Are normative probability judgments a “system two”-operation?

Joakim Carlberg
Thanks to Linnea Karlsson Wirebring and Linus Andersson
ARE NORMATIVE PROBABILITY JUDGMENTS A “SYSTEM TWO”-OPERATION?

Joakim Carlberg

Previous research on human judgment and decision making has demonstrated systematic and predictable biases of judgment in experimental settings. One example of this is the tendency to intuitively violate the conjunction rule - a simple rule of probability. This was well illustrated in the famous Linda-problem. (Tversky & Kahneman, 1983). According to the dual-process theory of reasoning, (Kahneman, 2011) reasoning fallacies such as the conjunction fallacy occurs when people fail to use analytic reasoning and instead overly rely on their intuition. The dual process theory proposes that cognitive processes underlying our intuitive impulses and our conscious reasoning constitute two different modes in the mind – system 1 and system 2- and that the intuitive system 1 are not able to compute probabilities. Furthermore, it is assumed that processes that are labeled system 1 are fast whereas system 2 are thought to be slow. We tested these time course assumptions of dual process theory in a within-subject design by comparing response time latencies between conjunction fallacy judgments and accurate probability judgments. The results showed that inducing accurate responding did not result in delayed response latency. This indicates that making accurate probability judgments does not require more processing time which goes against what would be expected by the dual-process framework.

Often, human judgment seems to be guided by intuitions about probabilities and mathematical puzzles. When faced with problems that contain large amounts of uncertainty, like deciding what car to buy or when choosing a life partner, people do not typically rely on mathematical algorithms for making an optimal decision (Davis, 2005; Slovic et al., 1977; Tversky & Kahneman, 1974). When assessing these kinds of subjective probabilities people rather rely on intuitions based on prior experiences than spontaneously conform to the classical normative laws of probability theory (Tversky & Kahneman, 1983). These observed tendencies probably mirrors facets that serves humans well in most daily life decisions although in certain circumstances this lack of normative intuition can lead to some systematic and predictable biases. This has been demonstrated in numerous
studies revealing cognitive biases such as the base rate fallacy, insensitivity to sample size, and the gambler’s fallacy (Kahneman & Tversky, 1972), which all are showing systematic discrepancies between regularities of intuitive judgment and the laws of probability theory.

One striking example of this is the apparent predisposition to violate the conjunction rule. According to the conjunction rule, the probability of a conjunction of two events A and B, or Pr(A∧B), cannot exceed the probability of any of its constitutive events, namely Pr(A) or Pr(B). When violating this rule in a direct comparison of A and A&B, it is called a conjunction fallacy. This was demonstrated by Tversky and Kahneman (1983) in the famous Linda problem. In this problem, individuals were asked to read the following description: “Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.” In the most transparent version of the problem, (Tversky & Kahneman, 1983, p. 297) the authors asked the participants to decide which of the two following alternatives they considered to be the most probable: (1) Linda is a bank teller. (2) Linda is a bank teller and is active in the feminist movement. Overall, 85% of respondents chose alternative (2) over (1) and thus, erroneously judged Pr(A∧B) as more probable than Pr(A). This has later been replicated in a variety of experimental contexts (e.g., Bar-Hillel & Neter, 1993; Andersson, et al., 2016; Stergiadis, 2015). The accumulated evidence suggests that the conjunction fallacy is a highly robust phenomena and that the Linda problem provides an illustrative paradigm.

A few different accounts for the conjunction fallacy have been discussed. Some have argued that the Linda problem contains linguistic ambiguities which would question whether the non-normative judgments really could be claimed to be fallacies. One such critique by Hertwig, Benz, & Krauss (2008) emphasize that the word and in the conjunction could lead to that the alternative p(A) is interpreted as A and not B - Linda is a bank teller which is not active in the feminist movement. Hertwig & Gigerenzer (1999) points to the word probability, claiming that, depending on the context, people make different semantic and pragmatic interpretations and that this is leading them to infer a non-mathematical meaning of the word. However, there is evidence to show that removing these ambiguities is far from a sufficient remedy (Tentori & Crupi, 2012). Although there are even more alternative explanations (see for example Gavanski & Roskos-Ewoldsen, 1991), a widely adopted explanation for the tendency to commit the conjunction fallacy and the occurrence of many other cognitive biases is because of the use of heuristics (Tversky & Kahneman, 1974).

