Received: 13 December 2022 | Revised: 13 October 2023 | Accepted: 3 January 2024

DOI: 10.1111/psyp.14532

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Teleological reasoning bias is predicted by pupil dynamics: Evidence for the extensive integration account of bias in reasoning

Martin Jensen Mækelæ[1] | **Isabel V. Kreis**[1,2] | **Gerit Pfuhl**[1,3]

[1]Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway

[2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[3]Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Martin Jensen Mækelæ, Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway.
Email: maekelae.m.j@gmail.com

Abstract

Teleological reasoning is the tendency for humans to see purpose and intentionality in natural phenomena when there is none. In this study, we assess three competing theories on how bias in reasoning arises by examining performance on a teleological reasoning task while measuring pupil size and response times. We replicate that humans ($N = 45$) are prone to accept false teleological explanations. Further, we show that errors on the teleological reasoning task are associated with slower response times, smaller baseline pupil size, and larger pupil dilations. The results are in line with the single-process extensive integration account and directly oppose predictions from dual-processing accounts. Lastly, by modeling responses with a drift-diffusion model, we find that larger baseline pupil size is associated with lower decision threshold and higher drift rate, whereas larger pupil dilations are associated with higher decision threshold and lower drift rate. The results highlight the role of neural gain and the Locus Coeruleus–Norepinephrine system in modulating evidence integration and bias in reasoning. Thus, teleological reasoning and susceptibility to bias likely arise due to extensive processing rather than through fast and effortless processing.

KEYWORDS

decision-making, drift-diffusion model, dual-process, extensive integration, Locus Coeruleus, neural gain, norepinephrine, pupillometry, reasoning bias, teleological reasoning

1 | INTRODUCTION

Human reasoning and decision-making are prone to bias. A salient example is the tendency to see purpose and intentionality in natural phenomena when there is none. This is known as teleological reasoning (Kelemen et al., 2013). As with other well-documented reasoning biases, what causes this non-normative reasoning remains elusive (Kelemen, 1999). In this paper, we assess three competing theories on how bias in reasoning arises by examining performance on a teleological reasoning task while measuring pupil size and response times.

Teleological reasoning is seen early in children's reasoning development as an explanatory default (DiYanni & Kelemen, 2005). This bias is so persistent that even physical scientists have been shown to endorse false

teleological explanations, such as "Trees produce oxygen so that animals can breathe." under time pressure (Kelemen et al., 2013). It is proposed that teleological reasoning remains a cognitive default throughout life (Kelemen et al., 2013). Teleological beliefs may be replaced later in life with scientific normative explanations such as "Oxygen produced by trees is a by-product of photosynthesis." It is not known if this new mindware (scientific explanations) becomes intuitive knowledge for smarter individuals (Raoelison et al., 2020; Stanovich, 2018) or if teleological reasoning always needs to be suppressed by deliberative processing (Evans, 2008; Kahneman, 2011). These two explanations are in line with the Smart intuitor and Default-Interventionist dual-process models, respectively, which have been highly influential in research on bias in reasoning (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Pennycook et al., 2015; Stanovich, 2009a, 2009b). We here briefly introduce two dual-process models, the Default-Interventionist account and the Smart intuitor account. Alternative dual-process models were not included as they failed to make clear and distinct predictions from the Default-Interventionist and Smart intuitor accounts in this task (Epstein, 1994; Sloman, 1996).

1.1 | Dual-process models

At the core, dual-processing accounts state that human reasoning can be separated into two different modes of processing (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011). Type 1 processing, often called intuitive or heuristic, is automatic and does not require working memory capacity, that is, measurable features of Type 1 processing are being fast and effortless. Type 2 processing, often called analytic or deliberate, relies on working memory resources and uses mental simulation to generate responses. Measurable features of Type 2 processing are being slow and effortful. Accordingly, these processes can be gauged by measuring response times and pupil dilations, as the pupil is known to dilate with increasing cognitive effort (Hess & Polt, 1964; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018).

1.2 | Default-interventionist account

The Default-Interventionist account (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011) proposes that Type 1 processes are the default. Type 2 processes are engaged at later stages of reasoning, or not at all. The Default-Interventionist account proposes that humans are cognitive misers because their default is to conserve effort expenditure by relying on Type 1 processing. Thus, bias in reasoning occurs due to overreliance on fast effortless Type 1 processing and failure to engage in slow, effortful Type 2 processing when called for. According to the Default-Interventionist account, an intuitive teleological explanatory default produced by Type 1 processes (e.g., "Trees produce oxygen so that animals can breathe.") would have to be inhibited and overridden by Type 2 processing to produce a normative scientifically accurate explanation (e.g., "Oxygen is a by-product of photosynthesis./Trees do not produce oxygen so that animals can breathe.") when trying to understand events and phenomena. Importantly, the Default-Interventionist account predicts that overriding a false teleological explanation would require longer response times and more effort, compared to accepting a false teleological explanation which should be fast and effortless.

1.3 | The smart intuitor account

The Smart intuitor account has evolved from the Default-Interventionist account as an increasing number of studies show evidence opposing predictions from the Default-Interventionist account (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison et al., 2020; Raoelison & De Neys, 2019; Thompson et al., 2011). An example of this was shown by Raoelison et al. (2020) with a two-response paradigm for the cognitive reflection test (Frederick, 2005). The cognitive reflection test has been developed to assess an individual's ability to override an initial intuitive incorrect response in order to produce a deliberate correct response (consistent with Default-Interventionist account). However, Raoelison et al. (2020) showed that most correct responses were made fast (intuitively), and very few correct responses were due to respondents' initial wrong response followed by a correction after deliberation. Accordingly, the Smart intuitor account proposes that Type 1 processing can produce many types of intuitions which were previously believed could only arise from Type 2 processing (Bago & De Neys, 2019; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996; Thompson et al., 2018). Importantly, the Smart intuitor account proposes that high cognitive capacity individuals are more likely to answer correctly on reasoning tasks by having "better" or more accurate intuitions (Bago & De Neys, 2017, 2019; Raoelison et al., 2020). A corrective deliberate process (as proposed by Default-Interventionist) can still happen, but most correct responses in decision-making tasks are due to accurate intuitions rather than overriding faulty intuitions (Raoelison et al., 2020). The Smart intuitor account predicts then that overriding of false teleological explanations is not always necessary. Both teleological intuitions and scientifically normative intuitions can be made intuitively through a fast and effortless Type 1 process. More

generally, the Smart intuitor account predicts that both normative responses and errors can be made fast and with little effort. However, when engaging in Type 2 processing, seen by longer response times and more effort, the normative response is more likely.

To distinguish between the Default-Interventionist and Smart intuitor accounts, we included individual difference measures of cognitive ability and cognitive motivation. According to the Default-Interventionist account, engaging in Type 2 processing increases the probability of normative responses. Therefore, performance on the teleological reasoning task should be associated with higher trait cognitive motivation (Cacioppo et al., 1996; Stanovich, 2009b; Toplak et al., 2011, 2014; West et al., 2008). However, if normative responses are made intuitively by individuals high in cognitive ability as proposed by the Smart intuitor account, then cognitive ability should be associated with performance and cognitive motivation should have less influence on normative responding (Raoelison et al., 2020).

Importantly, underlying both the Default-Interventionist and Smart intuitor account is the assumption that more effortful and extensive processing (Type 2) leads to more normative responses and less bias. However, a single-process framework, the Extensive integration account, makes the opposite prediction, namely that bias in reasoning is exacerbated by more extensive processing. Recently, Eldar et al. (2021) highlighted that dual-process theories and the Extensive integration account make opposing predictions regarding pupil dilation and found support for the Extensive integration account in three framing tasks.

