Tracking the Eye of the Beholder: Is Explanation Subjective?

Andrew J. Stewart
University of Manchester

Henrik Singmann
University College London

Matthew Haigh
Northumbria University

Jeffrey S. Wood
Birmingham City University

Igor Douven
CNRS / Sorbonne University

Abstract

There is much recent evidence showing that explanation is central to various cognitive processes. On the other hand, philosophers have argued that the notions of explanation and explanation quality are too subjective for explanation to play any role in science: what may be an adequate explanation for one person may fail to be so for another. We compare the results of a study tasking participants with rating explanation quality with those of an eye-tracking study, finding that ratings of explanation quality from participants in the former study were strongly predictive of the ease with which participants in the latter study processed text fragments presenting the same explanations that were used in the rating study. This finding undermines the thought that explanation is only in the eye of the beholder.

Keywords: comprehension; explanation quality; explanatory coherence; explanatory reasoning; eye tracking; processing.

1 Introduction

There is much recent research demonstrating the centrality of explanation to a variety of cognitive processes, including categorization (e.g., Williams & Lombrozo, 2010), generalization (e.g., Lombrozo & Gwynne, 2014), learning (e.g., Baillargeon & DeJong, 2017; Rittle-Johnson & Loehr, 2017), understanding (e.g., Walker & Lombrozo, 2017), semantic and pragmatic processing (e.g., Hobbs, 2004; Douven, 2016; Douven, Elqayam, Singmann, & van Wijnbergen-Huitink, 2018, 2020; Mirabile & Douven, 2020), and reasoning (Douven & Schupbach, 2015; Johnston, Johnson, Koven, & Keil, 2016; Douven & Mirabile, 2018; Douven, 2019, 2020, 2021). While that may make explanation an important psychological concept, a number of influential philosophers have argued that explanation cannot serve any serious scientific purposes.

Their main objection to assigning a substantive methodological role to explanation has been that explanation is deeply subjective. According to van Fraassen (1980), explanation is inherently interest-relative and context-dependent—"explanation is in the eye of the beholder," as Lycan (2002, p. 418) summarizes van Fraassen's position. Similarly, Brandon (1998, p. 79) argues that different people may content themselves with different stopping points in explanations: where one person may be satisfied with a given explanation, another person may raise further "why" questions.

Moreover, friends and foes agree that our judgments of explanation quality are based on the so-called theoretical virtues, most notably, on coherence; for a hypothesis to be a good explanation, or to be the best explanation, depends on the extent to which it is coherent, both internally and externally, with background knowledge (Thagard, 1989; Lombrozo, 2007; Johnston et al., 2016; Walker, Bonawitz, & Lombrozo, 2017). And some philosophers have argued that the theoretical virtues lack real content and play only a rhetorical role in science (see, e.g., Giere, in Callebaut, 1993, p. 232).

As Bird (1998, p. 64) remarks, this assessment of the nature of explanation would, if correct, have dire consequences for the status of our scientific knowledge:

> [E]xplanation is closely related to inference; the standard form of scientific inference is Inference to the Best Explanation. Inference to the Best Explanation says, for instance, that when seeking the cause of a phenomenon we should look for that possible cause that best explains the phenomenon. If explanation is subjective, then that which our best explanation is will also be a subjective matter. If that is right, our inferential practices and hence the strength of our claims to scientific knowledge will also be subjective.

Thus, anyone committed to the objectivity of science will want to deny that explanation is in the eye of the beholder.

But not all share that commitment. Indeed, van Fraassen (1989) is among the fiercest critics of Inference to the Best Explanation and, accordingly, holds that our scientific knowledge is much more limited than most believe. According to him, science is not in the business of discovering theories that are *true*; it aims to give us theories that are *empirically adequate*, roughly meaning that they are predictively accurate. The mechanisms and processes *producing* the data are forever beyond our ken.