Kahneman & Tversky (1974) proposed the use of heuristics when making judgments under uncertainty. These heuristics are described as information-processing shortcuts that make judgments computationally economic. They can sometimes provide good enough answers to complex problems but also lead to some systematic and predictable errors. Kahneman and Frederick (2002) expanded this original framework of heuristics and biases and proposed the
process of attribute substitution to explain how judgment heuristics work. The authors defined the process of attribute substitution and proposed that: “A judgment is said to be mediated by a heuristic when the individual assesses a specified target attribute of a judgment object by substituting a related heuristic attribute that comes more readily to mind.” (Kahneman & Frederick, 2002, p.4)

Thus, when people are confronted with a difficult question they substitute it with an easier one, which is answered instead. The substitution is believed to happen fast, unconsciously and effortlessly and inevitably introduces systematic biases.

One of the key heuristics that is believed to account for the conjunction fallacy is the \textit{representativeness} heuristic (Tversky & Kahneman, 1974; Tversky & Kahneman, 1983). Judgments mediated by representativeness is described as the tendency to judge the probability that A belongs to B by how representative or \textit{similar} A is to B (Kahneman & Tversky, 1972). When making subjective probability judgments as in the Linda-problem, people are believed to use representativeness as a heuristic attribute to assess the probability – the target attribute (Kahneman & Fredrick, 2002). Accordingly, since the description of Linda appears more representative of a feminist bank teller than a bank teller, people infer that it is also more probable, unaware that an attribute substitution has occurred.

Kahneman and Fredrick (2002) argues that judgments derived from attribute substitution originates from an intuitive mode of cognitive functioning and contrasts this with a controlled and \textit{deliberate mode of reasoning}. The authors used the terms “system 1” and “system 2”. They argued for a distinction between processes that are \textit{fast}, automatic and effortless (system 1) and those that are \textit{slow}, deliberate and effortful (system 2). A diverse set of proposals for the presence of such dual processing in higher cognition has been around for decades and some of them suggests that there may indeed be two architecturally and evolutionarily distinct cognitive systems underlying these two types of processing (for an overview, see Evans, 2008).

Kahneman and Frederic (2002) point out that they “use systems as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate” (Kahneman & Frederick, 2002, p.3). Kahneman (2011) gives a list of examples of the actions that are attributed to the different systems. System 1 is active when for example orienting to a sudden sound, completing the phrase “bread and…”, detecting hostility in a voice, and driving a car on an empty road. System 2 is active when for example searching the memory to identify a surprising sound, monitoring the appropriateness of your behavior in a social situation, and parking in a narrow space. According to Kahneman (2011) the diverse set of operations labeled as system 2 has one thing in common: they require attention and is disrupted when attention is drawn away (e.g. following a conversation between two persons in a crowded and noisy room). It is assumed that when making judgments, system 1 proposes quick and intuitive answers and system 2 monitors the quality of these proposals and either endorse, corrects or overrides them.
In the classical study and in numerous replications, the majority of people confronted with the Linda problem are committing the conjunction fallacy. From a theoretical perspective, based on the dual-process framework, this is a consequence of system 1 generating a representativeness based impression, while system 2 fails to detect and correct inferences based on similarity (Kahneman & Frederick, 2002; Kahneman & Tversky, 1972; Tversky & Kahneman, 1974; Tversky & Kahneman, 2013; Kahneman, 2011). This implies that avoiding the conjunction fallacy requires a system 2 process to be engaged. This in turn implies that if conjunction fallacies are a result of fast system 1 processing and normative probability judgments are a result of slow system 2 processing these two different judgments should be associated with different processing times. Normative probability judgment should be more time consuming than judgments based on similarity (Kahneman, 2011).

However, despite the pervasiveness of this argument, and the fact that this is a key prediction from many dual processing theories (e.g. Kahneman & Frederick, 2002), this has not been systematically investigated to a large extent. This can partly be attributed to the fact that many previous studies of the conjunction fallacy have used one-shot decision making, that is, one single trial per person.