1.4 | Extensive integration, neural gain, and the locus coeruleus–norepinephrine system

The Extensive integration account builds on a single-process framework where decision-making is seen as a dynamic process of gradual noisy evidence accumulation and integration leading up to a decision (Busemeyer et al., 2006; Busemeyer & Townsend, 1993; Krajbich & Rangel, 2011; Usher et al., 2013; Usher & McClelland, 2004). Here, bias accumulates if the decision-making process unfolds over many time steps. Thus, a small bias will have larger effects if each piece of evidence has lower weighting and the decision requires a longer evidence accumulation process. Thus, more extensive integration is associated with more bias (Eldar et al., 2021; Usher & McClelland, 2004). Importantly, it is proposed that evidence integration is influenced by the Locus Coeruleus–Norepinephrine system, as norepinephrine modulates neural gain (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Eldar, Cohen, et al., 2016; Eldar, Niv, et al., 2016; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016). Low neural gain leads to lower weighting of each piece of evidence, and thus more extensive integration is required to reach a decision (Eldar, Cohen, et al., 2016; Eldar et al., 2013, 2021). Conversely, high neural gain leads to increased weighting of each piece of evidence. Importantly, neural gain can be gauged with pupillometry as pupil diameter is highly correlated with Locus Coeruleus activity (Aston-Jones & Cohen, 2005; Eldar et al., 2021; Gilzenrat et al., 2010; Reimer et al., 2016). Smaller baseline pupil diameter indicates low tonic Locus Coeruleus activity, low norepinephrine levels, and low neural gain (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003; Eldar et al., 2013; Eldar, Niv, et al., 2016; Joshi et al., 2016; Reimer et al., 2016). Additionally, larger pupil dilations can also indicate low neural gain as baseline pupil size and baseline-corrected pupil dilations are inversely correlated (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Gilzenrat et al., 2010). Thus, according to the Extensive integration account, bias in reasoning occurs due to more extensive evidence integration, which is exacerbated by low neural gain. Therefore, the Extensive integration account predicts that biased responses (i.e., teleological reasoning errors) are associated with longer response times and larger pupil dilations (indicating low neural gain).

In this study, we assessed which of the three accounts best explains teleological reasoning bias by evaluating performance on a teleological reasoning task. A teleological reasoning bias is evident if participants make more errors when evaluating the truth of false teleological explanations compared to comparable control statements (such as physical explanations and true teleological explanations, see methods). Both dual-process models predict that slower response times and larger pupil dilations are associated with more normative responses, that is, rejecting false teleological explanations (e.g., "Trees produce oxygen so that animals can breathe"). The Extensive integration account makes opposing predictions, namely that normative responses are associated with fast responses and smaller pupil dilations. Additionally, the Extensive integration account predicts that larger baseline pupil size is associated with normative responding.

Table 1 summarizes the predictions across the three accounts.

1.5 | Exploratory analyses and pre-registration

As an exploratory measure we recorded pupil dilations following feedback (correct or incorrect) that

TABLE 1 Predictions of the three accounts for responses in the teleological reasoning task.

Parameter	Default-interventionist	Smart intuitor	Extensive integration
Response time	Slow responses are more likely normative. Fast responses are more likely errors	Slow responses are more likely normative. Fast responses can be both normative and errors	Fast responses are more likely normative. Slow responses are more likely errors
Pupil dilation	Larger dilations are more likely normative responses. Smaller dilations are more likely errors	Larger dilations are more likely normative responses. Smaller dilations can be both errors and normative responses	Smaller dilations are more likely normative responses. Larger dilations are more likely errors
Baseline pupil size	N/A	N/A	Larger baseline more likely leads to normative responses
Cognitive ability	High ability predicts better performance (but see Stanovich and West [2008])	High ability predicts better performance	N/A
Cognitive motivation	High cognitive motivation predicts better performance	Cognitive motivation has less impact on performance than cognitive ability	N/A

Note: Predictions where the three accounts make similar predictions are not included, for example, pupil dilation to feedback (see text).

participants received after their responses in the teleological reasoning task. Pupil dilation has been linked to decision uncertainty and the following surprise after feedback (Colizoli et al., 2018; de Gee et al., 2021; Preuschoff et al., 2011; Urai et al., 2017). We expected larger pupil dilation, signaling surprise, for error trials compared to trials with correct responses. Further, we expected larger pupil dilation where decision confidence was high, but the feedback indicated being incorrect, and smaller dilations on trials where decision confidence was low. Pupil dilation to feedback cannot confirm or disconfirm any account.

Lastly, in accordance with the Extensive integration account, we modeled responses on the teleological reasoning task with an established sequential sampling model of the decision process, the drift-diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). The drift-diffusion model allows for the investigation of latent psychological processes underlying decisions (Ratcliff & McKoon, 2008; Wiecki et al., 2013). Additionally, the drift-diffusion model enables investigation of the link between psychological processes and neural mechanisms by utilizing physiological measures (i.e., pupil dilation) as predictors of parameters in the drift-diffusion model (Cavanagh et al., 2011, 2014; Wiecki et al., 2013).

Pre-registration for this study is available on OSF (https://osf.io/vk7r4/). Our pre-registered hypotheses were in line with the Default-Interventionist dual-process account. Please note, we deviate from the pre-registration as the analysis plan was found to be inadequate. Additionally, the pre-registration included plans to assess heart-rate variability; however, due to low-quality recordings (Empatica E4), these data could not be analyzed and are hence not described further.

2 | METHODS

2.1 | Participants

Participants were non-psychology students, $N=45$ (27 female), and mean age was 23.35 years (range 18–37). Participants reported not having any neurological disorder, history of brain disease or surgery, and not taking any central nervous system medication or drugs. In addition, as all test stimuli were in English and participants had different native languages, self-rated English proficiency had to be higher than 4 on a scale from 1 to 7, where 1 = "understand a few words" and 7 = "Master it like native language". The threshold was set based on a previous study showing no difference in deliberate reasoning performance between native and second language, and no effect of English proficiency on deliberate reasoning for participants scoring above 4 on the same English proficiency scale (Mækelæ & Pfuhl, 2019). All participants gave written informed consent prior to participation. The study was approved by the institutional review board at the Department of Psychology, UiT, The Arctic University of Norway. Participants received a voucher worth 400 NOK (approximately 40 USD) for participating in two test sessions (from test session two we included two cognitive ability measures in the SOM where we report the relationship between performance on the teleological reasoning task and two cognitive ability measures).

2.2 | Materials

2.2.1 | Cognitive motivation

We used the 18-item Need for Cognition Scale (NFC) (Cacioppo et al., 1984), which measures a person's

tendency to engage in and enjoy cognitively effortful activities. An example item is "*I prefer complex to simple problems*." The 18 items are rated on a 5-point Likert scale from 1 = "*Extremely uncharacteristic of me*" to 5 = "*Extremely characteristic of me*." Total score can range from 18 to 90. Internal consistency was high, McDonalds $\omega = 0.86$. The scale was implemented in Qualtrics (Qualtrics, Provo, UT).

2.2.2 | Cognitive ability

We used a composite of rational reasoning tasks to measure cognitive ability. The battery of rational reasoning tasks was created with 14 items from the heuristics and biases literature. We used items 2–7 from the Cognitive Reflection Test (Toplak et al., 2014), one fully disjunctive reasoning problem; "the marriage problem" (Levesque, 1986), one probability matching task (Koehler & James, 2010), one probability estimation task; "the bus problem" (Teigen & Keren, 2007), one making sense of medical results problem (Gigerenzer et al., 2007), one Bayesian reasoning problem (Toplak et al., 2007), adapted from Fischhoff and Beyth-Marom (1983), one covariation detection problem (Stanovich & West, 1998), one knight and knave problem (Smullyan, 1978), and one conditional reasoning problem (Lehman et al., 1988). Correct answers were scored as 1, incorrect as 0. Total composite rational reasoning score ranged between 0 and 14. The task was implemented in Qualtrics (Qualtrics, Provo, UT).

