nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02227-0

How laypeople evaluate scientific explanations containing jargon

Received: 13 July 2024

Accepted: 24 April 2025

Published online: 12 June 2025

Check for updates

Francisco Cruz [1,2] & Tania Lombrozo [2]

Individuals rely on others' expertise to achieve a basic understanding of the world. But how can non-experts achieve understanding from explanations that, by definition, they are ill-equipped to assess? Across 9 experiments with 6,698 participants (Study 1A = 737; 1B = 734; 1C = 733; 2A = 1,014; 2B = 509; 2C = 1,012; 3A = 1,026; 3B = 512; 4 = 421), we address this puzzle by focusing on scientific explanations with jargon. We identify 'when' and 'why' the inclusion of jargon makes explanations more satisfying, despite decreasing their comprehensibility. We find that jargon increases satisfaction because laypeople assume the jargon fills gaps in explanations that are otherwise incomplete. We also identify strategies for debiasing these judgements: when people attempt to generate their own explanations, inflated judgements of poor explanations with jargon are reduced, and people become better calibrated in their assessments of their own ability to explain.

The sum of humanity's knowledge cannot be contained within a single mind. Instead, knowledge is distributed across social networks, with individuals specializing in some domains and relying on their communities for others[1,2]. This 'division of cognitive labour' requires individuals to learn who they can trust across domains[3–5], and when others' knowledge is reliable[6–8]. For example, both children and adults track who is likely to know what[1,9,10], and adults memorize novel information strategically, depending on whether they expect knowledge sources to be available in the future[11].

When it comes to science, those without specialized training in a given topic (that is, laypeople) often acquire scientific information in the form of explanations from experts[12]. This presents a challenge, since (by definition) laypeople lack the training typically required to evaluate the accuracy of such explanations directly[3,13]. How, for example, might a layperson evaluate explanations for the efficacy of COVID-19 mRNA vaccines without (substantial) knowledge of molecular biology? Since scientific knowledge is increasingly technical and specialized[14], and as sources of misinformation become ever more prevalent[15], this problem becomes even more acute.

In the present research, we take up the question of how laypeople evaluate scientific explanations despite lacking expertise. We focus on a salient cue that could inform these evaluations: whether explanations include domain-specific language used by expert groups, or 'jargon'. Jargon is a defining feature of expert communities[14,16] and thus potentially informative about whether a source possesses relevant expertise. In fact, previous work finds that, under some conditions, explanations that contain scientific jargon are judged more 'satisfying'[17–32]. On the other hand, understanding jargon itself requires expertise[3,13]; hence, for laypeople, its inclusion in explanations could hinder genuine scientific understanding. Consistent with this, previous work has also found that explanations with jargon are considered less 'comprehensible'[33–39]. Jargon thus provides a microcosm for a broader puzzle of ignorance—how laypeople can gain explanatory understanding from information they are ill-equipped to comprehend.

Our first aim is to reconcile these seemingly conflicting findings from previous work: that jargon has favourable effects on the evaluation of explanations under some conditions, but detrimental effects under others. Documenting favourable effects, research in cognitive psychology has found that explanations containing jargon are often found more satisfying than equivalent explanations without jargon[19,30], especially when they are relatively poor or circular[22,25,27,31,32]. These effects are found for explanations across scientific disciplines, including math[21], neuroscience[19,26,28] and other natural sciences[20,22,26]. They are also found despite jargon's unfamiliarity[32] and reflect more than the explanation's additional length[29,32]. Jargon can have positive effects even when it merely labels a concept with a made-up word[17,18,23], leading

[1]CICPSI, Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal. [2]Department of Psychology, Princeton University, Princeton, NJ, USA.
e-mail: franciscocorreiadacruz@gmail.com

https://doi.org/10.1038/s41562-025-02227-0

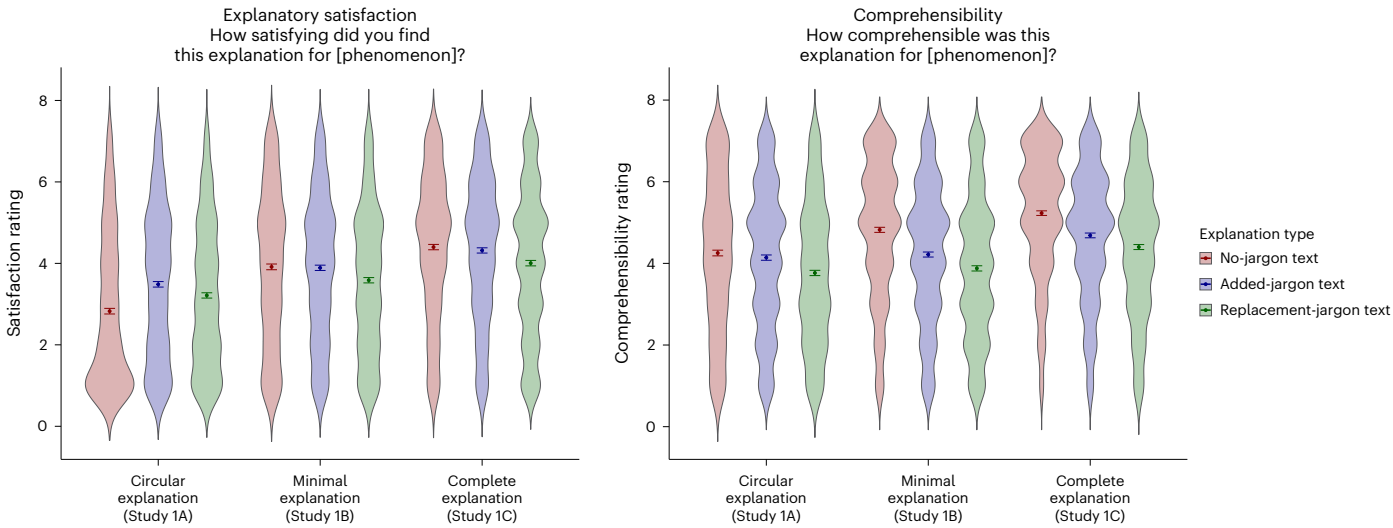

Fig. 1 | Explanatory satisfaction and comprehensibility as a function of jargon in Studies 1A–C. Violin plots depict the distribution of individual ratings and data are presented as means ± s.e.m. for $N = 737$ (Study 1A), $N = 734$ (Study 1B) and $N = 733$ (Study 1C). The key result is that circular explanations with jargon were rated more highly on explanatory satisfaction than those without; in all other cases, jargon decreases or has no effect on ratings (for both satisfaction and comprehensibility).

some researchers to argue that effects of jargon stem from inferences from the presence of jargon itself; for instance, that the explanation points to an underlying cause[23] or a community of experts[24].

Documenting detrimental effects, previous work on science communication has shown that when jargon replaces part of the text from a relatively complete explanation, perceived 'comprehensibility' decreases[33–39]. Negative effects of jargon on comprehensibility generalize across types of explanation[39], explanation source[37], domain complexity[36] and the presence of controversy[34,35,38]. Researchers have argued that these negative effects of jargon reflect the decreased processing fluency caused by lower frequency or unfamiliar words[33].

Note that research finding beneficial effects of jargon differs from that finding detrimental effects along three important dimensions: (1) whether the explanations considered are relatively poor or complete; (2) whether participants are asked to evaluate explanatory 'satisfaction' or an explanation's 'comprehensibility' and (3) whether jargon is 'added' to an explanation or 'replaces' non-jargon text. Across Studies 1A–C, we manipulate these differences systematically to reconcile the apparent contradictions from past work and to isolate the conditions under which jargon has beneficial versus detrimental effects. We find increases in explanatory satisfaction when jargon is added to poor or nearly circular explanations, but that this effect is mitigated, or even reversed, for relatively complete explanations. We also find that jargon can decrease an explanation's comprehensibility for relatively poor and complete explanations alike. These findings help us isolate the conditions under which jargon has favourable effects and refine our next question: Why is it that jargon heightens how satisfying relatively poor explanations are judged to be?

