When Does an Individual Accept Misinformation?

David Borukhson (borukhd@cs.uni-freiburg.de)
Department for Computer Science, University of Freiburg, Germany

Philipp Lorenz-Spreen (lorenz-spreen@mpib-berlin.mpg.de)
Center for Adaptive Rationality, Max Planck Institute for Human Development
Berlin, Germany

Marco Ragni (ragni@sdu.dk)
Danish Institute for Advanced Studies, South Denmark University, Odense, Denmark
Cognitive Computation Lab, Technical Faculty, University of Freiburg, Germany

Abstract

A new phenomenon is the spread and acceptance of "fake news" on an individual user level, facilitated by social media such as Twitter. So far, state of the art socio–psychological theories and cognitive models focus on explaining how the accuracy of fake news is judged on average, with little consideration of the individual. This paper takes it to a new level: A breadth of core models are comparatively assessed on their predictive accuracy for the individual decision maker, i.e., how well can models predict an individual's decision before the decision is made. To conduct this analysis, it requires the raw responses of each individual and the implementation and adaption of theories to predict the individual's response. We used two previously collected large data sets with a total of 3309 participants and searched for, analyzed and refined existing classical and heuristic modeling approaches. The results suggest that classical reasoning, sentiment analysis models and heuristic approaches can best predict the "Accept" or "Reject" response of a person. A hybrid model that combines those models outperformed the prediction of all individual models pointing to an adaptive tool-box.

Keywords: Predictive modeling; fake news detection; socio-psychological theories

Introduction

Misinformation and disinformation, such as "fake news", are phenomena reaching far back in the history of the media. It typically refers to intentionally or unintentionally false information presented in a way that is deliberately or accidentally designed to mislead people. In the relatively young information ecosystem of the internet, such pieces of information can achieve particularly high spread, especially through self-organized sharing on social media (Vosoughi, Roy, & Aral, 2018). As a result, individual decisions to believe and share information have reached new importance, as they can be a micro-level driver of the scaled spread of false information and collectively even put democratic decision-making at risk (Lewandowsky et al., 2020). Especially in recent years, misinformation has returned to prominence in connection with political events such as the 2016 UK European Union membership referendum, the 2016 US presidential election (Allcott & Gentzkow, 2017), or during the ongoing COVID-19 pandemic (Cinelli et al., 2020). Numerous theories have emerged describing the spread and acceptance of misinformation. Thus it appears a timely and relevant approach to empirically evaluate some prominent theories of acceptance of news as well as more general human reasoning models on prediction accuracy: If we better understand the motives and mechanisms of susceptibility to believe misinformation on an individual level, action can be taken to reduce acceptance and spread of such news (Lazer et al., 2018; Lorenz-Spreen, Lewandowsky, Sunstein, & Hertwig, 2020).

Modeling cognitive processes has long been of interest for understanding human reasoning and many theories from different fields of psychology have been formalized into computation models (Fum, Del Missier, & Stocco, 2007). The methods used in this paper follow the premise that models need to *predict* a future output of an individual and not just reproduce data. Hence, we implement the theories in a way that they first make predictions and then are later evaluated on the data (being previously trained on a different data set).

Susceptibility to misinformation in news items has been studied with different approaches describing spread (Del Vicario et al., 2016) and acceptance (Rampersad & Althiyabi, 2020). Some studies refer to rather simple implementation of cognitive reasoning models based on the correlation of measured features (Pennycook & Rand, 2019), however as experimental data on news reasoning tasks have only recently become available, there appears to be a dearth of empirical quantification for comparing decision making models in this domain. In this paper, we fill this gap by systematically comparing the predictive power of different influential theories of decision making on an experimental dataset covering news acceptance decisions.

The experimental datasets were collected and published by Pennycook and Rand (2019) and contain information on accuracy judgements of a number of "fake news" and real news items, as well as about the individual test participants. Participant–specific data from these sets can be used to test news item reasoning hypotheses put forth by Pennycook and Rand (2019) about motivated reasoning, and compare them to other, more general heuristic theories in the tradition of the Adaptive Toolbox of Gigerenzer and Selten (2002).

Reasoning about News and Cognitive Models

How can one formally quantify news–item related information and the effect it has on a human reasoner who is exposed to it?