2.2.3 | Teleological reasoning

The teleological reasoning task consisted of statements containing false teleological explanations (test items), as well as control statements (control items) that participants were asked to judge as true or false (Kelemen et al., 2013; Kelemen & Rosset, 2009). There were 77 items in total, 34 of which were test items consisting of false teleological explanations for natural phenomena (e.g., "Trees produce oxygen so that animals can breathe."). The 43 control items consisted of 24 physical explanations that were either true ("Objects fall downwards because they are affected by gravity.") or false ("Soup is hot because it is primarily liquid."), and 19 control teleological explanations that were either true ("Schools exist in order to help people learn new things.") or false ("Mice run away from cats in order to get exercise."). Thus, test sentences are false teleological explanations in the domain of natural phenomena where the stated explanations are inappropriate. Control sentences are teleological explanations concerning the social–conventional and artifact domains where these explanations are appropriate.

The task was computerized with stimulus sentences presented auditorily via noise-canceling headphones. The task was self-paced, and each trial was initiated by pressing the space bar. Trials started with a fixation cross appearing on screen, and the auditory stimulus was presented after a delay of 0.5 s (see Figure 1). Stimulus sentences varied in duration between 2.3 and 3.7 s. After the stimulus sentence ended, participants had 4 s to respond, indicating whether the statement was true or false by pressing "D" or "K" on a QWERTY keyboard, respectively. Participants received feedback 1.8–2.4 s after their answer, by a "V" or "X" appearing in place of the fixation cross (feedback duration 4.0–6.2 s, uniformly jittered), representing correct and incorrect responses, respectively. If a participant did not respond within the 4 s, the trial was amended to the end of the task for repetition. All stimuli presented on screen were isoluminant. Items were pseudo-randomized with the constraint of not more than three in a row of the same type (test items or control items).

Instructions, fixation cross, and feedback for the task were presented on a monitor (width 34 cm, height 27 cm, resolution 1280 × 1024). The teleological reasoning task was programmed in Python (version 3.7) and presented in Psychopy (Peirce et al., 2019), script available on OSF (https://osf.io/vk7r4/). The auditory test stimuli for the teleological reasoning task were created by entering the

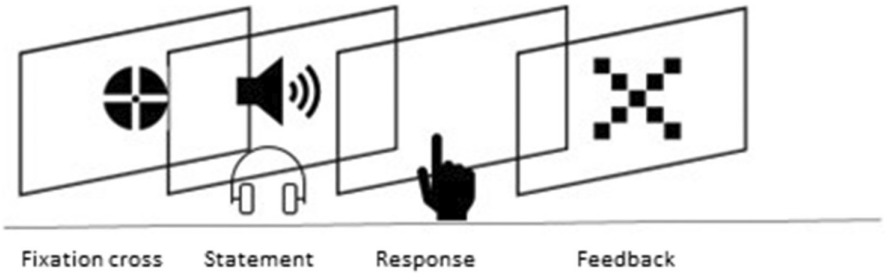

FIGURE 1 Teleological reasoning task. Trial structure of the Teleological reasoning task. Fixation cross, duration 200 ms. Statement (stimulus onset delayed by 0.5 s) presented auditorily (length 2.3–3.7 s). Feedback (onset 1.8–2.4 s after response, jittered) indicating correct and incorrect responses (here X for being wrong) presented on screen (4.0–6.2 s). Figures not to scale.

stimulus statements into Google Cloud 's speech-to-text API (Demo provided by Google Cloud online (available at: https://cloud.google.com/speech-to-text, [accessed 08.27.2019])). The resulting output was recorded with the audio recording and editing software Audacity® (Version 2.3.2, Audacity Team, 2019), audio files are available on OSF (https://osf.io/vk7r4/).

2.3 | Pupil recording

Pupil size was recorded during the teleological reasoning task with a desk-mounted Eyelink 1000 eye tracker (SR-Research, Ontario, Canada) with a sampling rate of 500 Hz. A chinrest was used to stabilize head position and viewing distance (65 cm from top of screen, 69 cm from screen bottom). A two-minute baseline measurement of pupil dilation was recorded in a sitting position in front of the computer before the teleological reasoning task started. Participants were instructed to fixate on the center of the screen.

2.3.1 | Procedure

Participants were recruited through flyers at UiT, The Arctic University of Norway. Participants were individually tested by a trained experimenter. The order of the tasks were cognitive ability, cognitive motivation, and teleological reasoning task. The test session included assessments for a separate replication project (Mækelæ et al., 2023), with a Demand selection task (Kool et al., 2010) followed by NASA task load index (Hart & Staveland, 1988) and another N-TLX assessment following cognitive ability, administered at the beginning of the session in a different room. However, these assessments are not relevant to the current study and were not expected to affect performance in any of the other assessments.

2.4 | Data processing

Data processing of pupil measurement was performed in the statistical environment R (version 4.1.2.) (R Core Team, 2021). Eyeblinks and other artifacts (rapid changes in pupil size, caused by head movements, lid flickering, etc.) were detected based on the signal's velocity (Mathot, 2018) and corrected using linear interpolation. Here, thresholds and on- and offset margins for the interpolation window were adapted on an individual basis, due to inter-individual differences in signal recovery (the speed at which the signal returns back to normal). The interpolated signal was smoothed with a 3 Hz low-pass Butterworth filter. If blinks or artifacts spanned more than 1000 consecutive milliseconds, the respective interpolated signal was treated as missing. Finally, the signal was visually screened, and trials with remaining artifacts were identified and excluded from further analysis if the artifacts occurred during time windows of interest (trial baseline, decision, and feedback; $n = 0.5$ trials per participant on average). For each trial, baseline pupil size ("Baseline pupil") was calculated as the average signal across the first 200 ms following the onset of the fixation cross. Pupil dilation during decision-making, that is, the time window from onset of the auditory stimulus until response made, and during feedback processing, that is, the time window between feedback onset and the subsequent 3000 ms, was baseline-corrected by subtracting baseline pupil from every sample within the respective time window of interest. Maximum pupil dilation during decision-making ("PDmax-BL") and during feedback ("Feedback PDmax-BL") were extracted. For decision-making, maximum pupil dilation was further calculated based on the raw signal, without prior baseline correction ("PDmax").

Baseline pupil, PDmax, PDmax-BL, and Feedback PDmax-BL measures were treated as missing (NA) if more than 50% of the signal within the respective time window were missing and/or interpolated.

2.5 | Data analyses

Linear mixed models were analyzed with the lme4 package (Bates et al., 2015). Modeling of responses on the base-rate tasks with the drift-diffusion model was performed with Python (version 3.9) (Patil et al., 2010). The model was implemented with the hierarchical drift-diffusion model, contained in the dockerHDDM (Pan et al., 2022; Wiecki et al., 2013).

First, we aimed to replicate that humans show a teleological reasoning bias. We assessed whether false teleological explanations (test condition) lead to more errors in reasoning compared to comparable explanations (control condition) by testing if there was a significant difference in accuracy between the test and control conditions.

Second, to investigate which of the three accounts best explains performance in the teleological reasoning task we applied separate generalized linear mixed models (GLMM) for response times and pupil dilations. All reported models successfully converged. We only report relevant estimates of fixed factors in the manuscript; for more details on the models, see SOM (Tables S1–S5 and S7–S10). For the pupil analysis, the main analysis is conducted with maximum pupil dilation with Baseline pupil subtracted ("PDmax-BL") as this is a common way

to report pupil dilation (Mathot, 2018), also referred to as phasic response. Further, we report analyses with Baseline pupil (also referred to as tonic response) and PDmax (uncorrected) entered separately as this is of particular interest for the Extensive integration account. We note that the latter approach may lead to multicollinearity issues; however, centering of the variables alleviates this. Assessment of variance inflation factor with the "caret" package (Kuhn, 2015) and visual inspection of the residuals with the "DHARMa" package (Hartig, 2022) showed no multicollinearity issues.

Third, to investigate how individual differences in cognitive ability and cognitive motivation influence susceptibility to false teleological explanations, we performed a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition.

Values for "Baseline pupil", "PDmax", "PDmax-BL", and "Feedback PDmax-BL" were separately *z*-scored within participants. Cognitive motivation and cognitive ability were *z*-scored across participants.