De Neys, (2006) tested the response time predictions in a between group design using the Linda problem and another similar problem called the Bill-problem. They reported that those making correct analytic inferences indeed demanded more processing time than did those participants making heuristic inferences. However, using a between group-design makes the results hard to attribute solely to the chosen strategy. It is quite possible that people who exhibit a normative intuition also tends to react slower in these tasks but that it is not the strategy per se that is slow, but rather individual differences in general processing speed confounded with factors that is yet unknown.

The purpose of the present study was to test the response time predictions of the dual system theory of Kahneman & Frederick (2002) by directly comparing probability judgments where the conjunction fallacy was committed with probability judgments where the conjunction fallacy was avoided, but in a within-subject design and with multiple repetitions of each judgment increase the reliability of the measurements.

To this end, we administered a paradigm used in several previous studies (see e.g. Andersson et al., 2016) with 30 Linda-like probability judgment tasks. We included three conditions with the aim of contrasting conditions where the conjunction fallacy was committed with conditions where it was not: 1) a probability judgment condition, 2) an informed probability condition and 3) a similarity condition. Based on previous studies (Andersson et al., 2016) it has been shown that the majority of respondents commit the conjunction fallacy in the first condition, avoid making the conjunction fallacy in the second condition and choose the conjunction in the third condition. Different methods to remedy the occurrence of the conjunction fallacy, that is, to observe conditions where the conjunction fallacy is
avoided, has been studied. Sloman, Over, Slovak, and Stibel, (2003) has proposed that making nested-set relations transparent will increase the coherence of probability judgment, for instance, by emphasizing that bank tellers that are active in the feminist movement are a sub-group of the superordinate group of all bank tellers. By presenting information in a way that allows people to extract sub-sets relative to super-sets in the problem structure helps the facilitation of normative reasoning. The nested-sets hypothesis predicts that any manipulation that increases the transparency of the nested-sets relation should increase correct responding. Therefore, in the informed probability condition, we used a hint about the nested-set structure of the task in order to try to induce avoidance of the conjunction fallacy (Stergiadis, 2015). The aim of the similarity condition was to make the participants return to the same strategy used when committing the conjunction fallacy in the first condition. Based on previous studies (Israelsson, 2014; Andersson et.al., 2015) it has been shown that judgements of similarity exhibits the same response time pattern as probability judgments that violates the conjunction rule. Thus, if the response accuracy of the tasks are proven to correspond with different response time latency in this study, the similarity condition makes it possible to draw more valid conclusions.

Moreover, the study payed regard to a hypothetical instruction effect. It can be expected that any new task and the appliance of a new rule or strategy would be a slow system 2 operation (Kahneman, 2011). It was expected that if revealing the nested set structure could induce normative probability judgments in participants that had already started committing the conjunction fallacy, this would correspond to increased response latency initially, due to the instruction effect. Our main hypothesis was that, even after controlling for the instruction effect, normative probability judgments would exhibit longer processing time compared to judgments that violated the conjunction rule.

Method

Participants
Twenty-four participants between the age of 21 and 38 were tested (M\text{AGE} = 28.29). Twelve of these were female (M\text{AGE} = 26.91) and 12 were male (M\text{AGE} = 25.50). The study utilized a convenience sample and the participants were mainly recruited through personal contact or via announcement boards on the campus of Umeå University. All participants signed an informed consent prior to participation and received 99 Swedish kronor in remuneration.

Material/Design
The study consisted of 30 Linda-like problems developed by Andersson and colleagues (2016). The problems were using descriptions with the same structure as the Linda-problem (Tversky & Kahneman, 1983). Stating name, age, education and interests/hobbies (eg. “Fabian, 33, has studied musicology and likes jazz”). In contrast to the Linda problem four alternatives were used instead of two. This was
to make the problem less transparent for the participant. It was judged that, because the participants were going to do multiple problems of the same character, they may otherwise start to rely on simple visual cues (e.g. selecting the shortest alternative), instead of reading and elaborate each alternative. The alternatives contained one unlikely constituent, (e.g. “taxi driver”), one extremely unlikely constituent (e.g. “ostrich farmer”), one conjunction with two unlikely constituents (e.g. “taxi driver and orienteer”) and one conjunction with one unlikely and one likely constituent (e.g. “taxi driver and record collector”). The question was presented like this:

Fabian, 33, has studied musicology and likes jazz.