Properties of the news items are known from an experimental pre-test that asked for political partisanship, other characteristics such as perceived familiarity and perceived importance of the items. The content of the pictures was not analyzed, as the primary focus is on news headline processing and introducing image recognition techniques appears to be beyond the scope. On the participant level, apart from demographic information, a score for cognitive reflection was measured for each participant.

Do some reasoning models predict news item acceptance decisions by individuals better than others? Among different classes of reasoning theories (Cognitive Models, Reasoning by Heuristics), a selection of relevant models is briefly presented. For each model, first a theoretical description is provided and then a mathematical formalization of its (expected) predictive function is presented that corresponds to the implementation of that model.

Note that while for all models there exist numerous other variants in implementation, often more complicated and with higher capacity than the ones presented, the goal of this paper is not to calculate highest–performance specifications for the given models, but to study and compare general approaches to modeling the processing of misinformation.

Experiments

The experiments used for evaluating the models were conducted by Pennycook and Rand (2019). They comprise accuracy judgements of participants about individual news items, consisting of an image and a headline, a shape in which they typically appear in a social media environment.

The methodical details for experimentation are further elaborated in Pennycook, Binnendyk, Newton, and Rand (2020).

Experiment 1

This experiment was completed as part of a study that took place in 2017 (Pennycook & Rand, 2019). 843 participants (763 with completed data) were recruited from Amazon Mechanical Turk. They were presented 20 partisan (10 Pro–Republican, 10 Pro–Democrat), and 10 neutral news items of which 15 were real news and 15 were fake news items and asked to rate them for accuracy. News items were compiled from a fact–checking website (for false information items) and mainstream news sources (for true items). Each item consists of a short headline underneath a picture. They were presented to participants sequentially, in a random ordering. Apart from questions about demographics such as education, gender and age, participants were asked to complete a cognitive reflection test (CRT) consisting of seven questions. The CRT was devised to measure the tendency of a person to "resist reporting the first response that comes to mind" when presented with a question . The questions in a CRT typically hint at one solution that springs to mind at the first glance, but proves incorrect on second thought, e. g. : "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" While $0.10 at first instance seems to be suitable, of course the correct answer is $0.05 (Frederick, 2005). In the experiments evaluated for this paper, a variant of the CRT with seven questions for logical reasoning was used.

The primary goal of this experiment was to compare the Motivated and Classical Reasoning accounts of processing news items. For Motivated Reasoning, CRT was expected to correlate positively with acceptance of fake news items of a partisanship corresponding to that of a test participant. However, in the analysis of the experiment, such correlation was not found: Instead CRT and accurate classification of news items were found positively correlated, thus hinting a Classical Reasoning explanation.

Data was included for the 763 participants who completed all stages of the experiment.[1]

Pretest For this study, data from a pretest of 195 different persons was used, who were asked to judge on a scale to what extent an item was perceived as partisan for Republicans or Democrats ("more favorable to Democrats" vs. "more favorable to Republicans") and whether the item appeared familiar or unfamiliar to the participant.[2]

Relevant Measured Features For each task, the following features relevant to models were measured:

- Perceived accuracy of headline ("To the best of your knowledge, how accurate is the claim in the above headline", 1 to 4 scale),
- perceived familiarity of headline ("Have you ever seen or heard about this story before?", 1 to 3 scale),
- reaction time for fake/real categorization response,
- CRT value as the mean of correct responses in all 7 CRT test questions (Thomson & Oppenheimer, 2016),
- conservatism of participant (1 to 7 scale) from the pretest in two separate questions on political ideology on social and economic issues,
- perceived political partisanship of news items (1 to 5 scale: "more favorable to Democrats" to "more favorable to Republicans") from the pretest both as an absolute value and by partisanship of individuals,
- highest completed education level of the participant.

Experiment 2

This experiment was identical to Experiment 1, only the number of participants and the number and selection of news headlines used was different: 12 fake and 12 real news items were presented to each of the 2644 participants; complete data is available for 2546 participants. The study also included other questions such as about trust in media and fact–checkers, yet these are not relevant to the current models

[1] Publicly accessible dataset: https://osf.io/h2kms/

[2] Publicly accessible dataset: https://osf.io/5dsf8/

and research question. This Experiment took place in 2017 (Pennycook & Rand, 2019)[3]. **Pretest** and **Relevant Measured Features** are the same as in Experiment 1.