2.6 | Exploratory analyses

Pupil dilations following feedback were recorded to investigate uncertainty and surprise in the teleological reasoning task. We applied a Linear mixed model (LMM) with Feedback PDmax-BL as outcome with condition and accuracy as fixed factors.

A drift-diffusion model was applied to investigate latent psychological processes underlying decision in the teleological reasoning task and the influence of pupil dynamics. The drift-diffusion model is an established computational model of the decision process consistent with the Extensive integration account (Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). We note that the drift-diffusion model was accuracy coded, meaning the decision boundaries are correct and incorrect responses, and accordingly do not include a bias parameter.

First, we assessed whether there was a difference in the decision process when evaluating false teleological explanations compared to control statements, by testing if there were significant differences in the main parameters of the drift-diffusion model in the test and control condition. Second, pupil data were applied as a linear predictor of trial-by-trial variation in drift rate, threshold, and drift-rate variability. We ran the analyses with both "PDmax-BL" and separately entered "PDmax" and "Baseline pupil" as predictors.

For each model, we ran five Markov chains with 20,000 samples each, 12,000 of which were burn-in. Every second sample was discarded as thinning in order to reduce autocorrelation in chains. Model convergence was assessed with visual inspection of the trace, autocorrelation, the marginal posterior, and the Gelman-Rubin R statistic. All parameters had an R-hat value below 1.01. Model comparison was conducted with the deviance information criterion (DIC). Lower DIC indicates better fit. However, we note results of models with fit in similar range as DIC has limitations when comparing fit. See SOM Table S11 for comparison of all models.

2.7 | Sample size

Our sample size rationale was based on a comparable study linking pupil responses to prediction-making in environments with changing stochastic structure (de Berker et al., 2016; Kreis et al., 2023). In this study, a pupillary sensitivity measure to uncertainty correlated highly positively with performance (Pearson correlation coefficient $r = .62$, $n = 22$). Assuming some regression to the mean, we based our sample size calculation on a smaller effect size, $r = .4$, α of 0.05 (two-sided test), power of 0.8, which yielded 44 participants in the analysis (G power 3.1). Regarding individual differences, Thompson et al. (2018) report large effect sizes (η^2 of 0.3 to 0.6), and thus a sample of 40 participants would be sufficient to find an effect. Our final sample after exclusions was deemed sufficient to continue with analyses.

3 | RESULTS

A total of six participants were excluded, two by their behavioral responses (one failed to respond, one mixed up buttons), and four had too low quality of their pupil data or calibration failed, leaving a total of 39 participants (see SOM for behavioral analysis prior to exclusions by low-quality pupil data, i.e., with $n = 43$).

Descriptive statistics for all variables can be found in Table 2.

3.1 | Accuracy

To assess if participants showed a teleological reasoning bias, we compared participants' performance in the test condition to the control condition. A Mann–Whitney U test showed that the percentage of correct responses in the control condition (Mdn = 91.9, $SD = 5.4$) was significantly higher than the percentage of correct responses in the test condition (Mdn = 75.0, $SD = 14.4$), U = 1315, $p < .001$. This indicates that participants on average showed a teleological reasoning bias and endorsed false teleological explanations.

TABLE 2 Descriptive statistics.

	Mean	*SD*	Minimum	Maximum
Baseline pupil (tonic response)	32.81	4.99	21.92	51.59
PDmax	35.73	5.65	24.37	58.49
PDmax-BL (phasic response)	2.91	1.93	−1.54	14.87
Feedback PDmax-BL	1.83	2.33	−11.92	14.29
Response time in seconds	1.21	0.80	0.01	3.96
Cognitive ability	7.21	2.48	3.00	13.00
Cognitive motivation	55.10	10.19	24.00	74.00

Note: Variables not *z*-scored.

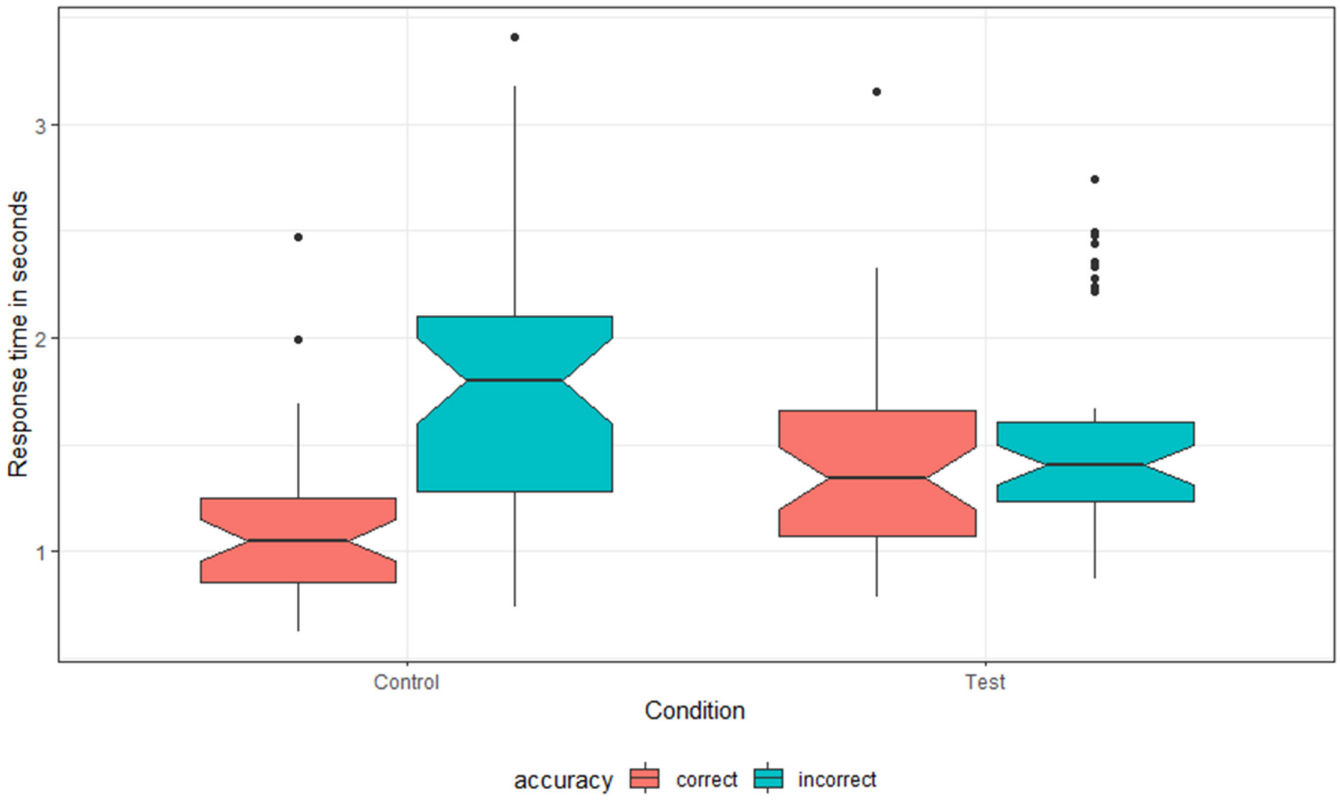

FIGURE 2 Response times separated by condition and accuracy. Response times, average per participant in seconds for the teleological reasoning task. Responses are separated by condition and accuracy.

3.2 | Response times

To assess if normative responses were associated with longer (as predicted by dual-process models) or shorter (as predicted by the Extensive integration account) response times, we applied a GLMM with accuracy as outcome (normative–error responses) and *z*-scored response times and condition as fixed factors and participants as random factors.

The results showed that correct responses were associated with shorter response times ($\beta = -0.48$, $SE = 0.06$, $z = -8.59$, $p < .001$), and that more errors were made in the test condition ($\beta = -1.18$, $SE = 0.13$, $z = -9.28$, $p < .001$), see Figure 2.