However, there is reason to hold that the above-mentioned skepticism regarding explanation is overblown. Over the past two decades, mathematicians and mathematically inclined philosophers have done much to show that the main theoretical virtues can be given rigorous formal definitions, making it hard to maintain that these virtues are empty. See, for instance, Li and Vitányi (1997) and Sober (2015) for promising formalizations of simplicity, and see Bovens and Hartmann (2003), Douven and Meijs (2007), Schippers (2015), and Koscholke (2016) for mathematically precise accounts of coherence. Schupbach and Sprenger (2011) even present a formal measure of explanatory strength directly in probabilistic terms, without invoking the notions of coherence or simplicity at all.

As to the possibility that people may disagree about how "deep" an explanation ought to be in order to qualify as satisfactory, note that it is really just that: a *possibility*. For all the critics of explanatory reasoning have shown, people tend to be satisfied at the same stopping points in an explanation. Whether this is actually the case is an empirical question, which to this date has not received the attention it deserves.

There is not just reason to doubt the criticisms leveled against explanation; there is already considerable evidence suggesting that explanation is, at a minimum, *intersubjective*. For example, Douven and Schupbach (2015), Colombo, Bucher, and Sprenger (2017), Douven and Mirabile (2018), and Mirabile and Douven (2020) elicited judgments of explanation quality from their participants, and the results these authors report show that those judgments were in broad agreement with each other.

The aforementioned studies all asked participants for their subjective judgments of explanation quality, but none of them made an attempt to relate those judgments to anything objectively measurable. In this paper, we do want to compare subjective ratings of explanation quality obtained from participants in one study to an objective measure in participants in another study. Specifically, we address the issue of the objectivity of explanation, and in particular of explanation quality, through combining the results of a study tasking participants with rating explanation quality with those of an eye-tracking study, querying whether judgments of the quality of a number of explanations obtained from one group of participants are predictive of eye-tracking data obtained from another group of participants reading text fragments presenting those same explanations. We focus on readers' eye movements as they process the explanations in the conditional form.

There is much work in the area of discourse processing examining the role of coherence in comprehension (e.g., Trabasso & van den Broek, 1985; Sundermeier, van den Broek, & Zwaan, 2005; Radvansky, Tamplin, Armendarez, & Thompson, 2014). Text that involves (for example) causal breaks, contradictions, and violations of semantic expectation results in processing disruption indexed by a number of measures including the N400, lower acceptability ratings, increases in reading time, and disruption in the eye movement record. But it is easy for sets of statements to be coherent in this sense without there being any *explanatory connections* among them, so without some of the statements making other statements in the set more *understandable*. The latter is the hallmark of *explanatory coherence* (Thagard, 1989), which nowadays is often analyzed in terms of mutual probabilistic support among statements (see above). It is reasonable to assume that discourse coherence is a necessary condition for explanatory coherence, but it is certainly not sufficient. In this paper, we focus on the effect of explanatory coherence on processing, which distinguishes our work from earlier studies that looked into the role of discourse coherence. In all of our materials below, discourse coherence (in terms of the semantic link between a target sentence and prior context) remains equivalent across conditions, while explanatory coherence (the degree to which the different discourse elements support each other) varies.

To further clarify how the focus of our studies is on explanatory coherence rather than on the more general discourse coherence, we note that the protagonists in the vignettes that were used in both studies are consistently described as *thinking* about the possible cause behind a particular event. To represent the content of a character's thoughts, readers build a mental space (Fauconnier, 1985, 1997) capturing the content of the character's reasoning about the possible explanation of the described event. If readers are sensitive to the quality of this explanatory reasoning, this should show up in the eye movement record and would suggest that readers evaluate the content of this mental space in light of knowledge that *they* have about the prior discourse. In other words, they would effectively be judging the quality of the fictional character's reasoning which would suggest that they are sensitive to explanatory coherence in the *absence* of a break in discourse coherence.

The question of how explanation quality influences online processing has, to date, not been addressed. This is an important topic to examine as explanation quality is central to explanatory reasoning; it is unknown whether sensitivity to explanation quality emerges in the eye-tracking record. To the extent that the ease with which participants processed the text fragments is an objective measure of explanation quality, and not up to their subjective judgment, our findings add to the aforementioned evidence undermining the thought that explanation is only in the eye of the beholder.