We propose that jargon is assumed to fill gaps in otherwise defective explanations.

Testing this hypothesis is the aim of Studies 2A–3B. We find that, in the absence of jargon, even laypeople can detect that circular scientific explanations contain explanatory gaps. When jargon is introduced, however, they assume the jargon fills those gaps, creating a sense of satisfaction. For instance, when learning that 'burning scented candles can cause changes in base excision repair mechanisms, leading to bladder cancer', they assume that 'base excision repair mechanisms' explain the link between burning scented candles and bladder cancer, and consequently find the explanation somewhat satisfying. This satisfaction is arguably misplaced, however, as it occurs alongside drops in reported comprehension and even when jargon is entirely unfamiliar (for instance, because it is made up, as in Studies 2B–2C).

Studies 3A and 3B thus turn to correction: What punctures inflated evaluations of explanations containing jargon? We borrow two strategies for decreasing illusions of explanatory depth[40–42], asking people to answer follow-up questions (Study 3A) or to generate explanations themselves (Study 3B). Both effectively reduce perceptions of explanation quality for explanations containing jargon, supporting our hypothesis that jargon boosts perceptions of explanation quality when it is taken to fill gaps in an otherwise defective explanation.

Finally, in Study 4, we explore downstream consequences of inflated evaluations of circular explanations: whether reading explanations with jargon makes people overestimate their own ability to 'generate' good explanations for scientific phenomena. We find that participants who receive explanations with jargon but fail to reproduce that jargon in their own explanations are especially poorly calibrated, in the sense that they fail to anticipate how good others will judge their explanations to be.

Together, these findings solve an important puzzle of ignorance, shedding light on how laypeople evaluate scientific explanations with jargon that reduces comprehensibility on the one hand, but fosters a sense of explanatory satisfaction on the other. We not only identify when and why these effects arise, but also document strategies for correction, with implications for science communication and the psychology of explanation.

Results

Preregistration and statistical approach

All studies were preregistered (https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591). Linear mixed models were conducted using R's lmerTest package (v.3.1.3)[43]; the number of experimental conditions informed contrast coding, and models included random by-subject and by-stimulus intercepts (except in Study 2C, due to the nature of its design). All relevant statistical tests were two-tailed. For simple and moderated mediations, we used SPSS's MEMORE macro (v.3.0)[44] (but PROCESS v.4.2 in Study 2C). Additional preregistered analyses are available in Supplementary Information.

Effects of jargon on circular explanations

In Study 1A, we explored the effects of jargon on 'circular explanations': explanations that simply restate the existence of the relationship being

https://doi.org/10.1038/s41562-025-02227-0

Table 1 | Linear mixed models on explanatory satisfaction and comprehensibility as a function of jargon in Studies 1A–C

	Explanatory satisfaction								Comprehensibility							
	Study 1A (circular)			Study 1B (minimal)			Study 1C (complete)			Study 1A (circular)			Study 1B (minimal)			Study 1C (complete)		
	b [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*
Predictors																		
Intercept	2.81 [2.19, 3.44]	8.83	<0.001	3.91 [3.19, 4.62]	10.75	<0.001	4.40 [3.78, 5.01]	14.11	<0.001	4.25 [3.87, 4.63]	22.04	<0.001	4.82 [4.45, 5.19]	25.34	<0.001	5.23 [5.06, 5.40]	60.11	<0.001
Added jargon (vs Control)	0.68 [0.51, 0.84]	8.07	<0.001	−0.02 [−0.18, 0.14]	−0.22	0.823	−0.08 [−0.23, 0.07]	−1.03	0.305	−0.10 [−0.26, 0.06]	−1.28	0.202	−0.60 [−0.75, −0.45]	−7.96	<0.001	−0.55 [−0.68, −0.41]	−7.73	<0.001
Replacing non-jargon (vs Control)	0.40 [0.24, 0.57]	4.79	<0.001	−0.31 [−0.47, −0.16]	−3.92	<0.001	−0.39 [−0.54, −0.24]	−5.02	<0.001	−0.48 [−0.64, −0.33]	−5.97	<0.001	−0.94 [−1.09, −0.79]	−12.39	<0.001	−0.83 [−0.97, −0.69]	−11.77	<0.001
Random effects																		
σ^2	2.61			2.36			2.20			2.43			2.10			1.82		
τ_{00}																		
Participant	0.52			0.67			0.77			0.76			0.70			0.69		
Text topic	0.29			0.38			0.28			0.10			0.10			0.01		
ICC	0.24			0.31			0.32			0.26			0.28			0.28		
N																		
Text topic	3			3			3			3			3			3		
Participants	737			733			732			737			733			732		
Observations	2,211			2,199			2,196			2,211			2,199			2,196		
Marginal/conditional R^2	0.022 /0.254			0.006/0.314			0.009/0.328			0.013/0.270			0.049/0.311			0.045/0.311		

σ^2 residual variance, τ_{00} random intercept variance; ICC, intraclass correlation coefficient; *N*, number of observations; R^2, explained variance.

explained. For instance, a circular explanation without jargon for why some candy generates sparks when it is crushed (candy triboluminescence) was: "Because candy is crushed, this can result in visible light". Participants received this explanation as is or with jargon added in one of two ways: as 'added' text ("Because candy is crushed and charged leptons collide, this can release visible light in the form of electromagnetic waves"); or as 'replacement' text (matched with the non-jargon explanation in length) ("Because charged leptons collide, visible electromagnetic waves can be released"). We expected explanatory satisfaction to be higher and comprehensibility to be lower, for explanations with (vs without jargon), replicating and extending previous results, respectively.

Consistent with these predictions (Fig. 1 and Table 1), explanations with jargon were rated more satisfying than those without jargon, regardless of how jargon was included in the explanation. Also as predicted, introducing jargon decreased explanation comprehensibility when jargon 'replaced' non-jargon content (although 'adding' jargon did not decrease comprehensibility).

Across experiments, we collected measures beyond satisfaction and comprehensibility, including perceptions of learning, confidence in the explanation's evaluation, intention to defer to experts and (as a manipulation check) perceived presence of jargon. These measures are described in the Methods section and reported fully in Supplementary Information.

Effects of jargon on more comprehensive explanations

In Studies 1B and C, we considered more comprehensive explanations. Study 1C included 'complete' explanations: explanations that detailed each step between cause and effect. For example, the complete explanation (without jargon) for candy triboluminescence described how crushing hard sugar produces static electricity, leading to discharges of energy that, after being absorbed, can be re-emitted as visible light. Study 1B included minimal explanations: an intermediate case between circular and complete explanations.

Across both studies (Fig. 1 and Table 1), we found no evidence that including jargon as additional text affects satisfaction, and we found that using jargon as replacement text 'decreased' satisfaction. In addition, adding jargon (through addition or replacement) decreased comprehensibility.

Taken together, the results from Studies 1A–C identify the conditions under which jargon shifts evaluations of explanations in more positive versus negative directions. Notably, jargon only elevates explanatory satisfaction for circular explanations, despite having negative or neutral effects on comprehensibility. Why?

For circular explanations, jargon is assumed to fill gaps

We propose that circular explanations without jargon are perceived to have explanatory gaps. When presented with authoritative jargon, however, participants assume the jargon fills perceived gaps and judge the explanation more satisfying. This hypothesis generates several predictions. First, perceived explanatory 'gappiness' should mirror effects of explanatory satisfaction, such that jargon reduces perceived gappiness for circular explanations, but has weaker or no effects for complete explanations. Second, perceived gappiness should partially or fully mediate the effect of jargon on explanatory satisfaction.