Is it more probable that the person being described is a:

1. Taxi driver and record collector (A&B)
2. Ostrich farmer (D)
3. Taxi driver (A)
4. Taxi driver and orienteers (A&C)

The correct answer would be taxi driver (A) whereas a choice of the alternative Taxi driver and record collector (A&B) would be evidence of the conjunction fallacy.

All the participants were tested with the same material in a within-subject design. The material was divided into three judgment conditions (blocks) with 10 items in each condition: Probability condition (block 1), Informed probability condition (block 2) and Similarity condition (block 3). To control for potential ordering effects, the problems was randomized into the different conditions using a computer software program. In the probability condition, the participants answered the question: “Is it more probable that the person being described is a (1, 2, 3 or 4)?” In this condition, it was expected that, in accordance to previous studies (Andersson et.al., 2016; Stergaidis, 2015), most of the participants would commit the conjunction fallacy. In the informed probability condition the question was the same as in the probability condition: “Is it more probable that the person being described is a (1, 2, 3 or 4)?” but these questions was preceded with an instruction which were giving hints about the nested set structure of the problem. (See Appendix A). Thus, by giving the subtle hint we tested whether it was possible to induce normative probability judgments in those who committed the conjunction fallacy in block 1. In the similarity condition, the question was “is the described person most similar to someone who is (1, 2, 3 or 4)?” These questions were preceded with a new instruction in order to interrupt the normative judgments induced in the previous condition(block 2). In short, they were asked to forget about previous information and instead focus on this new strategy (See Appendix B). Asking the participants to do judgments based on similarity was an attempt to make the them return to judgments based on similarity which would
theoretically, be performed by the same heuristic system –system 1- as used when committing the conjunction fallacy.

Subsequent to the main judgment tasks, participants answered an interactive PDF-questionnaire which contained a variety questions of which only data about age and sex were used for this study. The questionnaire collected data about age, sex and years of education. Participants also got to rate the perceived difficulty of/mental effort needed for each of the three conditions using a Borg CR-100 scale and answered questions about their perceived understanding of the word probability. They also answered a math problem to test their knowledge about joint-probabilities and answered three questions which would be relevant when conducting these kind of judgment tasks: 1) Are you familiar with Daniel Kahneman? 2) Are you familiar with the conjunction fallacy? and 3) Are you familiar of the Linda-problem?

Procedure
After the participant was welcomed and seated they got information about the procedure. They were told that their task was to answer questions on a computer using a numeric keyboard and to fill in an interactive PDF-questionnaire. Then they read and signed a consent stating that their participation was voluntary and that they could interrupt at any time during the experiment without having to explain why. Next, they got one trail item. After the trail item was done, participant was asked if everything was clear. If the answer to this question was yes, the test was started. First, the participants answered the ten probability questions -block 1- Then they got the reasoning hint (Appendix A) followed by another ten probability questions -block 2- Then they were presented with the instructions for the similarity task followed by ten similarity questions –block 3- When the main testing was done, the participants was seated in front of a laptop computer and asked to fill in the interactive PDF-questionnaire. Lastly, they were thanked for their participation and received a remuneration of 99 Swedish kronor.

Analysis/data
The results was presented in three sections: 1) Response accuracy, 2) first item response time, and 3) block condition response time. Each analysis had slightly different inclusion criterion depending on its purpose. In the response accuracy analysis, we were looking at the whole sample to determine whether the within-group manipulation was working and that the participants hearkened the reasoning hint about the nested set structure. In the first item response time, mean response time on the first item in each condition was compared. The purpose was to test the time course in a response pattern that resembles the many previous studies made on the conjunction fallacy in which one item is used in a one-shot decision. In this analysis, only participants that were responding as expected on the first item of each condition was included. Namely, those who choose the conjunction (A&B) on the first item of the probability condition, choose the constituent (A) on the first item in the informed probability condition and choose the conjunction (A&B) on the first item in the similarity condition. This sample
would correspond to only the participants that were committing the conjunction fallacy in for example the classic study by Kahneman, (1983). As mentioned earlier, this were not interested in any between-group variations, only the time course of the processes. Lastly, in the block condition response time, mean response time on the entire set of items I each condition was compared. This was considered the main analysis and the purpose was to test the time course of the two different types of judgments. In order to control for the initial latency effect expected following the instructions, for this analysis the first item of each condition was removed. Furthermore, participant that did not follow the response accuracy pattern expected in any of the conditions was excluded. Participants was determined not to respond as expected if they avoided the conjunction in the probability condition in more than 50% of the items, choose the conjunction more than 50% of the items in the informed probability condition or avoided the conjunction in more than 50% of the items in the similarity condition. Again, because we only were interested in time course differences between the different types of judgments, only those making the majority of the judgments as expected could be used.