3.3 | Pupil dilation – Decision

The most important question in this study is whether errors in teleological reasoning are associated with small or large pupil dilations. The Default-Interventionist account predicts that errors occur through a fast effortless process and would therefore be associated with smaller pupil dilations. The Smart intuitor account predicts that both

errors and normative responses can be associated with small pupil dilations; however, if pupil dilations are large, the account predicts that normative responses are more likely. The Extensive integration account, on the other hand, predicts that errors should be associated with larger pupil dilations.

To test if larger or smaller pupil dilations were predictive of correct responses on the teleological reasoning task, we applied a GLMM with accuracy as outcome, PDmax-BL (phasic response) and condition as fixed factors and by-participant random intercepts (see SOM Table S6 for analysis with pupil dilation and effort as outcome).

The results showed that smaller pupil dilations were a significant predictor of normative responses ($\beta = -0.19$, $SE = 0.06$, $z = 3.15$, $p = .002$), and that participants made more errors in the test condition ($\beta = -1.30$, $SE = 0.12$, $z = -10.51$, $p < .001$). Thus, the results indicate that errors are associated with larger pupil dilations (i.e., larger phasic responses). Figure 3 shows average pupil waveform for correct and incorrect responses (see also, SOM Figure S1 for phasic response (z-scored PDmax-BL) in the time window from stimulus sentence onset until response).

Next, the Extensive integration account specifically predicts that lower baseline pupil size and larger pupil dilations are associated with more bias and thus more incorrect responses. To assess the contribution of both Baseline pupil and PDmax, we applied a GLMM with accuracy as outcome, Baseline pupil, PDmax, and condition as fixed factors and by-participant random intercepts.

The results showed that higher Baseline pupil was associated with more correct responses ($\beta = 0.24$, $SE = 0.08$, $z = 3.06$, $p = .002$). Conversely, larger PDmax were associated with more errors ($\beta = -0.21$, $SE = 0.08$, $z = -2.76$, $p = .006$), and the test condition was associated with more errors ($\beta = -1.31$, $SE = 0.12$, $z = 10.53$, $p < .001$). The results showed that errors in teleological reasoning are associated with smaller baseline pupil size (tonic response) and larger pupil dilations (phasic response).

3.4 | Individual differences

To distinguish between the Default-Interventionist and Smart intuitor account, we included individual difference measures of cognitive ability and cognitive motivation. According to the Default-Interventionist account, engaging in Type 2 thinking, thus increasing probability of normative responses, is related to trait differences in cognitive motivation. However, if normative responses are made intuitively by individuals high in cognitive ability as proposed by the Smart intuitor account, then cognitive motivation should make little difference in normative responding.

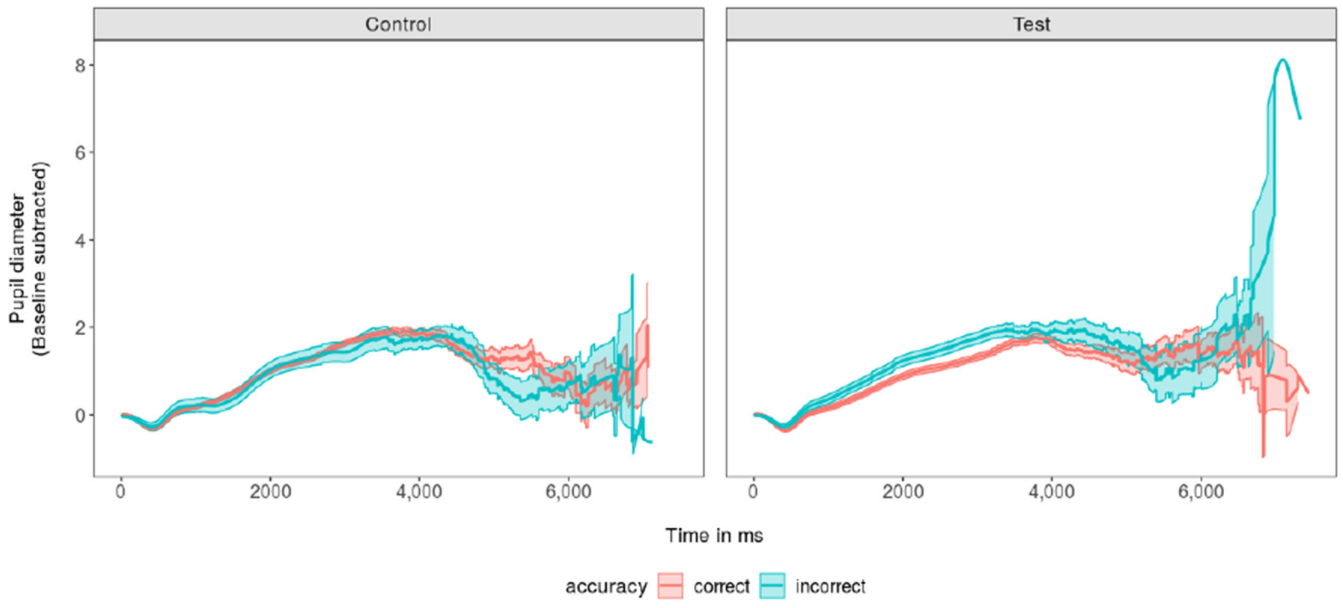

FIGURE 3 Pupil waveform for correct and incorrect responses in the control and test conditions during listening and until a response was made. Change in pupil waveform from onset of the statement until a response was made in the teleological reasoning task. Minimum duration is 2.4 s (shortest statement and immediate responding), maximum is 7.7 s. Pupil waveform is averaged across all participants and trials. Exclusions applied. Shaded area represents standard error. ms, milliseconds.