Study 2A replicated Studies 1A and C within a single design. We found the predicted interaction between jargon and explanatory completeness for explanatory satisfaction ($b = -0.81$, 95% CI [−1.09, −0.54], $t = -5.87$, $P < 0.001$; Fig. 2 and Supplementary Table 8), such that jargon increases explanatory satisfaction for circular explanations only ($b = 0.80$, 95% CI [0.60, 0.99], $t = 8.04$, $P < 0.001$; for complete explanations: $b = -0.04$, 95% CI [−0.22, 0.15], $t = -0.87$, $P = 0.709$; Supplementary Table 9). A mirrored interaction was found for judgements of explanatory 'gappiness' ($b = 0.47$, 95% CI [0.21, 0.73], $t = 3.59$, $P < 0.001$; Fig. 2 and Supplementary Table 8): circular explanations were

https://doi.org/10.1038/s41562-025-02227-0

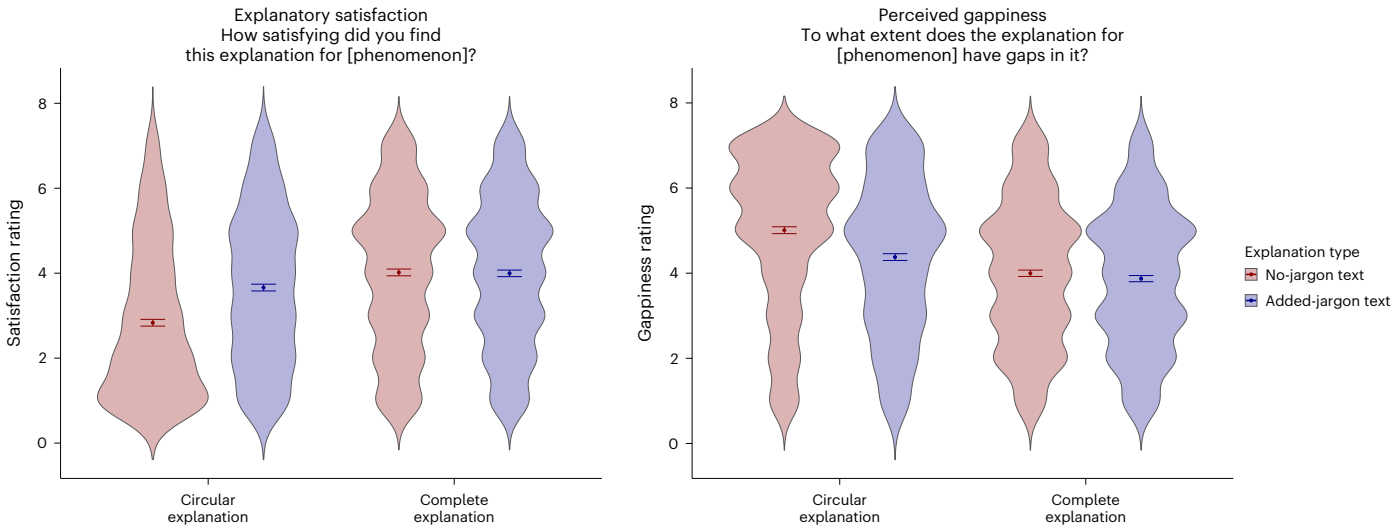

Fig. 2 | Explanatory satisfaction and perceived gappiness as a function of jargon and explanation completeness in Study 2A. Violin plots depict the distributions of individual ratings and data are presented as means ± s.e.m. for $N = 1{,}014$. The key result is that adding jargon to circular explanations increases explanatory satisfaction while decreasing perceived gappiness; this pattern is not observed for complete explanations.

perceived as less gappy when jargon was added, with a weaker effect when explanations were complete. Finally, we tested whether perceived gappiness mediated the effects of jargon on explanatory satisfaction. We found evidence for partial mediation for circular explanations (indirect effect = 0.48, 95% CI [0.32, 0.64]; direct effect = 0.37, 95% CI [0.17, 0.48]), but not for complete explanations (indirect effect = 0.11, 95% CI [−0.16, 0.39]). This difference across circular and complete explanations was itself significant (index of moderation = −0.36, 95% CI [−0.59, −0.14]).

We also considered an alternative to our explanatory gap hypothesis: that participants use jargon as a cue to source expertise, which in turn drives effects on satisfaction. Although jargon did increase perceived source expertise, it did so for 'both' circular and complete explanations (Supplementary Fig. 2, and Supplementary Tables 10 and 11).

Does jargon fill explanatory gaps because participants garner genuine understanding from the content of the jargon? Although we expected the jargon used to be largely unfamiliar, we tested this possibility more systematically in Studies 2B and C by replicating the benefits of added jargon for circular explanations with invented jargon (or pseudowords; for example, for food waste: 'hydrolysis' was replaced with 'dryholysis' or 'drylotic'). For Study 2B, we found once again that circular explanations with added jargon were judged more satisfying and less 'gappy' than those without (Supplementary Fig. 3 and Supplementary Table 13), and that the effect of jargon was fully mediated by perceived gappiness (indirect effect = 0.32, 95% CI [0.16, 0.49]; direct effect = 0.10, 95% CI [−0.03, 0.24]).

To ensure the robustness and real-world applicability of our effects, Study 2C differed from its predecessors in several ways: the explanations were presented in the form of social media posts, jargon was manipulated between participants, and the jargon manipulation check item was removed to ensure that participants' attention was not drawn to the jargon artificially.

Despite these changes, Study 2C replicated Study 2B's results for satisfaction and gappiness, as well as the mediation of the former through the latter (Fig. 3 and Supplementary Table 15).

Follow-up questions mitigate the effects of jargon

Given that positive effects of jargon on explanatory satisfaction occur for unfamiliar and invented jargon alike, they seem to reflect 'illusions' of understanding. How can these illusions be punctured? Our next studies aimed to identify effective interventions by targeting the perception of explanatory gaps, thus providing further support for our account of how jargon shifts explanatory satisfaction.

Study 3A experimentally manipulated perceived gappiness: participants first read and rated an explanation (either circular or complete; between participants), before answering a follow- up question about it and providing new ratings. Answering the follow-up question correctly required content exclusive to the complete explanations. Therefore, we expected follow-up questions to increase perceived gappiness and decrease explanatory satisfaction for participants exposed to circular explanations, especially when these contained jargon; no such effects were expected for complete explanations.

Overall, participants found circular explanations with jargon to be more satisfying than those without, but this effect was smaller after participants attempted to answer the follow-up question; as expected, we found no evidence of this interaction for complete explanations (Fig. 4 and Table 2). The reverse happened for perceived gappiness: circular explanations with jargon were judged less gappy, and this effect was smaller after participants attempted to answer the follow-up questions (there was a similar interaction effect for complete explanations). Finally, for circular explanations with jargon, we found that the drop in explanatory satisfaction after receiving the follow-up question was (fully) mediated by perceived gappiness, as our account would predict (indirect effect = −0.15, 95% CI [−0.23, −0.11]; direct effect = −0.07, 95% CI [−0.17, 0.04]).

Generating explanations mitigates the effects of jargon

Study 3B conceptually replicated Study 3A, but with a different debiasing intervention: asking participants to generate their own explanation for the initial question (rather than being asked a follow-up question, as in Study 3A). Participants first read and rated two circular explanations, one with added jargon and one without; then, they generated their own explanations for each phenomenon and provided new ratings.