Results

Response accuracy
All participants chose either the conjunction (A&B) or the constituent (A) in all problems. Overall, participants chose the constituent (A), thus, avoided the conjunction fallacy in 26% of the 10 items (M = 2.63, SD = 3.66) in the probability condition, in 77% of the 10 items (M = 7.71, SD = 3.60) in the informed probability condition and in 9.6% of the 10 items (M = 0.96, SD = 1.63) in the similarity condition.

First item response time
Nineteen of the 24 participants (79.2%) choose the conjunction (A&B) on the first item in block 1. (item 1) Thirteen of these 19 (68%) choose the constituent (A) and avoided the conjunction fallacy on the first item on block 2. (item 11) and choose the conjunction (A&B) on the first item on block 3 (item 21). After the exclusion 7 men and 6 women (n = 13) was left for this analysis. The mean response time of the first item in each condition was compared. (Figure 1). A paired sample t-test showed differences in mean response time between item 1 and item 11, t(12) = -2.22, p = .046 (two-tailed), but no difference in mean response time between Item 1 and item 21, t(12) = -1.54, p = .15 (two-tailed) or between item 11 and item 21, t(12) = .43, p = .68 (two-tailed).

Block condition response time
Nineteen of the 24 participants (79%) choose the conjunction (A&B) in more than 50% of the items in block 1. Fourteen of these 19 (74%) choose the constituent (A) and avoided the conjunction fallacy in more than 50% of items in block 2. Thirteen of these 14 (93%) choose the conjunction (A&B) in more than 50% of the
items in block 3. After this exclusion 7 men and 6 women (n = 13) was left for this analysis. Mean response time between each condition was compared (Figure 2). A paired sample t-test showed no difference in mean response time between the probability condition (block 1) and the informed probability condition (block 2), $t(12) = 1.39, p = .190$ (two-tailed), no differences in mean response time between the probability condition (block 1) and the similarity condition (block 3), $t(12) = 1.63, p = .13$ (two-tailed) and no difference in mean response time between the informed probability condition (block 2) and the similarity condition (block 3), $t(12) = -.036, p = .97$ (two-tailed).

Figure 1. Mean response time for the first item in each condition (milliseconds). Probability (item 1), informed probability (item 11) and similarity (item 21). Error bars denotes one standard error of the mean.
Figure 2. Mean response time in each condition (milliseconds). Block-probability (item 2-10), Block-informed probability (item 12-20) and Block-similarity (item 22-30). Error bars denote one standard error of the mean.

Discussion

On basis of the dual-process theory, we expected normative judgments to be slower than making judgments based on heuristics. We tested this claim by inducing normative responses in people committing the conjunction fallacy and by comparing the response latency between these different types of judgments.

In our first analysis we investigated the response accuracy on the whole sample. The main purpose of this analysis was to test whether it was possible to induce accurate responding in participants that had started committing the conjunction fallacy. No prior research has tested revealing the nested set structure in a within-subject design. Thus, the rate of which the participants would embrace the
conjunction rule was an open question. In our experiment 68% of participants that had committed the conjunction fallacy hearkened the instruction and started respond correctly. No test of statistical significance was conducted but these result is similar to that was observed in Stergiadis, (2015) between-group design, which showed an improvement rate of (67%). This indicates that it is possible to interrupt the automated response and to replace it with a new one.