2 Study I: Rating Explanation Quality

2.1 Method

2.1.1 Participants

One hundred and seventy-six participants were recruited via the CrowdFlower platform (https://www.crowdflower.com) and 164 consented to participate. Data from 15 volunteers were excluded as they did not pass one or both of our data seriousness checks. The final sample consisted of 149 participants. Each participant was paid €0.70.

Table 1: Example Item

<table>
<tr><th>Rating task</th><th>Eye-tracking task</th></tr>
<tr><td>Debbie sees Mark pull up in a new car. His friend Debbie suspects his parents would buy him a car if he passed the recently held first year examination at college. Debbie doubts that Mark will have passed the first-year examination.
|She also knows Mark saved some money, but not enough to buy a car himself. High Fit|
|She also knows that Mark has recently received a large inheritance. Low Fit|
Debbie thinks that if
[a] Mark is driving a new car,
then
[b] he passed the first-year examination.
Mark does look good behind the wheel of the car though.
Suppose [a] is true. Then how well, in your opinion, does [b] explain [a]?</td><td>Debbie sees Mark pull up in a new car. His friend Debbie suspects his parents would buy him a car if he passed the recently held first year examination at college. Debbie doubts that Mark will have passed the first-year examination. (She also knows Mark saved some money, but not enough to buy a car himself) vs. (She also knows that Mark has recently received a large inheritance). Debbie thinks that if Mark is driving a new car, |THEN HE PASSED THE FIRST-YEAR EXAMINATION. Critical Region| MARK DOES LOOK GOOD BEHIND THE WHEEL OF THE CAR THOUGH. Spillover Region|</td></tr>
</table>

Note. The left column shows an example item indicating the two versions (High Fit / Low Fit) presented in the rating task from Study I. Participants in this study saw only one version per item. The right column shows the same item as it was presented in the eye-tracking task from Study II. The two analysis regions used in the eye-tracking task appear in small capitals and are delimited by vertical bars. The full list of items is available at: https://osf.io/vrhjg/

2.1.2 Design and materials

The experimental items consisted of 18 vignettes (see Table 1 for an example). Each vignette was six sentences long. The first three sentences provided a context. The fourth sentence manipulated the explanatory goodness of fit by varying the degree to which the situation described by this sentence made the consequent of the following conditional appear a better or worse explanation of the antecedent of the same conditional, in two levels: high explanatory fit versus low explanatory fit. The fifth sentence contained the conditional and the final sentence contained additional contextual information. For each vignette, there were two versions of sentence four, thus giving 36 different vignettes in total. The 36 vignettes were split into two Latin squared lists. For each vignette, participants then answered two questions using seven-point Likert scales. The first question asked about the acceptability of the explanation for the described event. For a conditional (e.g., “If the village has been flooded, then the dam has broken”), participants were asked to determine how well the consequent (e.g., “the dam has broken”) explained the antecedent (e.g., “the village has been flooded”), supposing the antecedent was true indeed. Secondly, participants were asked to indicate their confidence in this response. One question was inserted at the beginning of the study to assess whether participants were reading the instructions (“Below is a list of hobbies. If you are reading these instructions, please write ‘I read the instructions’ in the ‘Other’ box”). At the end of study, participants were asked if they had completed the task seriously (“Please indicate whether or not you responded seriously to the questions in this experiment”). These two questions provided us with a data quality check.

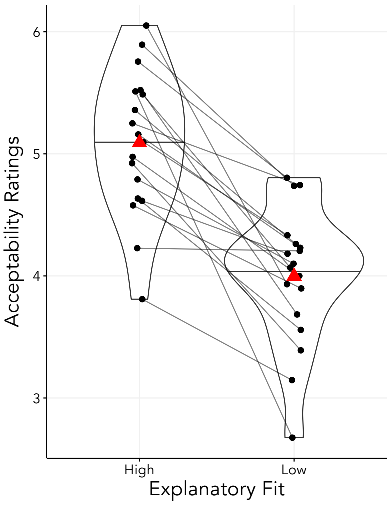

Figure 1: Participants' explanatory fit ratings as a function of the experimenter assigned explanatory fit. The violin plot shows a (mirrored) density estimate, the horizontal bar the median, the red triangle the mean. Individual points show mean ratings for each scenario and are slightly jittered on the x-axis to avoid overlap. Points of the same scenario in both fit conditions are connected by a line.