Once again we observed the predicted interactions (Fig. 4 and Table 2; Supplementary Table 21 for the effects below). Generating explanations reduced explanatory satisfaction when the original explanations contained jargon ($b = -0.42$, 95% CI [−0.53, −0.30], $t = -6.92$, $P < 0.001$), but the reverse occurred for explanations without jargon ($b = 0.13$, 95% CI [0.02, 0.25], $t = 2.27$, $P = 0.023$). Conversely, generating explanations only increased perceived gappiness for original explanations with jargon ($b = 0.54$, 95% CI [0.42, 0.66], $t = 8.73$, $P < 0.001$;

https://doi.org/10.1038/s41562-025-02227-0

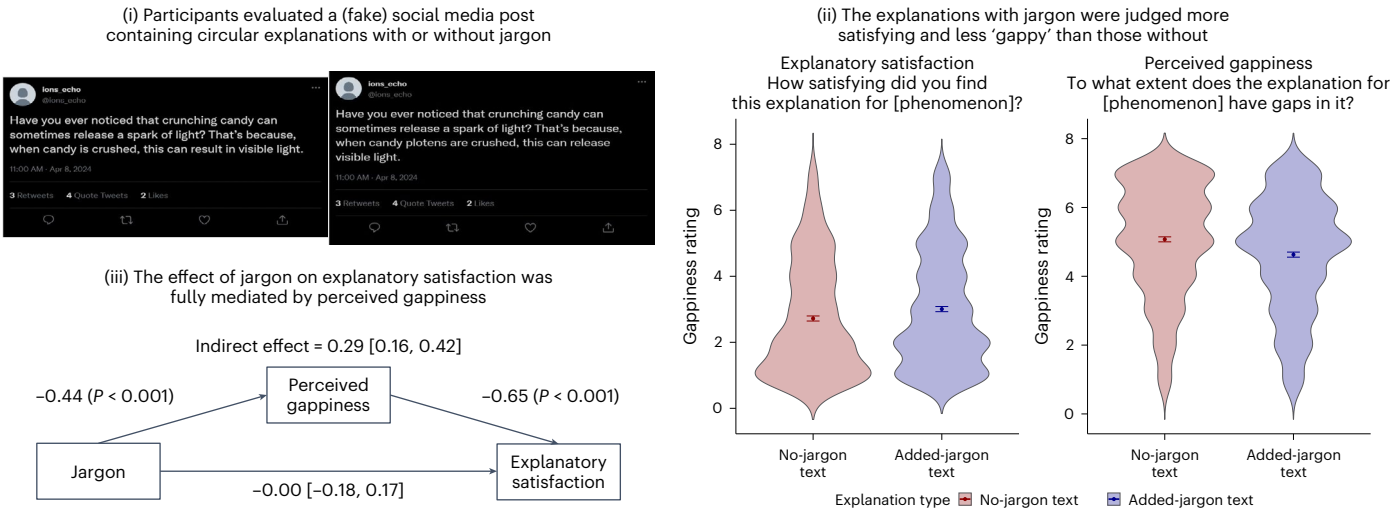

Fig. 3 | Sample stimuli and results for explanatory satisfaction and perceived gappiness as a function of jargon in Study 2C. Violin plots depict the distributions of individual ratings and data are presented as means ± s.e.m. for $N = 1{,}012$. Example stimuli are given in (i). The key result is that adding jargon to circular explanations increases explanatory satisfaction while decreasing perceived gappiness (as depicted visually in (ii)). The effect of jargon on explanatory satisfaction is fully mediated by perceived gappiness (see (iii) for results).

without jargon: $b = 0.06$, 95% CI [−0.06, 0.18], $t = 1.01$, $P = 0.313$). As in Study 3A, the debiasing effect of generating explanations on explanatory satisfaction was mediated by perceived gappiness (indirect effect = −0.28, 95% CI [−0.36, −0.21]; direct effect = −0.13, 95% CI [−0.23, −0.04]).

Participants in Study 3B were also asked to report how well they themselves could explain each target phenomenon. When participants received explanations with jargon, they reported considerably higher ratings than when they received explanations without (Table 2). However, after actually generating an explanation, participants decreased their ratings for explanations that initially contained jargon ($b = -0.42$, 95% CI [−0.53, −0.31], $t = -7.50$, $P < 0.001$), but 'increased' their ratings for explanations that did not ($b = 0.17$, 95% CI [0.06, 0.28], $t = 3.06$, $P = 0.002$; Supplementary Table 21).

In an (exploratory) analysis, we coded all explanations generated after receiving an original explanation with jargon in a subsample that generated high-quality explanations ($N = 374$; see Study 3B 'Participants' subsection in Methods). We found that only 22.3% of generated explanations included jargon ('preserved jargon'), while 77.7% did not ('eroded jargon').

Explanations were coded as containing jargon if they included terms (1) likely to appear or be defined in a science textbook while (2) being uncommon in everyday language or unknown to many laypeople. The explanations were coded by one of the authors, and agreement with a masked rater was high (92.7%; based on a random sample of 20% of the explanations).

First, we replicated all previously reported results when considering only this subsample of participants (see Supplementary Analyses and Supplementary Tables 22–26). In addition, both 'preserved jargon' and 'eroded jargon' participants rated their explanations higher in quality than did participants who received explanations without jargon (before and after they generated explanations, Table 3). Although we did not find evidence that participants in the eroded and preserved jargon groups initially rated their explanatory ability differently ($b = 0.11$, 95% CI [−0.16, 0.39], $t = 0.81$, $P = 0.421$), generating explanations introduced a discrepancy such that ratings after explaining were lower for the eroded group relative to the preserved group ($b = 0.40$, 95% CI [0.14, 0.67], $t = 2.97$, $P = 0.003$; Fig. 5 and Supplementary Table 26). In other words, generating an explanation had the greatest debiasing effect for participants who received jargon but were unable to produce it.

Failing to reproduce jargon predicts miscalibration

In Study 4, participants evaluated the explanations generated by participants in Study 3B. First, this replicated the finding that jargon can boost explanation ratings for poor explanations, but using a measure of explanation 'quality' (Fig. 5 and Table 3; for results on satisfaction: Supplementary Fig. 7 and Supplementary Table 28). Specifically, explanations with preserved jargon were rated better than explanations generated from no-jargon texts (whereas explanations with eroded jargon were not; see Table 3). Second, these explanation ratings allowed us to assess the 'calibration' of participants who generated explanations in Study 3B. Specifically, did they accurately anticipate how good others would judge their explanations to be, both before and after they generated them?

Our initial (preregistered) prediction—that Study 3B's participants would overestimate the quality of their explanations after receiving explanations with jargon—was not borne out. In fact, participants underestimated the quality of their explanations across conditions (although to varying degrees; see Supplementary Analyses and Supplementary Table 29). Instead, in exploratory analyses, we quantified calibration as the 'correlation' between quality ratings across Studies 3B and 4.

We found evidence of calibration (that is, significant correlations) for a subset of Study 3B's participants. For those exposed to no-jargon explanations and those exposed to explanations with jargon that they successfully preserved, correlations were significant whether we considered explanation ability ratings before or after participants in Study 3B actually generated explanations (Fig. 5). For Study 3B's participants who generated explanations with eroded jargon, however, we found something distinct: their ratings were significantly correlated with Study 4 ratings only 'after' they attempted to explain.