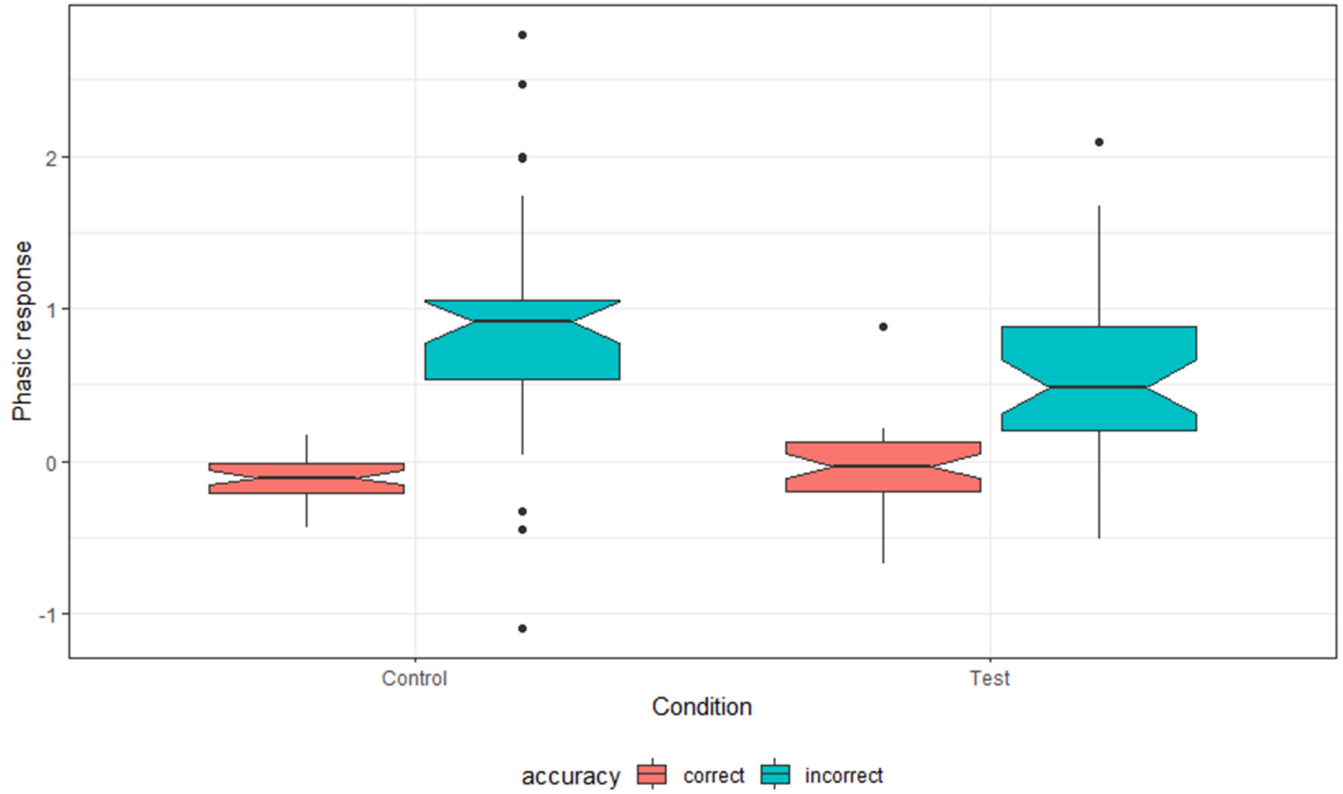

FIGURE 4 Phasic response (*z*-scored maximum pupil dilation with baseline subtracted) during feedback. Phasic response during feedback is the *z*-scored maximum pupil dilation with baseline subtracted and averaged per participant. Responses are separated by condition and accuracy.

To investigate how individual differences in cognitive motivation and cognitive ability influence performance, we conducted a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition. The model explained 28.1% of the variance in accuracy, with cognitive ability ($\beta = 0.08$, $SE = 0.02$, $t = 3.74$, $p = .001$) but not cognitive motivation ($\beta = -0.01$, $SE = 0.02$, $t = -0.38$, $p = .710$) as a significant predictor of performance in the test condition. The results show that higher cognitive ability, but not higher cognitive motivation, is associated with successfully rejecting false teleological explanations.[1]

3.5 | Exploratory analyses

3.5.1 | Pupil dilation to feedback

As an exploratory investigation we looked at pupil dilation following feedback, as pupil dilation has been known to signal decision uncertainty and surprise after feedback (de Gee et al., 2021). We interpret large pupil dilations here to indicate more surprise (see Figure 4).

To assess decision uncertainty and surprise for errors and normative responses in the two conditions, we conducted a linear mixed model with Feedback PDmax-BL as outcome and response, and condition and their two-way interaction as fixed factors and by-item[2] random intercepts.

The results yielded a significant interaction ($\beta = 0.36$, $SE = 0.12$, $t = 2.89$, $p = .004$), that is, pupil dilation was largest for incorrect responses in the control condition and smallest for correct responses in the control condition. On average, correct responses were associated with smaller pupil dilations to feedback ($\beta = -0.84$, $SE = 0.10$, $t = -8.58$, $p < .001$) compared to incorrect responses, and pupil dilations were on average larger in the control condition ($\beta = -0.31$, $SE = 0.11$, $t = -2.73$, $p = .006$) compared to the test condition. The result from the analyses of pupil dilation to feedback showed larger pupil dilations for errors, and this effect was larger in the control condition than in the test condition.

[1]SOM contains analysis for two additional measures of cognitive ability for a sub-sample of participants which participated on a separate day for a separate project.

[2]By-item random intercepts were applied as the model failed to converge when including by-participant random intercepts.

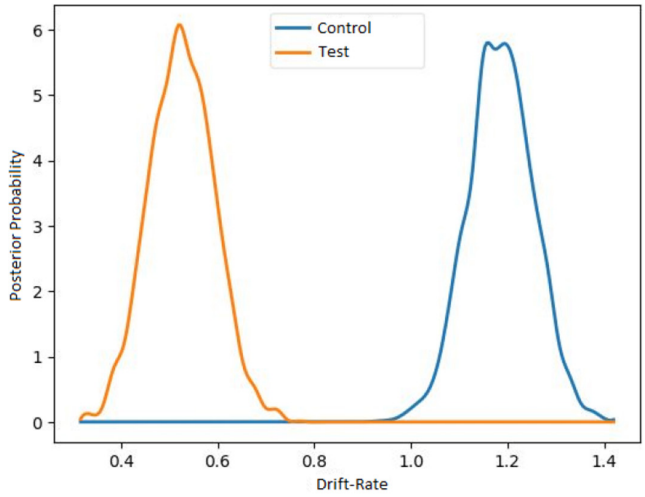

FIGURE 5 Posterior estimate of group mean drift rate in the test and control condition. Significant difference in posterior estimates of group mean drift rate in the test and control condition in the Teleological reasoning task.

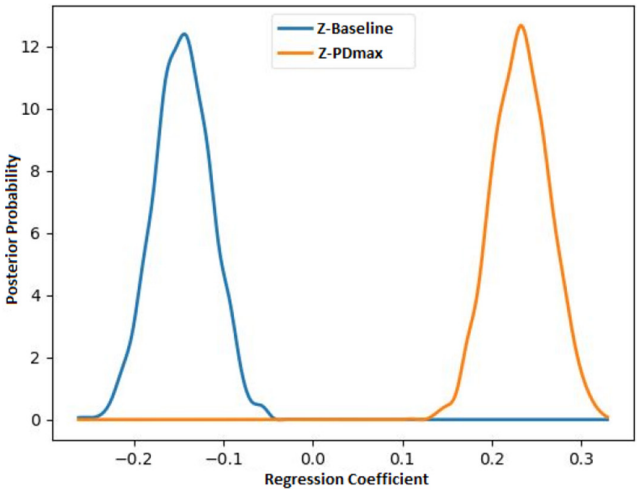

FIGURE 6 Effect of *z*-scored baseline pupil and PDmax on decision threshold. Posterior estimates of regression coefficients for *z*-scored trial-baseline pupil size and *z*-scored maximum pupil dilation as predictors of trial-by-trial variation in threshold.

3.5.2 | Drift-diffusion model

To find the model with the best fit, we analyzed the models in two steps. First, we assessed whether any of the main parameters of the drift-diffusion model differed between the test and control condition. In the second step, we assessed whether pupil measures could predict trial-by-trial variation in parameters of the drift-diffusion model.

In the first step, we found that drift rate was significantly lower in the test condition compared to the control condition (probability of drift rate in test condition being larger than mean in control = 0.01). Posterior estimates of drift rate in test and control conditions can be seen in Figure 5. Threshold was not significantly different in the two conditions (although, near significance level for the threshold being higher in the test condition), with a 0.077 probability of the mean threshold in the test condition being higher than the mean threshold in the control condition (see SOM Figure S2).[3]

In the second step, we applied pupil measures as predictors of trial-by-trial variation in parameters of the drift-diffusion model. According to the Extensive integration account, lower baseline pupil size, and thus also larger pupil dilations (as they are inversely correlated), should be linked to more extensive integration. More extensive integration in the drift-diffusion model can be achieved from either decreased drift rate (lower rate of accumulation toward decision boundary) or increased threshold (response caution) or both.

[3]Including drift-rate variability to the model or both separate threshold and drift rate was evaluated as not adding significant improvement to the model.

The winning model indicated by lowest DIC value was the model with *z*-scored Baseline pupil and *z*-scored PDmax as predictors of threshold, with separate drift rate by condition. As can be seen from Figure 6, Baseline pupil and PDmax had opposite effects on the decision threshold. Higher Baseline pupil was linked to lower threshold, whereas higher PDmax was associated with higher decision threshold.

We note that the winning model (Figure 6, DIC = 6088) showed only slightly better fit compared to the model with PDmax-BL as a predictor of drift rate (DIC = 6098) and the model with Baseline pupil and PDmax as predictors of drift rate (DIC = 6100). Importantly, the effect of pupil measures on drift rate was opposite to the effect these measures had on threshold (see SOM Figure S3). That is, higher PDmax-BL was associated with both lower drift rate and higher decision threshold (see SOM Figures S4 and S5). Posterior predictive modeling supported that PDmax-BL as a predictor of threshold had slightly better fit compared to PDmax-BL as a predictor of drift rate (see SOM Figures S7 and S8). Lastly, PDmax-BL was not related to drift-rate variability (see SOM Figure S6).