Modeling

We define some technical terms for clarity: A **task** is a stimulus requiring a response by the experiment participant. In the present case, a task consists of a news item together with the request to evaluate the item's correctness as it was presented to individuals during the experiment. A task is responded to by experiment participants. We refer to the combination of a task (a particular news item to categorize) and an individual it is presented to as a **trial**. Given an individual and a model that to be trained, a model's **prediction** is "Accept" or "Reject". Models attempt to maximize the number of predictions that correspond to the response in a respective trial of the same individual and task.

The **expected prediction** is a probability value between 0 and 1 that represents the model's chance to respond "Accept" on a trial. Given an individual and a task, every model m internally computes this value on the basis of a function $P_m : T \times I \rightarrow \mathbb{R}$, where T is the set of tasks and I is the set of participants. The expected prediction is then $max(0, min(P_m(t,i), 1)$ given $t \in T$ and $i \in I$, ensuring a mapping to the domain $[0,1]$. In the experiments, there is exactly one trial per person and task, so that P, expected prediction and prediction are also uniquely identified by a trial.

Before querying prediction function of a model, instantiated on the data of a participant, executes a **Pre–Training** function once. This function may optimize model parameters or data structures given data for the current individual and thus optimizes per participant. In the given setting, both Pre–Training and evaluation use data consisting of the complete set of trials per participant from the experiments.

The Dual–System–Theory and Classical Models

A Dual–System–Theory (Kahneman, 2011) essentially describes that cognition is divided into two separate classes of processes, two "systems": *System 1* activity is typically unconscious and describes intuitive processes and decision–making. Kahneman (2003) characterizes System 1 operations as "fast, automatic, effortless, associative, implicit [,] often emotionally charged; [...] governed by habit" . *System 2* accommodates intentional reasoning, such as decision–making through symbolic or logical inference. They are time–consuming, "serial, effortful" and assumed to be "deliberately controlled" by the individual reasoner (Kahneman, 2003).

Classical Reasoning

In the context of accuracy judgments about news items, the term classical reasoning as used by Pennycook and Rand (2019) and formulated by Kohlberg (1969) refers to the assumption that the extent to which people tend to think analytically, increases their likelihood to correctly classify "fake news" as misinformation and real news as real. In terms of the dual-process theory or Two–Systems View (Kahneman, 2003), measures of high system 2 activity such as the Cognitive Reflection Test should be correlated with correct classification of news items.

Implementation

$$P_{CR}(t,i) = \begin{cases} \kappa_R + \alpha_R * CRT_i & t \textit{ is real}, \\ \kappa_F + \alpha_F * CRT_i & t \textit{ is fake}. \end{cases} \qquad (1)$$

t refers to a task or trial (there is exactly one trial per task for every participant), CRT_i is the score achieved by participant i in the cognitive reflection test, addends and scaling factors κ_R, α_R, κ_F, α_F are parameters determined in Pre–Training: The equations model a linear approximation of CRT and mean participant response for real and "fake news" items, respectively. There are no free parameters. Yet notably, this model includes information on the truthfulness of the news item. Due to the correlation of truthfulness and participant responses (most items are categorized correctly) this gives it some advantage with respect to some other models, such as the heuristic presented below.

Classical Reasoning & Reaction Time

Following the Dual–System Theory account, slower responses can indicate a usage of System 2 processes which are expected to give more consciously reflected and thus accurate classifications (Kahneman, 2011).

Implementation This theory was implemented as en extension of the Classical Reasoning model: The reaction time *reac* of a person i on a given stimulus in task t is multiplied by a free parameter factor α and added to the expected response that the Classical Reasoning model yields.

$$P_{CR\&time}(t,i) = P_{CR}(t,i) + reac_{t,i} * \alpha$$

Motivated Reasoning

Motivated reasoning is related to the confirmation bias (Dawson, Gilovich, & Regan, 2002). It proposes that individuals that are "motivated to arrive at a particular conclusion [...] construct a justification for their desired conclusion" actively (Kunda, 1990). Thus, under the assumption that motivated reasoning is a System 2 activity (Motivated System 2 Reasoning, MS2R) someone who thinks analytically would tend to classify information as correct that is favorable with respect to their own opinion and tend to reject information contradicting their previous convictions. In the given setting, higher System 2 activity would thus increase likeliness to accept news items the more they seem favorable for the political party the participant supports (Pennycook & Rand, 2019).