To explore whether the emergence of calibration for participants who generated explanations with eroded jargon was statistically meaningful, we tested for differences in correlations before and after explanation generation (see refs. 45,46 for analytic approach). Generating explanations indeed (significantly) improved calibration for participants who generated explanations with eroded jargon (that is, correlations between ratings in Study 3B and 4 were significantly higher for ratings provided after vs before explanation generation; difference = 0.10, 95% CI [0.02, 0.18]). This was also the case for no-jargon explanations (difference = 0.15, 95% CI [0.09, 0.22]), but there was no

https://doi.org/10.1038/s41562-025-02227-0

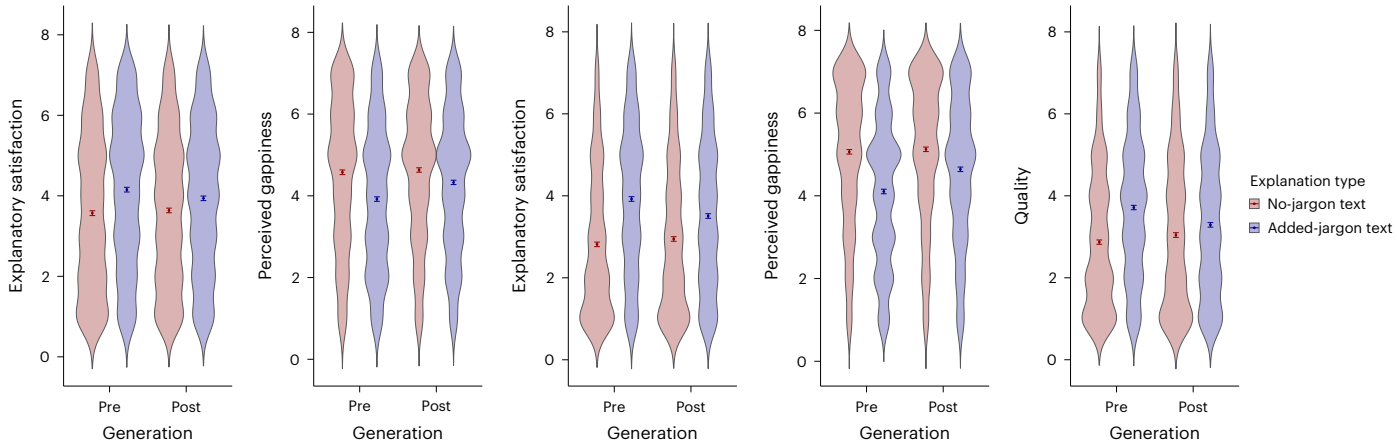

Fig. 4 | Explanatory satisfaction, perceived gappiness and explanation quality in Studies 3A and B. Two leftmost plots: Study 3A. Three rightmost: Study 3B. Explanatory satisfaction, perceived gappiness and explanation quality (Study 3B only) were measured both before and after the debiasing intervention of prompting participants to answer a follow-up question (Study 3A) or explain the phenomenon themselves (Study 3B). Violin plots depict the distributions of individual ratings and data are presented as means ± s.e.m. for $N = 1{,}026$ (Study 3A) and $N = 512$ (Study 3B). For circular explanations with jargon, attempting to answer a follow-up question (Study 3A) or to generate one's own explanation (Study 3B) resulted in a decrease in explanatory satisfaction and an increase in perceived gappiness (although this was not the case for relatively complete explanations, Study 3A). In both cases, the debiasing effect of generating explanations on explanatory satisfaction was mediated by perceived gappiness. In Study 3B, perceptions of quality increased with explanation generation for circular explanations without jargon, but decreased for circular explanations with jargon. The wording of the explanation quality item was the following: "If another person evaluated your explanation for [phenomenon], how good do you think that other person would think it is?".

difference in calibration for preserved-jargon explanations (difference = 0.05, 95% CI [−0.08, 0.18]). In sum, participants who received explanations with jargon that they subsequently failed to reproduce were uniquely uncalibrated, but recovered calibration after attempting to explain.

Discussion

Humans face a puzzle of ignorance: How can we evaluate explanations when doing so requires expertise that we lack? This puzzle holds across communities with distributed expertise and will become increasingly prevalent as humans interact with artificial intelligence that may exceed human capacities. We investigated this puzzle in the context of scientific explanations that contain expert language, or jargon.

First, our findings reconcile seemingly contradictory results from previous work, with jargon increasing explanatory satisfaction while decreasing comprehensibility. We find that jargon increases explanatory satisfaction for circular explanations (Study 1A), but not for more complete explanations (Studies 1B and C). Yet explanations with jargon are judged less comprehensible regardless of their completeness.

Second, we propose an explanation for why jargon inflates perceptions of explanatory satisfaction for poor explanations. When faced with explanatory gaps (as in our circular explanations), laypeople assume that authoritative jargon fills those gaps, even though they do not understand it (Studies 2A–C). This results in an elevated sense of explanatory satisfaction, even when comprehensibility is impaired. Interventions that expose explanatory gaps further support this account, with participants moderating their explanatory satisfaction after they are asked to answer follow-up questions (Study 3A) or provide their own explanations (Study 3B).

The perception of explanatory gaps uniquely explains the 'interactions' observed between the effects of jargon and explanatory completeness on satisfaction. By contrast, effects on comprehensibility and perceived expertise were not restricted to circular explanations. Jargon decreased comprehensibility for 'both' circular and complete explanations, plausibly because its unfamiliarity decreased fluency[33]. Similarly, jargon increased the inferred expertise of the explanation's source for circular and complete explanations alike (albeit more strongly for the former). These findings suggest that effects of jargon on satisfaction cannot be reduced to their effects on expertise. These findings are broadly consistent with previous work suggesting that (1) explanatory satisfaction is driven by perceived learning[27], and (2) people have a bounded understanding of science[47], relying on preconceived theories[48–52] and heuristic cues[31,38,53–55] (such as the presence of jargon) when evaluating explanations. In the current case, participants who received a complete explanation could have evaluated it on the basis of plausibility, mechanistic detail, or other criteria that were not available to participants who received circular explanations.

Third, our findings point to successful strategies for debiasing inflated judgements of an explanation's merit. When participants were prompted to answer follow-up questions (Study 3A), those who initially received circular explanations with jargon showed the largest drops in explanatory satisfaction. The same was found when participants generated their own explanations (Study 3B), not only for explanatory satisfaction, but also for participants' perceived ability to produce explanations that would be judged high quality by others. These findings suggest that the act of explaining can itself act as a corrective force, improving metacognitive calibration[27,40–42,55,56].

Our results can also be considered in light of the Dunning–Kruger effect, or the tendency for those lacking knowledge to overestimate their knowledge the most[57]. In the context of scientific explanations, this lack of calibration extends to those with intermediate levels of scientific knowledge[58–60]. Crucially, the Dunning–Kruger effect emerges with minimal learning, when complete ignorance is replaced by minimal knowledge[61,62]. Poor explanations may function similarly: circular explanations lacking jargon can be easily identified as deficient, but the inclusion of jargon is assumed to offer (minimal) explanatory knowledge, supporting perceptions of learning that bolster (over) confidence. Our debiasing procedures in Studies 3A and B counteract these inflated perceptions just as repeated task exposure increases calibration and reduces the Dunning–Kruger effect[61,62].

Consistent with these metacognitive effects, our final study (Study 4) considered whether effects of jargon on participants' judgements of their ability to generate high quality explanations were reflected in others' actual judgements of these explanations. In Study 3B, participants expected that their explanations would be more positively evaluated after having read original explanations with jargon, regardless of

Table 2 | Linear mixed models on explanatory satisfaction, perceived gappiness and quality as a function of jargon and timing in Studies 3A (for circular and complete explanations separately) and B