2.1.3 Procedure

After consenting to take part, participants completed the first data quality question, which assessed whether they were reading the instructions. Participants were randomly assigned to one of the two Latin squared lists. Each contained 18 vignettes and each participant saw a subset of 10 vignettes selected at random. One vignette was presented on screen at a time, with the two questions presented immediately below. A final question asked participants whether they had responded seriously.

2.2 Results and discussion

Figure 1 shows the mean acceptability ratings for each scenario as a function of the experimenter assigned explanatory fit. As can be seen, participants' responses agreed with the intended explanatory fit and were larger when a scenario was presented with a high fit than with a low fit. This pattern was also observed for each individual scenario (i.e., all slopes connecting two acceptability ratings for one scenario were negative). The acceptability ratings were highly correlated with the confidence ratings, $r = .67, p < .0001$, for the data aggregated across scenarios. Given this high degree of multicollinearity, we focus exclusively on the acceptability ratings in the following. Note that the mean acceptability ratings only ranged from 2.7 to 6.2 on the scale from 1 to 7.

However, Figure 1 also shows that within each explanatory fit condition there was some variability in the acceptability ratings. Furthermore, this variability was so strong that there were even scenarios with low fit that had higher acceptability ratings than other scenarios with high fit. Does this pattern not undermine our main claim of intersubjective agreement on the quality of explanations? The answer is *no*, as this conflates differences in the overall acceptability of the relationship expressed in the conditional (which was identical in both fit conditions) with differences induced by the explanatory fit manipulation that preceded the conditional. More specifically, some scenarios were formulated such that there were salient disablers for the relationship expressed in the conditional independent of the fit manipulation and thus overall acceptability was high, whereas other scenarios had additional salient disablers to the relationship expressed in the conditional and thus overall acceptability was generally low. As a result, these average differences are immaterial to our main claim. Furthermore, the same argument holds for average differences across participants. Some participants might generally use high values on the acceptability scale, some might use lower values, but the relevant result is that participants on average assign larger acceptability ratings to high fit explanations and lower ratings to low fit explanations.

3 Study II: Eye-tracking Task

3.1 Method

3.1.1 Participants

Twenty-four native English speakers from the University of Manchester, none of whom had taken part in the first study, and with no diagnosed language disorder, were recruited via opportunity sampling. The eye-tracking task lasted around 30 minutes and each participant was compensated with £ 5.

3.1.2 Design and materials

The items in this study were generated from the items used in Study I. There were two versions of each item (36 permutations in total). For any one item, there were both high and low explanatory fit versions for the described event. Using a Latin-square design, the vignettes were split into two lists with each participant seeing 18 experimental items. Each list also contained 9 filler passages that did not contain conditionals. Twelve participants were randomly assigned to each list. Comprehension questions followed 25 percent of the items.

3.1.3 Procedure

This study used an Eyelink 1000 in the desktop mount configuration to record eye movements during reading. During the study participants placed their head in a chin and forehead rest to stabilize their head position. Participant viewing was binocular but all recordings were only taken from the right eye which was sampled at 1,000 Hz. We presented the vignettes on a 22 inch ViewSonic monitor 60 centimeters from the participants' eyes. Each character was presented in size 22 Arial font and subtended 0.741° of visual arc. Before commencement of the study the eye-tracker was calibrated using nine fixation points presented in a grid format on the screen; calibration was also repeated as needed throughout the study. During the study itself, each trial appeared automatically following fixation upon a gaze trigger. After reading the vignette, at their normal reading rate for comprehension, participants pressed a button on a hand-held controller to reveal either a comprehension question or the next trial.