[3] Publicly accessible dataset: `https://osf.io/f5dgh/`

Implementation A formalization of motivated reasoning must differentiate between 3 situations: The participant's partisanship corresponds with the partisanship of the news item headline, the participant's partisanship contradicts the news item partisanship or either one of these is unknown or neutral. In the latter case, the theory MS2R does not have a predictive implication; in the first two cases, the prediction respectively depends on the prevalence of analytical thinking in a participant. As the first case by MS2R is expected to yield a positive or at least stronger correlation with responding "Accept" than the second, linear parameters were introduced for both cases.

$$P_{MS2R}(t,i) = \begin{cases} \kappa_C + \alpha_C * CRT_i & part_i = part_t, \\ \kappa_N + \alpha_N * CRT_i & part_i \neq part_t, \\ 0.5 & part_i \text{ } or \text{ } part_t \text{ } unclear. \end{cases}$$

C, N stands for matching or non–matching partisanship of test participant and presented news item: $C =$ confirming view, $N =$ contradicting view of news item with respect to the persons political orientation. $part_t$ and $part_i$ scores (partisanship of news item and participant) are determined in experimental pretesting, which divided news items in neutral, favorable for Democrats and favorable for Republicans and questioned participants about their political orientation. CRT_i is the cognitive reflection test score achieved by individual i. Addends κ and CRT scaling factors α are free parameters.

Weighted Sentiments

We included a linear combination of sentiments that were globally optimized as an additional model. The sentiment analysis was conducted using the `Empath` library for Python (Fast, Chen, & Bernstein, 2016) which assigns words in a text to pre-built categories. These are generated by deep learning methods over a large volume of text from modern fiction. To avoid overfitting to headlines, only such sentiment categories were considered that had an occurrence count of 4 or higher on the concatenation of all headlines.

Implementation

$$P_{SENTIMENTS}(t,i) = \begin{cases} \sum_{c=1}^{n} s_c(t) * \alpha_{c,i} & \geq 0, \quad return \ 1 \\ \sum_{c=1}^{n} s_c(t) * \alpha_{c,i} & < 0, \quad return \ 0 \end{cases}$$

$\alpha_{c,i}$ are free weighting parameters for a set of n sentiment measures $s_1(t) \ldots s_n(t)$ per headline of task t and for participant i. 10 sentiments features were weighted for each headline. So, this model is a linear combination of the sentiments.

Heuristics for Reasoning

Reasoning with heuristics as prominently treated in the Adaptive Toolbox by Gigerenzer and Selten (2002), means using a set of simple, yet comparatively high performing ("satisficing") rules. This bounded rationality approach was originally devised to facilitate high–risk decisions under time pressure; it acts both as a cognitive model and as an assistance tool for decision making (Luan, Schooler, & Gigerenzer, 2011).

Recognition Heuristic

This approach states that a stimulus is more likely be chosen or accepted by a person, if the person knows something about it, even if the information is not causally related to a reason to accept the stimulus. Here we assume that recognition values correspond to perceived familiarity measures of a news item from the pre–test; this too was suggested in a study by Schwikert and Curran (2014).

Implementation A formalization of recognition should yield "Accept" values for more familiar stimuli and "Reject" values for others; both a threshold model and a linear combination model seem plausible. Variants 1 (threshold) and 2 (linear):

$$P_{RECOG}(t,i) = 1_{FAM_t > \kappa_i}$$

$$P_{RECOG}^{linear}(t,i) = \kappa_i + \alpha_i * FAM_t$$

FAM_t is the perceived familiarity of stimulus t measured in a pretest, κ_i and α_i are free parameters.

Fast–and–Frugal Decision Trees

Fast–and—Frugal decision trees (FFTs) are intentionally simple, binary decision trees, where each node is connected to an output (Martignon, Vitouch, Takezawa, & Forster, 2003). They are used in various disciplines for categorizing an object in order to make decisions with relatively little information, making them easy to construct and execute and thereby a class of heuristics (Martignon, Katsikopoulos, & Woike, 2008; Raab & Gigerenzer, 2015). Figure 1 shows an example FFT for a decision problem involving conditions on 3 features.