	Study 3A												Study 3B								
	Circular explanations						Complete explanations														
	Explanatory satisfaction			Perceived gappiness			Explanatory satisfaction			Perceived gappiness			Explanatory satisfaction			Perceived gappiness			Quality		
	b [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*
Predictors																					
Intercept	3.33 [2.83, 3.82]	13.25	<0.001	4.82 [4.37, 5.26]	21.23	<0.001	4.33 [3.75, 4.92]	14.47	<0.001	3.90 [3.34, 4.46]	13.67	<0.001	3.30 [2.81, 3.78]	13.23	<0.001	4.74 [4.27, 5.20]	19.97	<0.001	3.23 [2.81, 3.65]	15.16	<0.001
Jargon	0.75 [0.63, 0.86]	12.49	<0.001	−0.70 [−0.82, −0.59]	−12.09	<0.001	0.13 [0.02, 0.24]	2.28	0.023	−0.24 [−0.35, −0.13]	−4.24	<0.001	0.81 [0.72, 0.90]	17.90	<0.001	−0.69 [−0.78, −0.60]	−15.28	<0.001	0.53 [0.44, 0.61]	12.62	<0.001
Timing (Post vs Pre)	−0.02 [−0.13, 0.10]	−0.30	0.766	0.16 [0.05, 0.28]	2.81	0.005	−0.13 [−0.24, −0.02]	−2.24	0.025	0.30 [0.19, 0.41]	5.40	<0.001	−0.14 [−0.23, −0.05]	−3.15	.002	0.30 [0.21, 0.39]	6.58	<0.001	−0.12 [−0.21, −0.04]	−3.00	0.003
Jargon × Timing	−0.41 [−0.64, −0.17]	−3.40	0.001	0.48 [0.26, 0.71]	4.17	<0.001	−0.16 [−0.39, 0.06]	−1.43	0.153	0.23 [0.01, 0.44]	2.03	.042	−0.55 [−0.72, −0.37]	−6.08	<0.001	0.48 [0.30 – 0.65]	5.25	<0.001	−0.60 [−0.76, −0.43]	−7.15	<0.001
Random effects																					
σ^2	1.80			1.70			1.71			1.60			2.08			2.10			1.78		
τ_{00}																					
Participant	1.23			0.88			1.11			1.03			0.89			0.73			1.08		
Text topic	0.30			0.24			0.43			0.39			0.30			0.27			0.21		
ICC	0.46			0.40			0.47			0.47			0.36			0.32			0.42		
N																					
Text topic	5			5			5			5			5			5			5		
Participants	508			508			518			518			512			512			512		
Observations	2,032			2,032			2,072			2,072			4,096			4,096			4,096		
Marginal/ conditional R^2	0.043/0.482			0.049/0.428			0.003/0.476			0.013/0.478			0.054/0.398			0.048/0.355			0.030/0.439		

https://doi.org/10.1038/s41562-025-02227-0

https://doi.org/10.1038/s41562-025-02227-0

Table 3 | Linear mixed models on quality ratings as a function of jargon (coded) in Studies 3B (pre- and post-generation separately) and 4

	Study 3B						Study 4l		
	Pre-generation			Post-generation					
	b [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*	*b* [95% CI]	*t*	*P*
Predictors									
Intercept	2.81 [2.32, 3.31]	11.15	<0.001	2.97 [2.57, 3.38]	14.38	<0.001	3.31 [2.79 – 3.83]	12.55	<0.001
Eroded jargon (vs Control)	0.79 [0.64, 0.93]	10.55	<0.001	0.16 [0.02, 0.29]	2.30	0.021	0.14 [−0.01, 0.30]	1.78	0.075
Preserved jargon (vs Control)	1.10 [0.86, 1.35]	8.91	<0.001	0.62 [0.39, 0.84]	5.40	<0.001	0.92 [0.66, 1.17]	7.10	<0.001
Random effects									
σ^2	1.75			1.44			2.21		
τ_{00}									
Participant	0.80			1.27			0.71		
Text topic	0.30			0.19			0.33		
ICC	0.39			0.50			0.32		
N									
Text topic	5			5			5		
Participants	374			374			421		
Observations	1,496			1,496			1,684		
Marginal/conditional R^2	0.063/0.424			0.012/0.509			0.023/0.335		

whether their generated explanations preserved the originals' jargon or not. However, participants in Study 4 only rated generated explanations higher in quality than control explanations when they preserved original jargon, which happened less than a quarter of the time. In addition, when jargon was not preserved, Study 3B's participants did not reliably predict Study 4's ratings, potentially because they were swayed by the jargon in the original explanations that participants in Study 4 had no access to. Only after our debiasing intervention did calibration emerge. These findings suggest that people are influenced by jargon when estimating explanation quality, regardless of whether they are generating or receiving explanations.

While we refer to judgements of explanatory satisfaction that outstrip personal understanding as 'inflated' (since they do not reliably track the quality of explanations), it need not follow that they are irrational. In fact, scientific explanations often serve purposes beyond direct transmission of scientific content, such as scaffolding everyday problem-solving[47]. Even if people recognize that they personally lack the knowledge to understand jargon, they could expect others in their community to possess the relevant understanding[24], and thus treat the jargon as a placeholder for this understanding accessible in other minds. Thus, despite inflated judgements of explanatory 'satisfaction', jargon could help people recognize their dependence on experts. This is supported by previous work: while explanations without jargon generate confidence in one's personal understanding, explanations with jargon replacing non-jargon text bolster deference to experts[38] (see also ref. 63). We find this effect of jargon on deference to experts in several of our own studies (1A–C, see Supplementary Information), and this is arguably a beneficial effect: in most cases, laypeople are not well equipped to judge the veracity of scientific explanations on their own.

Laypeople could also extract higher-level information from the presence of jargon itself; for instance, that the phenomenon has an underlying cause[23], is reducible to a lower-level science[25], is connected to a larger explanatory system[32], or is highly complex[64]. If this is the case, then even circular explanations with unfamiliar jargon convey some information that participants can use to guide further judgements and inquiry.

Our findings have important implications for education, science communication and interactions with experts more broadly. Deficit models of expert communication place the locus of public–expert differences on the former's lack of knowledge, suggesting that it can be overcome if this knowledge is transmitted by experts[65,66]. However, laypeople and experts differ in more than the knowledge they possess[67,68], and belief revision depends on more than knowledge alone, such as motivation and context[69–73]. The present research contributes to these discussions by highlighting the role of opaque markers of expertise. While we focus on jargon, other cues (such as inscrutable graphics or formulas) could play similar roles[21]. We expect such cues to be superfluous and even harmful when interactions allow experts to convey complete explanations, since they decrease comprehensibility without promoting satisfaction. On the other hand, including jargon in more incomplete explanations is a double-edged sword[74]. It can boost explanatory satisfaction and deference to experts, but can also decrease comprehensibility and calibration. Accordingly, the inclusion of jargon in scientific explanations should ultimately depend on context and goals.

Our findings are also important for the domain of generative artificial systems, such as large language models. Individuals mistake available knowledge (such as that provided by these systems) with knowledge that they themselves possess[11,75–77]; over-relying on these systems could thus bolster people's propensity for illusions of (scientific) understanding (even in scientists[78,79]). In fact, merely expecting that an algorithm can provide an explanation increases perceptions of understanding[80]. These illusions may be particularly problematic for complex topics whose description require jargon, as its presence may nudge people towards prematurely ending information search. This research also informs the alignment problem, or the challenge of aligning these systems' functioning with human goals[81]. While humans may seek understanding, the models are geared towards providing engaging explanations, and these may be at odds with one another. The abundant training data used in generative artificial intelligence systems might capture the features of the 'minds of the public'[82], biasing them towards the cues individuals pick up on when reading explanations. Our work suggests that jargon is such a cue and, to the extent that models