3.2 Results

We analyzed two regions of interest. These were the consequent of the conditional and the subsequent sentence, which acted as the spillover region (see Table 1). The consequent region is the earliest region

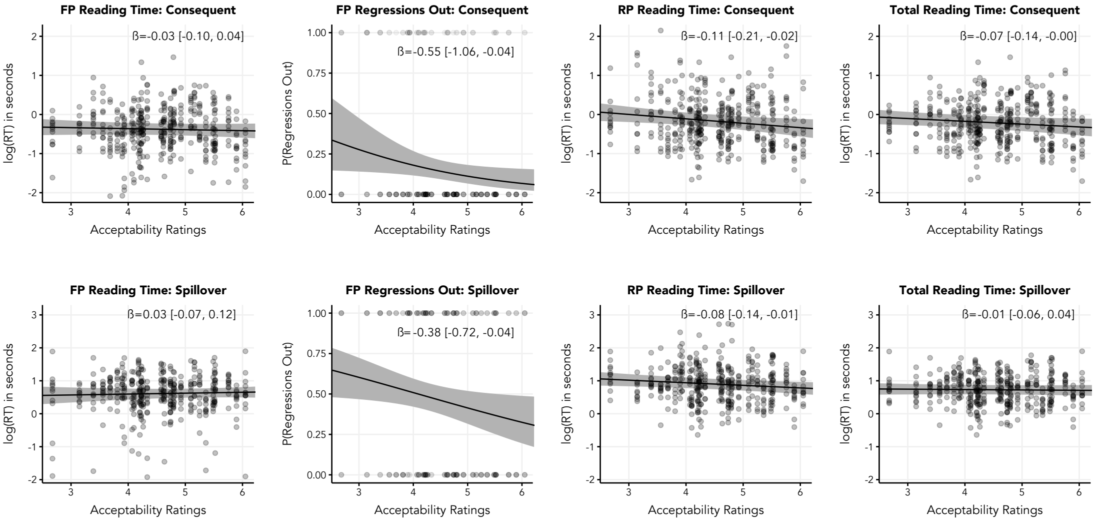

Figure 2: Eye-tracking measures as a function of acceptability ratings from Study I. The upper row shows results for the consequent region, the lower row shows results for the spillover region. Each column shows data from one eye-tracking measure. Individual data points are plotted semi-transparently such that areas with more points appear darker. The black line shows the fixed-effects estimate with 95 % confidence regions. The value in the upper right of each plot is the slope of the acceptability effect with 95 % confidence interval. Mean acceptability ratings only ranged from 2.7 to 6.2 on the scale from 1 to 7. FP = First Pass, RP = Regression Path.

of text for which we might expect to find a sensitivity to our manipulation emerging, while the spillover region is able to capture any delayed effects (Rayner, 1998). An automatic procedure pooled fixations shorter than 80 ms with adjacent fixations. It also excluded fixations that were shorter than 40 ms if they were not within three characters of another fixation, and truncated fixations longer than 1,200 ms. Due to a coding error, one item was removed from the analysis.

We analyzed four processing measures. First pass reading time is the sum of all the fixation durations in seconds from the eye first entering the region until first exiting either to the left or right. First pass regressions out is the percentage of trials in which regressive saccades were made from the current most rightward fixation into an earlier region. Regression path reading time is the sum of all fixation durations in seconds from the eye first entering a region until first exiting the region to the right (including all re-reading of previously read text). Total reading time is the sum of all fixation durations in a region in seconds. The first two measures provide information about processing as a region of text is first encountered, the third information about both early and intermediate processing, and the fourth information about early, intermediate, and later processes.

Analyses were performed using linear mixed models (LMM; Baayen, Davidson, & Bates, 2008) for the reading time measures, which were log-transformed before the analyses. This transformation resulted in approximately normal residuals. The regressions out were analyzed with binomial generalized linear mixed models (GLMM; Jaeger, 2008) with a logistic link function. In all models, mean acceptability ratings (i.e., by-item aggregated responses from the rating task in Study I) were entered as the sole fixed effect. For the random-effects, we started the analyses with a maximal structure with crossed-random effects for participants and items (Barr, Levy, Scheepers, & Tily, 2013). This entailed by-participant random intercepts and random slopes for acceptability as well as by-item random intercepts and random slopes for the binary fit measure.[1] We then iteratively simplified the random effects structure until the optimal structure was identified (Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). All analyses were performed using lme4 (Bates, Mächler, Bolker, & Aust, 2015). Tests of fixed effects were based on the Kenward–Roger approximation for LMMs and likelihood-ratio tests for the GLMMs using the methods implemented in afex (Singmann, Bolker, Westfall, & Aust, 2017). All data, analyses scripts, and detailed description of the optimal models can be found at: https://osf.io/vrhjg/.

