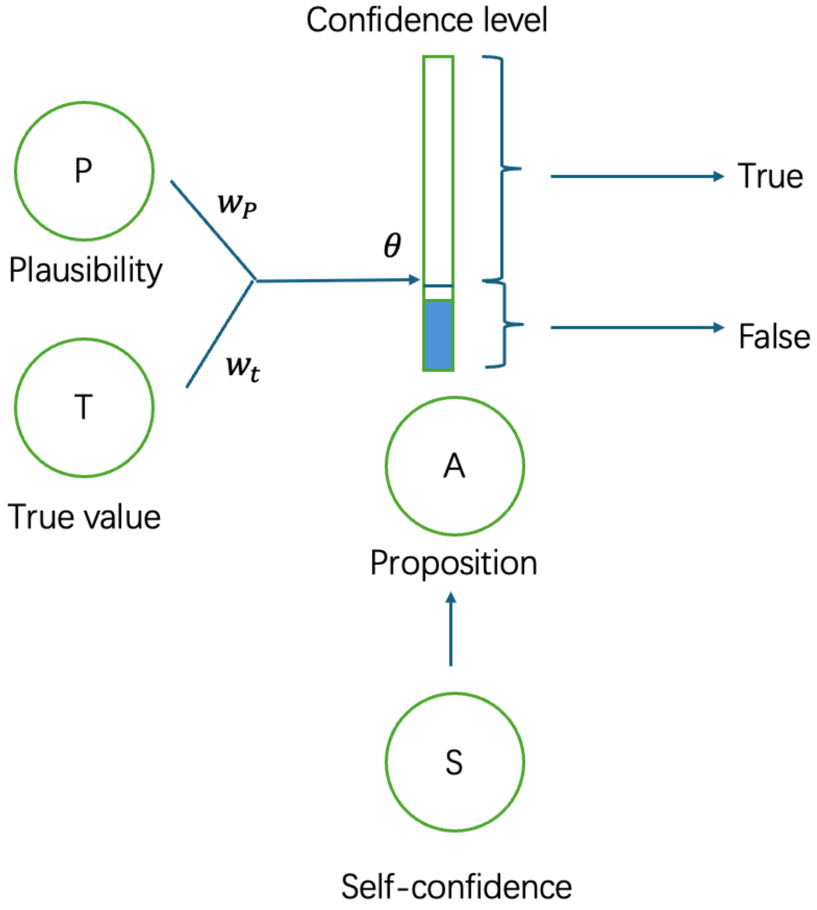
There is a special case here. Implausible true propositions can be expressed as s++-, due to the surprise effect (here I argue that the surprise effect strengthens the weighting of truth values). Because the surprise effect made the participants remember the true value better, this has been confirmed in previous Vorm's experiments. Thus, when the proposition is s++-, the correct identification rate is higher than +-- and approximately equal to +++ and ---. Also, the proposition reversal rate is reduced by the surprise effect and is 7% lower than plausible false proposition.

Section 6 Neural model hypothesis

One might suspect that the participants should be remembering the truth value rather than the confidence level. I think the reason why this is somewhat counter-intuitive is that the confidence level is not perceivable at the conscious level, but rather acts on the subconscious process of the information in the brain.

In response, I would like to propose a model hypothesis for how confidence level is stored and manipulated in neurons. The reason is that the structure of neurons allows this storage to be easily achieved, and the subconscious part of many psychological phenomena have been explained by neural networks. Specifically, as shown in the figure below, a neuron A can be considered as proposition A, and the confidence level of this proposition can be considered as the threshold θ of this neuron. The threshold θ of this neuron is jointly determined by neuron P and neuron T, and their influence is determined by their respective weights $\mathrm{w_P}$ and w_t (neuron P represents the plausibility, and neuron T represents the true value). As seen in the figure, this proposition has a high confidence level θ (70% probability that the proposition is true and 30% probability that the proposition is false). Neuron S can be seen as the current level of self-confidence of the individual. When the individual's self-confidence level is high, neuron S activates neuron A intensely so that its confidence level exceeds the threshold θ to enter the firing state and output a judgement that the proposition is true. When the individual's self-confidence level is low, neuron S is not enough to activate neuron A's confidence level to reach the threshold, so neuron A will remain in the default state and output a judgement that the proposition as false.

The introduction of neuron S (self-confidence) somehow also gives this system the ability to self-question. In another word, the flexibility to judge propositions as true or false. Self-confidence neuron S activates propositional neuron A less when the situation requires the individual to be cautious about what he or she already knows or when the individual's self-confidence is weak. This is when the individual's judgement of proposition A hovers between true and false (near threshold θ). When the situation does not require the person to be cautious, the activation of the propositional neuron A by the self-confidence neuron S is stronger. This is when the individual's judgement of proposition A is consistently true (well above the θ).



On the other hand, w_P and w_t weaken over time, but the rate of this weakening is related to the nature of the neuron. The wa of a plausibility neuron weakens more slowly over time because plausibility conforms to most propositions in the world, and it can often be strengthened. Whereas the w_t of a truth-value neuron will then weaken faster over time because it is a kind of memory for what is displayed on the screen. This explains why knowledge can only be remembered well through understanding. And knowledge that is directly remembered without understanding tends to be easily forgotten. Therefore, I predicted that when the time was further extended and the participant no longer recalled the content, the participant would gradually forget the truth values provided by the experiment. When retested in this situation, participants would be more inclined to judge the truth of the propositions only by their plausibility as in the pretest.

At the same time, w_P and w_t will enhance by the credibility of the source of information. This explains why the effect of truth value on confidence is stronger than plausibility. Because participants trust the authority of the experimenter more than their experience.

Conclusion

Overall, while Gilbert's experiments provide evidence for the Spinozan theory, Vorm's experiments come to the opposite conclusion and introduce additional details such as the plausibility of propositions. In response, I interpret Vorm's experimental results in terms of confidence level of Bayesian-theorem. In addition, I propose the feasibility of confidence level on neural networks with the corresponding modelling hypothesis. Specifically, the process of belief formation may not be as simple as the Spinozan theory. Individuals understanding a new proposition require extraction of old knowledge to form the plausibility and keep incorporating new evidence to updated judgement of the proposition. The result of their judgement is also not true or false, but a confidence level stored in a neural structure which is difficult to be aware of. This confidence level is adjusted according to the current context, so that one's judgements about propositions in different contexts have different outcomes, thus creating a kind of flexibility in beliefs.

Reference

Gilbert, D. T., Krull, D. S., & Malone, P. S. (n.d.). Unbelieving the Unbelievable: Some Problems in the Rejection of False Information.

Mandelbaum, E. (2014). Thinking is Believing. Inquiry, 57(1), 55–96. https://doi.org/10.1080/0020174X.2014.858417

Vorms, M., Harris, A. J. L., Topf, S., & Hahn, U. (2022). Plausibility matters: A challenge to Gilbert's "Spinozan" account of belief formation. Cognition, 220, 104990. https://doi.org/10.1016/j.cognition.2021.104990