https://doi.org/10.1038/s41562-025-02227-0

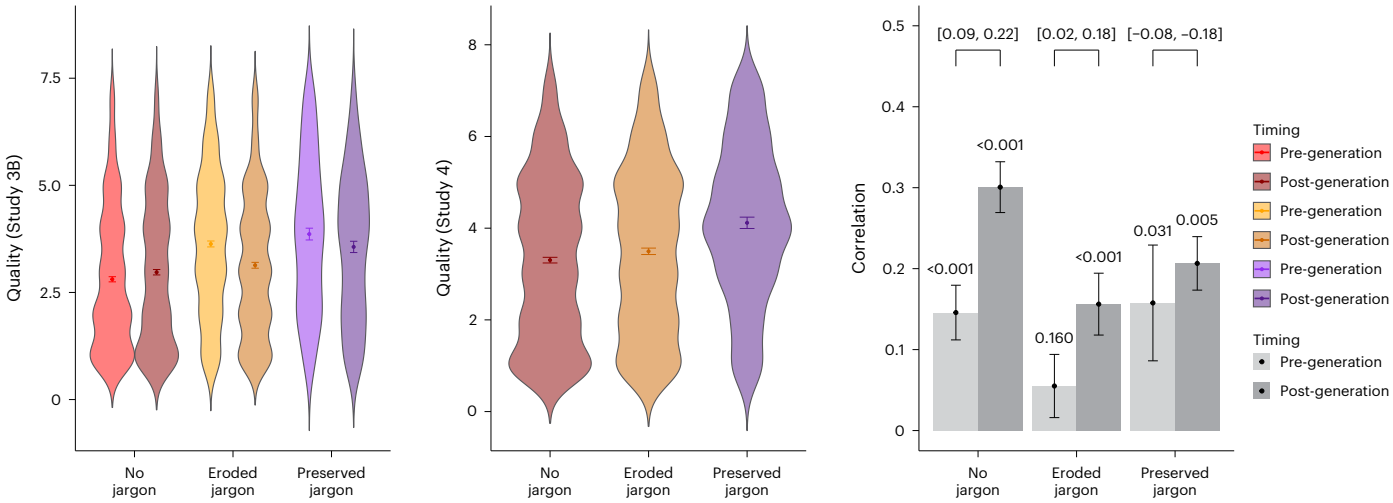

Fig. 5 | Quality ratings from Studies 3B and 4, as well as correlations between those ratings as a function of (coded) jargon. Violin plots depict the distributions of individual ratings and data are presented as means ± s.e.m. for $N = 374$ (Study 3B) and $N = 421$ (Study 4); bar plots depict the mean correlation and error bars represent ±s.e. Left: Study 3B; the key result is that explanations with eroded and preserved jargon were both rated higher than those based on explanations without jargon (both before and after the debiasing intervention). Middle: Study 4; explanations that preserved jargon were rated higher in quality than no-jargon explanations (but eroded-jargon explanations were not rated higher in quality than no-jargon explanations). Right: calibration between Study 3B and Study 4; the key findings are: (1) participants who generated explanations that failed to include jargon were uniquely uncalibrated and (2) the debiasing intervention increased, and induced, calibration for these participants who generated eroded-jargon explanations (such that the difference in correlations between pre- and post-debiasing intervention was significant). The first finding derives from significance testing of the correlations in each condition, with the number above each bar indicating the corresponding P value: (1) we find no evidence of correlation between Study 3B and Study 4 quality scores for pre-generation eroded-jargon explanations ($r(654) = 0.06$, $P = 0.160$), (2) but we reject the null hypothesis for the absence of an effect in all other conditions (no-jargon pre-generation: $r(842) = 0.15$, $P < 0.001$; no-jargon post-generation: $r(842) = 0.30$, $P < 0.001$; eroded-jargon post-generation: $r(654) = 0.16$, $P < 0.001$, preserved-jargon pre-generation: $r(188) = 0.16$, $P = 0.031$; preserved-jargon post-generation: $r(188) = 0.21$, $P = 0.005$); all tests were two-sided and no correction for multiple comparisons was implemented. The second finding derives from the values above brackets, which correspond to 95% confidence intervals for the difference in correlations across the corresponding bars (calculated using the method of ref. 46; significance is inferred if the interval does not contain 0): (1) the difference in the magnitude of correlations (from pre- to post-generation) is significant, indicating increased calibration, for explanations without jargon: 95% CI = [0.09, 0.22], and explanations with eroded jargon: 95% CI = [0.02, 0.18]; (2) but we find no evidence for such in explanations that preserve the original jargon: 95% CI = [−0.08, 0.18].

can identify that, they could potentially default to including jargon in explanations. This would constitute a problematic misalignment not only because individuals' preferences do not necessarily engender better learning, but also because vigilance for such a mismatch would be low[83].

The present work is not without limitations. Notably, we focus on explicit judgements of explanations, rather than their use in real-world decision-making over time (for instance, in solving everyday problems[47]). It thus remains a largely open question how such explanations translate into real-world consequences. Nevertheless, our effect sizes are not negligible (see Supplementary Table 29) and they extend to more naturalistic settings (that is, explanations in the form of social media posts in Study 2C). It is also plausible, although needing verification, that the miscalibration in explanation ability documented in Studies 3B and 4 shapes whether and how laypeople share the explanations they have been exposed to with others.

Another limitation concerns the scope of our materials. While we aimed to generate a diverse but representative set of largely novel scientific explanations for our participant population, we cannot confidently generalize to all scientific content. We also expect important boundary conditions to our effects of jargon to arise with different populations and in different contexts. For example, our findings probably hinge on participants' assumption that the jargon included in explanations is legitimate and relevant. A more informed or sceptical population could question these assumptions, resulting in different patterns of effects. On the other hand, the role of jargon in 'fixing' defective explanations could extend beyond the simple case of circularity: participants might assume that jargon compensates for other explanatory defects, such as weak evidence or an unspecified mechanism.

All in all, the present research proposes an account of when and how jargon influences perceptions of (scientific) explanations. Although always detrimental for perceptions of comprehensibility, jargon can increase perceptions of explanatory satisfaction for low-quality explanations. This is because jargon decreases the perception that the explanation contains explanatory gaps. This account not only identifies strategies for bringing explanatory satisfaction into closer alignment with actual understanding, but also offers a roadmap for studying puzzles of ignorance more broadly—how individual minds deal with the complex knowledge distributed across communities of minds with variable expertise.

Methods

The present research complied with all relevant regulations and was granted ethics approval by Princeton University's Institutional Review Board (Protocol 10662 IAA: Explanations and concepts). All participants provided informed consent before participating in the experiment. Although the experimental conditions that participants were assigned to were masked from participants, conditions were not masked from the researchers for the purposes of data analysis (it was masked for coding purposes, nonetheless; see 'Study 3B' in Methods).

Study 1A

This experiment was preregistered on 14 December 2023, and data were collected between 14 and 20 December 2023.

Participants. Our final sample included 737 participants recruited on Prolific ($M_{age} = 40.61$ years, $s.d._{age} = 13.95$ years; 338 men, 376 women, 20 non-binary and 3 did not disclose). An additional 38 participants

were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid US$0.70 for their participation (that is, at a US$10.50 hourly rate; all $ symbols hereafter refer to US dollars).

We preregistered a final sample of 728 participants. This number was informed by a sample size estimation conducted on G*Power (v.3.1)[84], considering an effect half as large as that obtained by ref. 23, for power set at 80%. The effect in question was that corresponding to the difference between the no-jargon and added-jargon conditions (see below). Since our design is within participants, this target sample actually yields power above 80% for an effect size of the magnitude described.

Materials. We created sets of explanations for three scientific phenomena: the formation of wine diamonds, candy triboluminescence, and the relationship between scented candles and cancer. The first two topics were drawn from previous research[21]; the third was selected after reviewing multiple online sources that we deemed to be reliable (for example, governmental websites, policy reports, scientific papers). These three phenomena were selected because they concern relatively unfamiliar scientific phenomena; hence we expected participants to be unfamiliar with their explanations.

For each phenomenon, we created three circular explanations: one without jargon ('no-jargon text'), one with jargon added to the no-jargon text ('added-jargon text'), and one with jargon added in lieu of non-jargon text ('replacement-jargon text'). All explanations were circular in the sense that they simply restated the phenomenon or relationship being explained. For example, for scented candles, the question was "Why do scented candles cause bladder cancer?" and the explanations were the following: "Burning scented candles causes changes in the replication of bladder cells – leading to bladder cancer" (no-jargon text), "Burning scented candles causes changes in the replication of bladder cells (specifically, changing base excision repair) – leading to bladder cancer" (added-jargon text) and "Burning scented candles causes base excision repair changes in bladder cells – leading to bladder cancer" (replacement-jargon text).