Secondly, looking at only those who committed the conjunction fallacy on item 1 -first item-probability- avoided the conjunction fallacy on item 11 –first item informed probability- and committed the conjunction fallacy on item 21 –first item similarity- we observed differences between item 1 and item 11, in which item 11 was significantly slower. This indicates that the very first heuristic judgment is faster than the first normative judgment. This could be seen as a replication of De Neys, (2006) between groups design, which used a one-shot decision paradigm. Their conclusion was that correct analytic inferences demanded more processing time than did those participants making heuristic inferences. Although to support this conclusion we should have observed a difference in response time between item 11 and item 21 but no such tendency was observed in this study.

In the third analysis -block condition response time- we included only those that displayed response patterns that was expected in each condition. Our data showed no differences in time course between probability judgments that violated the conjunction rule, normative probability judgments or judgements of similarity. This disproved our main hypothesis that correct responding would require more time consuming system 2 processes. It contradicts earlier observations (De Neys, 2006) and goes against predictions derived from the dual-process framework.

In summary, although the analysis of the first items in each condition seems to support findings in previous studies of probability judgments, when we analyzed repeated trials and mean response times the results where actually contradictory. The findings in this study therefore indicates that normative probability judgments may not be a system 2 operation.

Recently, Bago and De Ney's, (2017) published a study that is in accordance with these conclusions. Through a series of experiments, they constructed a paradigm which allowed them to compare response accuracy between system 1 and system 2 operations. They frequently observed correct, logical answers in the immediate response and concluded that following the intuitive responses can lead to normative probability judgments.

One alternative explanation for the observations made in this study, which also may account for its biggest limitations, is the issue about statistical power. Only 13 participants out of 24 could be included in the main analysis –this makes for an attrition rate of 46% - and there is a lack of prior knowledge about the expected effect sizes in the mechanisms of fast and slow processes. The theoretical framework does not depict any absolute values of these attributes, only relations. Namely, either it is faster or it is slower. Possibly, we are looking for small effect
sizes which would need a much bigger sample size to make sure not committing a type 2 error.

Another limitation in our study, which could question the validity of the first item-analysis, is that the first item in the probability condition actually was the trial item, item nr 0. This may affect the timing for the first item in the condition, making the response time faster because the participants had one trial before to train the response.

Gaining knowledge about the circumstances in which humans tend to make good decisions and when to expect biased judgments has many obvious implications in daily life situations. One important part in contributing to this knowledge bank is the development of its theoretical underpinnings. The dual-process theory of human cognition is an attempt to give a comprehensive model of the human judgment. In order to either strengthen or weaken its position as a useful model of the mind, it's claims has to be supported or falsified with empirical data. This thesis contributes to that.

References


Hertwig, R., & Gigerenzer, G. (1999). The “conjunction fallacy” revisited: How intelligent inferences


Let’s take a break. Please read the following text. You don’t have to hurry, there is no time limit.

Imagine buying a bag of wine gums with almost exclusively black candies – improbable, but not impossible. When picking a candy at random it is probable that you will get a black candy.

However, if someone would ask you whether it would be more probable to pick a black candy or a candy, the correct alternative would be a candy. That alternative contains all subcategories – red candy, green candy, yellow candy and so on – but also black candy, which was the other alternative. As black candy is a subcategory to candy, it can never be more probable to randomly pick a black candy than to pick a candy.

For the same reason, it is always more probable that a person is “tall” than “tall and thin”: to say that a person is tall includes all subcategories -tall and thin, tall and fat, tall and normal weighted etc.

For the same reason, it is as well always more probable that a person is “blond” than that it is “blond with blue eyes”: to say that a person is blond includes all subcategories – blond with blue eyes, blond with brown eyes, blond with green eyes.

In some, but not all of the tasks in this study, you will be able to utilize the information above.

Now the task is changed. Earlier you did probability judgments. Now we want you to do similarity judgments. Similarity judgments does not follow the same rules as probability judgments. Don’t care about the instructions you got in the previous break.

If you imagine a tall person, maybe you tend to also think that the person is thin. If I would ask you to judge whether the category “tall” or “tall and thin” is more similar to the picture in your imagination, the latter would not be considered wrong.

In the same way, imagine a person that is blond. If I would ask you to judge whether “blond” or “blond with blue eyes” is more similar to your pictured person. I could not say that the latter is wrong. That’s up to you. When I ask you to make judgments about similarity, there is no right or wrong.

Now, your task is to judge which of the following four alternatives that are most similar to the person that you picture when reading the description.