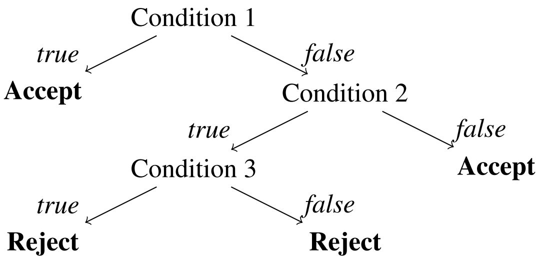

Figure 1: Example of a Fast-and-Frugal Decision Tree

There are multiple strategies for selecting the ordering of features used for conditions and the respective direction of exits. Some strategies in the literature like ifan or dfan (Phillips, Neth, Woike, & Gaissmaier, 2017) aim for optimal accuracy and they have been originally designed to facilitate decision making in situations of limited time. Others such as Max or ZigZag (Martignon et al., 2003, 2008) are derived from the Take–the–Best heuristic and optimize for best predictive performance; they do not consider conditional probabilities when selecting cues but only a greedy estimating measure of information gain.

Implementation FFT generation algorithms were implemented following the specifications by Martignon et al. (2008) for Max and Woike, Hoffrage, and Martignon (2017) for ZigZag (Z+ variant). They do not involve a depth limit. The features used per trial are all those listed as "relevant measured features" in the experiment description.

Algorithm 1 Fast–and–Frugal Tree Generation: Max

```
Given: A training set of stimuli with a set $C$ of numerical or categorical features
for feature $c \in C$ do
    calculate $v_c^1 = \frac{\textit{\#items correctly classified to “Accept” by cue}}{\textit{\#total items classified to “Accept” by cue}}$ for a cue to decide in $c$
    calculate $v_c^0 = \frac{\textit{\#items correctly classified to “Reject” by cue}}{\textit{\#total items classified to “Reject” by cue}}$ for a cue to decide in $c$
    calculate $rank(c) = max(v_c^1, v_c^0)$
order list of features $c \in C$ by decreasing value of $rank(c)$
        ▷ use this order to incrementally construct FFT from root to leaves
for $c \in C_{ordered}$ do
    if $rank(c) = v_c^1$ then
        the cue of $c$ leads to an “Accept” exit
    else ($rank(c) = v_c^0$)
        the cue of $c$ leads to a “Reject” exit
```

The ordering process in the for loop of Algorithm Max is considered an application of the Take–The–Best heuristic (Martignon et al., 2008). But Algorithm Max may yield "rake"–structured trees (trees where the ratio of "Accept" and "Reject" exits is strongly imbalanced) that might be unlike to cognitive representations. To avoid these, ZigZag enforces a binary alternating order of "Accept" and "Reject" exits. ZigZag uses the same Take–The–Best measure to determine feature/cue ranks, but the ordering of features is only the secondary specification for the resulting zig–zag shaped FFT.

Predictive Accuracy of Models

Our evaluation approach focuses on testing, how accurate models predict a response of each individual participant (Ragni, 2020). This allows to possibly falsify models and compare their predictive performance. To ensure a modeling evaluation standard, we used the CCOBRA-framework[4] that has been recently proposed (Ragni, Riesterer, & Khemlani, 2019). It automatically generates distinct test and training data and provides the models the same experimental test scenario, participants have been presented with. In the analyzed studies participants were presented with news item headlines and expects a response from the model. Then for each trial, the framework compares the participant's actual reply with the prediction given by the model. We use a "coverage" setting with coinciding pre-training and testing data, thus comparing how well models can account for individual decisions. Our model implementations and evaluation scripts are freely accessible.[5]

Evaluation

We assessed the predictive accuracy of the models. Table 1 shows predictive performance and other measures of implemented models for the datasets of Experiments 1 and 2. FFT models were not optimized per participant, as for a very limited number of trials constructing a decision tree per participant appears to be very prone to overfitting. Their Predictive Performance measures refer to a globally optimized version that nevertheless performed comparatively well.

Table 1: Predictive accuracy of models for both experiments.