4 | DISCUSSION

The purpose of this study was to investigate theoretical frameworks that explain bias in reasoning, in particular, teleological reasoning. The participants in the study did show a teleological reasoning bias, as evidenced by their acceptance of false teleological explanations for natural phenomena at a significantly higher rate compared to errors made on comparable control statements. This is

consistent with previous studies on teleological reasoning (Kelemen et al., 2013). By modeling responses with a drift-diffusion model, we found further support for false teleological explanations being harder to evaluate as the test condition yielded a lower drift rate.

Errors in reasoning were associated with slower response times and larger pupil dilations. Further, smaller baseline pupil size and larger pupil dilations were associated with errors in reasoning. Thus, the results strongly support the extensive integration account of bias in reasoning and provide no support for dual-processing accounts.

The extensive integration account relies on a framework where decision-making is seen as a noisy sequential sampling process where evidence is accumulated over time toward decision bounds, and a response is made when the evidence reaches a decision boundary. In this task, a possible mechanism for the decision-making process is that the statement presented is compared to pieces of knowledge about the world represented in memory. This comparison results in a weighting where the probability can favor the statement being true or false. Each comparison is counted as a piece of evidence with varying strength for the statement being true or false. Evidence is accumulated over time until the relative evidence weighting is strongly favoring the statement either being true or false (accumulation reaches decision boundary), and a response is made for the favored option. A small bias favoring acceptance of teleological explanations for each piece of evidence increases the chance of accepting a false teleological explanation with more extensive accumulation. Alternatively, the mechanism through which biases arise may be weighting too heavily information that should not influence the outcome of the decision. For example, when evaluating the test statement "The sun makes light so that plants can photosynthesize" the piece of knowledge that plants use light in the photosynthesis process can bias the evaluation of the statement as a whole toward being true, when it is not. This is coherent with evidence showing that low neural gain can broaden attention, which could allow irrelevant information to influence and bias decisions (Eldar et al., 2013, 2021).

The extensive integration account further draws on research showing that the Locus Coeruleus–Norepinephrine system modulates neural gain in the brain which influences neural communication, such that when gain is high, activated neurons become more active, and inhibited neurons become less active (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003; Eldar, Niv, et al., 2016; Joshi et al., 2016). In the sequential sampling process, this means that when gain is high each piece of evidence is more heavily weighted, and fever pieces of evidence are needed to reach a decision boundary (Eldar et al., 2021; Eldar, Niv, et al., 2016). By analyzing trial-by-trial variation in pupil size with the drift-diffusion model, the results strongly support that pupil dynamics reflect changes in neural gain. Larger baseline pupil size was associated with both lower decision threshold and higher drift rate. Thus, larger (tonic) baseline pupil size, indicating higher gain, was associated with less evidence accumulation which led to faster responses and importantly, fewer errors. Conversely, larger phasic pupil dilations were associated with higher decision threshold and lower drift rate. Thus, larger phasic pupil dilations, indicating low neural gain, were associated with more evidence accumulation which led to slower response times and more errors in reasoning. Accordingly, the results corroborate predictions from the extensive integration account.

According to dual-process theories, when Type 2 processes are engaged the normative answer should be more likely. Type 2 processes are indicated by longer response times and more effort, reflected in larger pupil dilations. In this study, we found that normative responses were associated with shorter response times and less effort as reflected in smaller (phasic) pupil dilations, which contradicts dual-process predictions.

Response time in this study was limited but not speeded, that is, time was sufficient as the mean response time was more than two standard deviations below the time limit. The error rate in the test condition in this study was comparable to the error rate in the unspeeded condition in Kelemen et al. (2013). We have no indication of participants having felt time-pressured. But even if so, the speed-accuracy trade-off would have affected the test and control condition similarly (Kelemen et al., 2013).

Pupil dilation leading up to the decision was predicted by response accuracy. On one hand, higher baseline pupil size could indicate an optimal level of arousal and attention (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003). On the other hand, larger pupil dilations could reflect higher uncertainty (Colizoli et al., 2018; Preuschoff et al., 2011; Urai et al., 2017; Yu & Dayan, 2005) on subjectively more difficult trials, where errors indeed are more likely. These explanations are not mutually exclusive but describe separate processes. A less likely explanation, in a dual-process framework, explains the results by rationalization of intuitive errors (however, the authors advise against post-hoc justifications). Additionally, the results could be explained by unsuccessfully invested effort in trials where errors were made. However, there were no differences in effort by condition (see SOM Table S6 for analysis of pupil dilation/effort), which speaks against an explanation of unsuccessfully invested effort. Finding no difference by condition in pupil dilation could be explained by participants not experiencing a difference with

regard to conditions in terms of difficulty or not recognizing a need to spend more effort.

Trial-by-trial variation in pupil dilation (and trial baseline and pupil dilation separately) was associated with changes in both threshold and drift rate in the drift-diffusion model, with the model for threshold showing a slightly better fit. This is coherent with findings from Cavanagh et al. (2014) who found that pupil dilation predicted threshold and found a slightly worse fit for drift rate. Other studies have linked pupil dilation to bias and variability in drift rate (de Gee et al., 2020; Leong et al., 2021; Murphy et al., 2014). In this study, pupil dilation had no relation to variability in drift rate. Bias in drift rate was not investigated. The difference in results across studies is probably due to task differences which influence the parameters of the drift-diffusion model, as well as different influences on pupil dilation, that is, arousal, surprise, reward, uncertainty, cognitive effort, and more (Beatty & Lucero-Wagoner, 2000; Laeng et al., 2012). Considering variation in both tasks and influence on pupil dynamics, it is unlikely that pupil dilation would converge on influencing a single parameter of the drift-diffusion model. However, within the context of this study, the influence of both baseline pupil size and pupil dilation on drift rate and threshold fit the predictions from the Extensive integration account.

Feedback-evoked pupil dilations were larger for errors compared to normative responses, which is consistent with an account of pupil dilation signaling uncertainty and surprise (Colizoli et al., 2018; de Gee et al., 2021; Preuschoff et al., 2011; Urai et al., 2017). Additionally, pupil dilations to errors were larger in the control condition indicating higher degree of surprise and higher confidence in the control condition. Higher uncertainty in the test condition compared to the control condition is consistent with the results from drift-diffusion model showing lower drift rate in the test condition indicating higher stimulus difficulty. The results also reflect the behavioral finding of the test condition being more difficult than the control condition.

Individual difference measures of cognitive ability and cognitive motivation were included in the study as predictions from the Default-Interventionist and Smart intuitor accounts differed. Performance on the teleological reasoning task was associated with higher cognitive ability and not cognitive motivation, supporting the Smart intuitor account. The measures of cognitive ability (see SOM for all measures) in this study were included as a convenient indicator of cognitive ability. However, the measures have several limitations and should only be interpreted as indicators of cognitive ability. They should not be interpreted as valid measures of general intelligence. The results should therefore be evaluated with caution. Rational reasoning tasks have been used as a measure dependent on both cognitive ability and cognitive motivation (Stanovich, 2016; Trippas et al., 2015). However, recent evidence suggests performance can be explained by cognitive ability and is not related to cognitive effort (Mækelæ et al., 2023; Otero et al., 2022). We also note that sample size was low and results from individual difference measures should be considered exploratory.

4.1 | Limitations

A limitation of this study is that performance on the teleological reasoning task was not assessed both speeded and unspeeded but with a fixed 4-s time limit for responding. Participants might differ in how time-pressured they felt. Hence, we do not know participants' maximum performance, or how the decision process would unfold without any time restrictions. However, the time to evaluate the truth of statements about the world in real life may not be much longer as there are often implicit time constraints such as flow of conversation, opportunity costs, in addition to cognitive effort costs. Importantly, we do note that there is no known anatomical link between the pupil and the Locus Coeruleus, and the relationship is likely related to common downstream influences (Nieuwenhuis et al., 2011). We therefore have no direct measures of neural gain or the Locus Coeruleus–Norepinephrine system. Variation in pupil size may also be influenced by other factors.