Individual data points as well as model estimates of the fixed-effects are shown in Figure 2. Acceptability had no effect on the first pass reading times for either the consequent region, $F(1, 362.78) = 0.52$, $p = .47$, or the spillover region, $F(1, 17.11) = 0.35, p = .56$. In contrast, acceptability affected first pass regression outs for both the consequent region, $\chi^2(1) = 4.70, p = .03$, as well as the spillover region, $\chi^2(1) = 4.76, p = .03$. The lower that one group of participants judged the acceptability of the explanation related to the conditional, the more regressive saccades were observed for a different group of participants when reading the conditional. Likewise, acceptability affected regression path reading times for both the consequent, $F(1, 17.55) = 5.90, p = .03$, and the spillover region, $F(1, 368.21) = 5.78, p = .02$; lower acceptability ratings were associated with longer reading times. For the total reading time of the consequent region we also observed such a relationship, $F(1, 36.07) = 4.37$, $p = .04$, but not for the total reading time of the spillover region, $F(1, 380.81) = 0.22, p = .64$.

4 General Discussion

Our goal was to determine whether explanation quality as determined by one group of participants was reflected in the online comprehension of those explanations as measured in another group of participants. Finding evidence of such an effect would indicate that explanation quality is not subjective and at a minimum intersubjective. Across a number of eye-tracking measures, a clear and consistent pattern emerged. On first-pass regressions out during reading, we found an increase in the number of leftward regressions as explanation quality decreased. This disruption in the eye-movement record was found on the processing of explanations themselves suggesting a rapidly emerging sensitivity to explanation quality. The effect persisted during reading of the subsequent text, indicating that explanation quality has a lingering influence on comprehension. Additionally, in cases of poorer quality explanations, increased regression path times emerged during reading of the explanations themselves and on the subsequent regions of text. This pattern was also found on total reading times to explanations. Together these findings provide strong evidence of early sensitivity to explanation quality during reading. This sensitivity extends to the processing of text downstream of the explanation itself. The findings reported above add to the literature on conditional processing as they provide clear evidence that sensitivity to explanation quality emerges rapidly during online comprehension. Crucially, the pattern of data supports the hypothesis that there is strong intersubjective agreement on explanation quality, contrary to what various prominent philosophers of science had suggested. Our results also indicate that explanation quality is a key factor that influences reading times during text comprehension. While it has been known for some time that plausibility in the input affects sentence comprehension (e.g., Matsuki, Chow, Hare, Elman, Scheepers, & McRae, 2011), our data suggest that explanation quality may be a component of what constitutes plausibility at the level of

[1] Acceptability only had two values for each item, one for high fit and one for low fit. Therefore, we estimated random slopes for the categorical variable explanatory fit rather than for the continuous acceptability ratings.

discourse coherence. All in all, we have provided a decisive refutation of the claim that explanation is too subjective to be assigned any serious role in the practice of science.

We are told, time and again, that beauty is in the eye of the beholder. Eye-tracking studies have given the lie to this platitude (Valuch, Pflüger, Wallner, Laeng, & Ansorge, 2015), showing that alleged matters of taste—specifically the appreciation of facial beauty—are to an important extent objectively measurable. Our results show that something very similar pertains to explanation and explanation quality. Aggregate judgments of explanation quality obtained in one study were highly predictive of the ease with which participants in a different study processed text fragments containing the explanations used as materials in the first study. That finding would itself cry out for explanation if the judgments from the first study were subjective in the way and to the extent that various philosophers have claimed such judgments to be.[2]

References

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412. doi:10.1016/j.jml.2007.12.005

Baillargeon, R. & DeJong, G. F. (2017). Explanation-based learning in infancy. *Psychonomic Bulletin & Review* 24: 1511–1526.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv[stat]. arXiv:1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67. doi:10.18637/jss.v067.i01

Bird, A. (1998). *Philosophy of science*. Abingdon UK: Routledge.