Model	Predictive Performance
Hybrid Model (best)	0.79, $MAD = 0.06$
Sentiments	0.75, $MAD = 0.12$
Recognition Heuristic	0.75, $MAD = 0.12$
CR&ReactionTime	0.67, $MAD = 0.08$
Recognition Heuristic-Lin.	0.67, $MAD = 0.12$
Classical Reasoning	0.65, $MAD = 0.12$
FFT Zigzag (Z+)	0.62, $MAD = 0.19$
S2 Motivated Reasoning	0.55, $MAD = 0.05$
FFT Max	0.46, $MAD = 0.12$
Data Baselines	
Correct Categorization	0.72, $MAD = 0.12$
Always "Reject"	0.61, $MAD = 0.15$
Random	0.50, $MAD = 0.00$
Two-Model Hybrids	
Recognition & Sent.	0.79, $MAD = 0.09$
Recogn.-Lin. & Sent.	0.75, $MAD = 0.09$
CR&ReactionTime & Sent.	0.75, $MAD = 0.09$

The first value in "Predictive Performance" refers to fitting per person: The first value shows the median predictive performance for participants and *MAD* is the respective median absolute deviation.

As seen in Table 1, most models indeed perform distinctively better than random. Interestingly, the ZigZag (Z+) FFT version achieves much better results than Max. This was also the case in the study by Martignon et al. (2008) and could have to do with lower robustness of Max. Figure 2 visualizes model performances by median over all participants.

Recommender models proved difficult to optimize, as they tended converge to a model that gathers prediction means over all participants for each item. As in 72% of trials participants categorized news items correctly, such recommenders applied this probability to each news item and as a result always yielded "Accept" for real news an "Reject" for fake news — as does the baseline model Correct Categorization.

[4] `https://github.com/CognitiveComputationLab/ccobra`

[5] `https://github.com/borukhd/Accept-Misinformation`

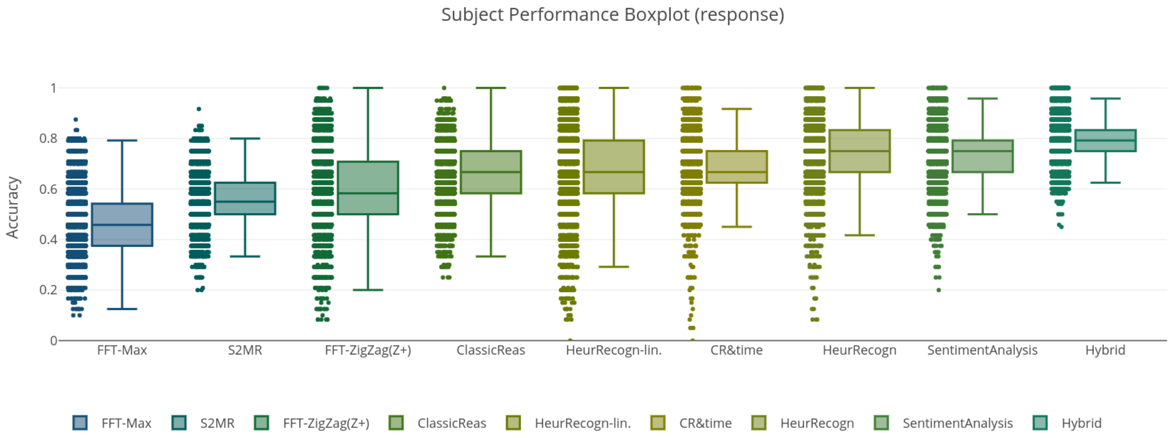

Figure 2: The predictive accuracy of each model for each individual participant (represented as a dot) in Exp. 1 and Exp. 2.

Hybrid Model

A hybrid model combines individual models by detecting for each participant p the reasoning model m that performs best and using m for predicting (respectively fitting) all trials of p. Percentage values identify the share of predictions of a model identified as the best among all participants.

The lower part of Table 1 shows three best performing two–model hybrids over outputs of all models optimized per participant. Also, an overall predictive performance of the decision of an individual participant of **0.79** with a strong leveraging effect on the $MAD = 0.06$ can be reached by an ensemble model approach. This indicates that different “cognitive tools” are employed by different participants.