5 | CONCLUSION

Teleological reasoning bias measured as errors in a teleological reasoning task was associated with larger pupil dilations and slower response times. The results support the extensive integration account of bias in reasoning and directly oppose predictions from dual-processing accounts.

AUTHOR CONTRIBUTIONS

Martin Jensen Mækelæ: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Isabel V. Kreis:** Data curation; software; validation; writing – review and editing. **Gerit Pfuhl:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision;

validation; visualization; writing – original draft; writing – review and editing.

FUNDING INFORMATION
None.

CONFLICT OF INTEREST STATEMENT
None.

DATA AVAILABILITY STATEMENT
https://osf.io/vk7r4/.

ETHICS STATEMENT
The study was approved by the institutional review board at the Department of Psychology, UiT, The Arctic University of Norway.

ORCID
Martin Jensen Mækelæ https://orcid.org/0000-0002-6791-1218
Isabel V. Kreis https://orcid.org/0000-0002-1022-699X
Gerit Pfuhl https://orcid.org/0000-0002-3271-6447

REFERENCES
Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J., & Lucero-Wagoner, B. (2000). *The pupillary system*. Cambridge University Press.

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*(1), 33–84. https://doi.org/10.1016/S0165-0173(03)00143-7

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks: The Official Journal of the International Neural Network Society*, *19*(8), 1047–1058. https://doi.org/10.1016/j.neunet.2006.05.043

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. https://doi.org/10.1037/0033-295X.100.3.432

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, B. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467. https://doi.org/10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and Pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Colizoli, O., de Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-31985-3

de Berker A. O., Rutledge R. B., Mathys C, Marshall L, Cross G. F., Dolan R. J., Bestmann S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*. *7*(1):10996. https://doi.org/10.1038/ncomms10996

de Gee, J. W., Correa, C. M. C., Weaver, M., Donner, T. H., & van Gaal, S. (2021). Pupil dilation and the slow wave ERP reflect surprise about choice outcome resulting from intrinsic variability in decision confidence. *Cerebral Cortex*, *31*(7), 3565–3578. https://doi.org/10.1093/cercor/bhab032

de Gee, J. W., Tsetsos, K., Schwabe, L., Urai, A. E., McCormick, D., McGinley, M. J., & Donner, T. H. (2020). Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains. *eLife*, *9*, e54014. https://doi.org/10.7554/eLife.54014

DiYanni, C., & Kelemen, D. (2005). Time to get a new mountain? The role of function in children's conceptions of natural kinds. *Cognition*, *97*(3), 327–335. https://doi.org/10.1016/j.cognition.2004.10.002

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Eldar, E., Cohen, J. D., & Niv, Y. (2016). Amplified selectivity in cognitive processing implements the neural gain model of norepinephrine function. *Behavioral and Brain Sciences*, *39*, e206. https://doi.org/10.1017/S0140525X15001776

Eldar, E., Felso, V., Cohen, J. D., & Niv, Y. (2021). A pupillary index of susceptibility to decision biases. *Nature Human Behaviour*, *5*(5), 653–662. https://doi.org/10.1038/s41562-020-01006-3

Eldar, E., Niv, Y., & Cohen, J. D. (2016). Do you see the Forest or the tree? Neural gain and breadth versus focus in perceptual processing. *Psychological Science*, *27*(12), 1632–1643. https://doi.org/10.1177/0956797616665578

Epstein S. (1994) Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*. *49*(8):709–24. https://doi.org/10.1037/0003-066X.49.8.709

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*,

59(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260. https://doi.org/10.1037/0033-295X.90.3.239

Frederick S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*. *19*(4):25–42. https://doi.org/10.1257/089533005775196732

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. *52*, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models* (Version 0.4.6) [R package]. http://florianhartig.github.io/DHARMa/

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461–468. https://doi.org/10.1016/S1364-6613(99)01402-3

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143. https://doi.org/10.1016/j.cognition.2009.01.001

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083. https://doi.org/10.1037/a0030399

Kreis I., Zhang L., Mittner M., Syla L., Lamm C., Pfuhl G. (2023). Aberrant uncertainty processing is linked to psychotic-like experiences, autistic traits, and is reflected in pupil dilation during probabilistic learning. *Cognitive, Affective, & Behavioral Neuroscience*. *23*(3):905–19. https://doi.org/10.3758/s13415-023-01088-2

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667–676. https://doi.org/10.3758/MC.38.6.667

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665–682. https://doi.org/10.1037/a0020198

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Kuhn, M. (2015). caret: Classification and regression training. *Astrophysics Source Code Library*, ascl:1505.003.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist*, *43*, 431–442. https://doi.org/10.1037/0003-066X.43.6.431

Leong, Y. C., Dziembaj, R., & D'Esposito, M. (2021). Pupil-linked arousal biases evidence accumulation toward desirable percepts during perceptual decision-making. *Psychological Science*, *32*(9), 1494–1509. https://doi.org/10.1177/09567976211004547

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*(1), 81–108. https://doi.org/10.1016/0004-3702(86)90068-8

Mækelæ, M. J., Klevjer, K., Westbrook, A., Eby, N. S., Eriksen, R., & Pfuhl, G. (2023). Is it cognitive effort you measure? Comparing three task paradigms to the need for cognition scale. *PLoS One*, *18*(8), e0290177.

Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLoS One*, *14*(1), e0211428. https://doi.org/10.1371/journal.pone.0211428

Mathot, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-linked arousal determines variability in perceptual decision making. *PLoS Computational Biology*, *10*(9), e1003854. https://doi.org/10.1371/journal.pcbi.1003854

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology*, *48*(2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614. https://doi.org/10.1016/j.intell.2021.101614

Pan, W., Geng, H., Zhang, L., Fengler, A., Frank, M., Zhang, R., & Chuan-Peng, H. (2022, November 1). A Hitchhiker's Guide to

Bayesian Hierarchical Drift-Diffusion Modeling with docker-HDDM. https://doi.org/10.31234/osf.io/6uzga

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in python. *Journal of Statistical Software*, *35*(4), 1–81.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*(6), 425–432. https://doi.org/10.1177/0963721415604610

Preuschoff, K., 't Hart, B., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for Noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*, 115.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2012.

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*, 170–178.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*(1), 13289. https://doi.org/10.1038/ncomms13289

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Smullyan, R. M. (1978). *What is the name of this book?: The riddle of Dracula and other logical puzzles*. Prentice-Hall.

Stanovich, K. E. (2009a). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

Stanovich, K. E. (2009b). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (1st ed., pp. 55–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist*, *51*, 23–34. https://doi.org/10.1080/00461520.2015.1125787

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Teigen, K. H., & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition*, *103*(3), 337–357. https://doi.org/10.1016/j.cognition.2006.03.007

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*, 945–961. https://doi.org/10.1037/xge0000457

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Toplak, M. E., Liu, E., Macpherson, R., Toneatto, T., & Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. *Journal of Behavioral Decision Making*, *20*, 103–124. https://doi.org/10.1002/bdm.544

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning*, *21*(4), 431–445. https://doi.org/10.1080/13546783.2015.1016450

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*, 757–769. https://doi.org/10.1037/0033-295X.111.3.757

Usher, M., Tsetsos, K., Lagnado, D., & Yu, E. (2013). Dynamics of decision-making: From evidence accumulation to preference and belief. *Frontiers in Psychology*, *4*, 785.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*, 930–941. https://doi.org/10.1037/a0012842

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1: Supporting Information.

How to cite this article: Mækelæ, M. J., Kreis, I. V., & Pfuhl, G. (2024). Teleological reasoning bias is predicted by pupil dynamics: Evidence for the extensive integration account of bias in reasoning. *Psychophysiology*, *00*, e14532. https://doi.org/10.1111/psyp.14532