Bovens, L. & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Callebaut, W. (1993). *Taking the naturalistic turn*. Chicago: University of Chicago Press.

Colombo, M., Bucher, L., & Sprenger, J. (2017). Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*. doi:10.3389/fpsyg.2017.01430

Douven, I. (2016). *The epistemology of indicative conditionals*. Cambridge: Cambridge University Press.

Douven, I. (2019). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence* 275: 235–251.

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science* 79: 1–14.

Douven, I. (2021). *The art of abduction*. Cambridge MA: MIT Press, in press.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology* 101: 50–81.

[2] We are grateful to Gabriel Radvansky and an anonymous referee for valuable comments on a previous version of this paper.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2020). Conditionals and inferential connections: Toward a new semantics. *Thinking & Reasoning* 26: 311–351.

Douven, I. & Meijs, W. (2007). Measuring coherence. *Synthese* 156: 405–425.

Douven, I. & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44: 1792–1813.

Douven, I. & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition* 142: 299–311.

Douven, I. & Wenmackers, S. (2017). Inference to the best explanation versus Bayes' rule in a social setting. *British Journal for the Philosophy of Science* 68: 535–570.

Fauconnier, G. (1985). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge MA: MIT Press.

Fauconnier, G. (1997). *Mappings in thought and language*. Cambridge: Cambridge University Press.

Hobbs, J. R. (1984). Abduction in natural language understanding. In L. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 724–741). Oxford: Blackwell.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59: 434–446. doi:10.1016/j.jml.2007.11.007

Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2016). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*. doi:10.1111/desc.12483

Koscholke, J. (2016). Carnap's relevance measure as probabilistic measure of coherence. *Erkenntnis* 82: 339–350.

Li, M. & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Heidelberg: Springer.

Lombrozo, T. & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience* 8. doi:10.3389/fnhum.2014.00700

Lycan, W. G. (2002). Explanation and epistemology. In P. K. Moser (Ed.), *The Oxford handbook of epistemology* (pp. 408–433). Oxford: Oxford University Press.

Matsuki K., Chow T., Hare M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37: 913–934.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 94: 305–315. doi:10.1016/j.jml.2017.01.001

Mirabile, P. & Douven, I. (2020). Abductive conditionals as a test case for inferentialism. *Cognition* 200: 104232. doi.org/10.1016/j.cognition.2020.104232

Radvansky, G. A., Tamplin, A. K., Armendarez, J., & Thompson, A. N. (2014). Different kinds of causality in event cognition. *Discourse Processes* 51: 601–618.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124: 372–422. doi:10.1037/0033-2909.124.3.372

Rittle-Johnson, B. & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review* 24: 1501–1510.

Schippers, M. (2015). Towards a grammar of Bayesian coherentism. *Studia Logica* 103: 955–984.

Schupbach, J. N. & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science* 78: 105–127. doi:10.1086/658111

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). afex: Analysis of factorial experiments. R package version 0.18-0. http://cran.r-project.org/package=afex

Sober, E. (2015). *Ockham's razors: A user's manual*. Cambridge: Cambridge University Press.

Sundermeier, B. A., van den Broek, P., & Zwaan, R. A. (2005). Causal coherence and the availability of locations and objects during narrative comprehension. *Memory & Cognition* 33: 462–470.

Thagard, P. R. (1989). Explanatory coherence. *Behavioral and Brain Sciences* 12: 435–467.

Trabasso, T. & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24: 612–630.

Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., & Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology* 6, Article 42. https://doi.org/10.3389/fpsyg.2015.00042

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

Walker, C. M. & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition* 167: 266–281. doi:10.1016/j.cognition.2016.11.007

Williams, J. J. & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science* 34: 776–806. doi:10.1111/j.1551-6709.2010.01113.x