Discussion and Conclusion

The results of this paper allow us to make a few assumptions about modeling misinformation processing. First, the evaluations of models CR and S2MR reflect a result achieved by Pennycook and Rand (2019); CR turns out to yield much better predictions than the motivated reasoning account. Although as expected CR outperformed motivated reasoning, yet the implemented S2MR model does have a prediction accuracy higher than random (see Table 1). An interesting finding is that a participant’s perceived familiarity with a news item appears to play a major role in judging its accuracy. The recognition heuristic relies on this measure and achieve successful predictions. This is consistent with the finding that repeated exposure to a news item increases its perceived accuracy (Pennycook, Cannon, & Rand, 2018).

Further, sentiment analysis provided interesting findings: While yielding very good results, it seems to suggest that word fields implying negative rather than positive emotions in some way receive more “Accept” responses.

Finally, the recommender optimized in a way that includes many participants rather than just a few specific ones and gets close to a model that always classifies news items correctly. This may indicate that a linear combination of the measured features does not reliably explain participants’ “Reject” vs. “Accept” classification behavior and their classification success; instead, pooling decisions of many different individuals leads to both comparatively successful prediction and often accurate classification of a news item.

In conclusion, numerous heuristic models perform reasonably well in explaining news item acceptance decisions of a participant and improving predictions of classical reasoning theory, even without information on whether the item is misinformation or not. Here, the features most significant for a participant’s acceptance decision appear to be perceived familiarity, partisanship, importance, perhaps “thrillingness” of a news item, and to a lesser extent, time spent on the decision.

Many models perform well although uncorrelated, which indicates that there may be different kinds of underlying processes in the present kind of decision making in a single individual or among groups that have not been identified yet. The improvements in predictive performance achieved with some hybrid models support this interpretation of an adaptive toolbox of strategies to evaluate news on the individual level.

Our findings also offer avenues for successful interventions to improve the accuracy of online decisions by considering the decision-making process and its context explicitly, going beyond third-party fact-checking. For example, if familiarity is an important consideration, providing related articles alongside news could be a way forward; encouraging deliberate decisions through friction or pooling judgments could also be promising avenues (Lorenz-Spreen et al., 2020). Further studies could focus on clustering participants by parameters fitted and implement more sophisticated strategies than binary decision trees for choosing among models.

Evaluating the power of models in a predictive setting is a new, rigorous, and promising method to systematically test cognitive models. At the same time, it provides a step forward to systematically construct new and better models to capture the specifics of the individual participant and automatically enriching an adaptive model toolbox.

Acknowledgements

We thank Gordon Pennycook for helpful clarifications of the datasets and providing us with data, that was not yet publicly available at the time this paper was written. Support to MR by the Danish Institute of Advanced Studies and the DFG (RA 1934/4-1, RA1934/8-1) is gratefully acknowledged. PLS acknowledges support by the Volkswagen Foundation (initiative ”Artificial Intelligence and the Society of the Future”).

References

Allcott, H., & Gentzkow, M. (2017, May). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211-36.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., … Scala, A. (2020). The covid-19 social media infodemic. *Scientific Reports*, *10*(1), 1–10.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4647–4657).

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, *8*(3), 135–142.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kohlberg, L. (1969). *Stage and sequence; the cognitive-developmental approach to socialization*.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … others (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., … others (2020). Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 1–8.

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*(2), 316.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*(6), 352–361.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. *Thinking: Psychological perspective on reasoning, judgment, and decision making*, 189–211.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. (2020). A practical guide to doing behavioural research on fake news and misinformation.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Phillips, N. D., Neth, H., Woike, J. K., & Gaissmaier, W. (2017). Fftrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making*, *12*(4), 344–368.

Raab, M., & Gigerenzer, G. (2015). The power of simplicity: a fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology*, *6*, 1672.

Ragni, M. (2020). Artificial intelligence and high-level cognition. In *A Guided Tour of Artificial Intelligence Research* (pp. 457–486). Springer.

Ragni, M., Riesterer, N., & Khemlani, S. (2019). Predicting individual human reasoning: The PRECORE-Challenge. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proc. of the 41th CogSci-Conference* (pp. 9–10). Erlbaum.

Rampersad, G., & Althiyabi, T. (2020). Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, *17*(1), 1–11.

Schwikert, S. R., & Curran, T. (2014). Familiarity and recollection in heuristic decision making. *Journal of Experimental Psychology: General*, *143*(6), 2341.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and testing natural frequencies, naïve bayes, and fast-and-frugal trees. *Decision*, *4*(4), 234.