Explanations without jargon and with jargon replacing non-jargon content (that is, the no-jargon and replacement-jargon texts) were matched in word count (range: 10–15 words), while added jargon texts were longer by virtue of the jargon added to the no-jargon text (range: 20–22 words). We incorrectly preregistered that explanations without jargon would be matched in length with explanations with jargon added to them, instead of those for which jargon replaced non-jargon content. We meant to preregister what we implemented (and what is noted above), as otherwise explanations with jargon replacing non-jargon content would necessarily be shorter than versions without jargon and thus vary in more than just whether they had jargon or not.

This error extends to the preregistrations for Studies 1B and C. All explanations across all studies are available in a dedicated OSF folder (https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591).

Design and procedure. After providing their consent, participants were presented with three explanations, one at a time, in random order. Each explanation concerned a different phenomenon (wine diamonds, candy triboluminescence, or scented candles), and each explanation involved a different role for jargon (no-jargon text, added-jargon text, replacement-jargon text). Thus, phenomenon and text type were both within-participant variables, with the assignment of phenomena to text type randomized across participants.

After reading each explanation, and on a different page, participants responded to a set of measures using 7-point Likert scales. The measures were presented in the order described below (except when noted). The first item was a 'jargon manipulation check' to verify that explanations with jargon were perceived to contain more specialized language than those without ("To what extent do you think that the text that you just read is jargony? By jargony we mean that it is written with words or expressions used by a group specialized in the topic that others may not know about or may find difficult to understand."; from 1 = Not jargony at all, to 7 = Very jargony). The second set of items was adapted from previous literature finding favourable effects of jargon on explanation ratings[18–31]. Specifically, participants reported 'explanatory satisfaction' ("How satisfying did you find this explanation for [phenomenon]?"; from 1 = Not at all satisfying, to 7 = Very satisfying) and 'perceived learning' ("To what extent has this explanation for [phenomenon] taught you something new?"; this item was presented last and was answered from 1 = Definitely nothing new, to 7 = Definitely a lot new). The final set of items was drawn from previous work documenting negative effects of jargon[32–38] and assessed 'comprehensibility' ("How comprehensible was this explanation for [phenomenon]? By comprehensible we mean that you perceive the text as vivid, that you feel you can distinguish essential from rather unimportant information, that you think you can judge whether the single statements are consistent and not in conflict with one another, and that you feel you can clearly and comprehensively understand and connect the single statements made by the author."; from 1 = Very incomprehensible, to 7 = Very comprehensible), 'confidence' ("To what extent do you agree with the following statement: Based on my present knowledge about the topic, I am confident to decide whether it is correct that [phenomenon]."; from 1 = Strongly disagree, to 7 = Strongly agree), and 'deference to experts' ("To what extent do you agree with the following statement: Before I decide whether the explanation I just read is true, I would like to seek further advice from an expert."; from 1 = Strongly disagree, to 7 = Strongly agree).

Participants were debriefed and dismissed after answering two simple attention check questions (for example, "In this experiment, you read a text about: (1) The theory of natural selection, (2) Human contributions to tectonic activity, (3) A potential cause for cancer, or (4) Australian wildfires and governmental policies") and providing demographic information.

Study 1B

This experiment was preregistered on 6 December 2023, and data were collected between 6 and 14 December 2023.

Participants. Our final sample included 734 participants recruited on Prolific (M_{age} = 40.67 years, s.d.$_{age}$ = 12.48 years; 369 men, 348 women, 13 non-binary and 4 did not disclose). An additional 41 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $0.70 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 728 participants, on the basis of the same procedure used in Study 1A and described above.

Materials. We created new explanations for the same phenomena used in Study 1A (wine diamonds, candy triboluminescence and scented candles); there were three different explanations for each phenomenon with no-jargon, added-jargon and replacement-jargon texts. This time, however, the explanations were 'minimal' in the sense that they offered a partial but incomplete mechanism.

For example, for scented candles, the explanations were the following: "Burning scented candles can cause a series of physical, chemical and biological changes that result in uncontrolled cell replication in the bladder – or bladder cancer." (no jargon), "Burning scented candles can cause a series of physical, chemical and biological changes (through the creation of deoxyribonucleic acid adducts that compromise base excision repair mechanisms), leading to uncontrolled cell replication in the bladder – or bladder cancer." (added jargon), and "Burning scented candles can create deoxyribonucleic acid adducts that compromise base excision repair mechanisms, leading to uncontrolled cell replication

in the bladder – or bladder cancer." (replacement jargon). No-jargon and replacement-jargon texts were matched in length (range: 24–25 words); added-jargon texts were slightly longer (range: 34–37 words).

Design and procedure. The design and procedure followed Study 1A, except that the explanations used were minimal instead of circular (see 'Materials').

Study 1C

This experiment was preregistered on 20 December 2023, and data were collected between 20 December 2023 and 2 January 2024.

Participants. Our final sample included 733 participants recruited on Prolific (M_{age} = 41.07 years, s.d.$_{age}$ = 13.91 years; 352 men, 366 women, 13 non-binary and 2 did not disclose). An additional 68 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $0.70 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 728 participants, on the basis of the same procedure used in Study 1A and described above.

Materials. We created new explanations for the same phenomena used in Studies 1A and B (wine diamonds, candy triboluminescence and scented candles). We used the same approach from previous studies to generate three versions of each text: no jargon, added jargon and replacement jargon. These explanations were 'complete' in the sense that they identified the key processes involved in generating each phenomenon. For example, for scented candles, the explanations were the following: "Burning scented candles releases dangerous chemicals into the air. When these chemicals are ingested, they can link themselves to our genetic material and damage it, leading to uncontrolled cell replication in the bladder – or bladder cancer." (no jargon), "Burning scented candles releases aromatic hydrocarbons, a class of dangerous chemicals, into the air. When these chemicals are ingested, they can link themselves to our genetic material, creating deoxyribonucleic acid adducts, which damage the genetic material (specifically, the DNA's base excision repair mechanism), leading to uncontrolled cell replication in the bladder – or bladder cancer" (added jargon), and "Burning scented candles releases aromatic hydrocarbons into the air. When aromatic hydrocarbons are ingested, they can create deoxyribonucleic acid adducts, damaging DNA's base excision repair, leading to uncontrolled cell replication in the bladder – or bladder cancer" (replacement jargon). Again, we matched the no-jargon and replacement-jargon texts in word count (no jargon: 39–44 words; added jargon: 55–62 words; replacement jargon: 39–44 words).

Design and procedure. The design and procedure followed Studies 1A and B, except that the explanations used were complete (see 'Materials').

Study 2A

This experiment was preregistered on 27 February 2024, and data were collected between 28 February and 3 March 2024.

Participants. Our final sample included 1,014 participants recruited on Prolific (M_{age} = 39.71 years, s.d.$_{age}$ = 13.31 years, not considering a participant who answered '1981' but was included for analyses nonetheless, since there were no additional concerns about the quality of their data; 427 men, 563 women, 22 non-binary and 2 did not disclose). An additional 76 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered.

Participants were paid $0.70 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 1,000 participants. This number was informed by power simulations conducted using the Superpower package (v.0.2.0)[85], considering the data from Studies 1A and C in combination. For this final sample, we would have 98.5% power to detect a target jargon × completeness (see below) interaction effect matching that found in our data, 79.5% power to detect one with 3/4 of its magnitude and 48.5% to detect one with half the magnitude of the original effect.

Materials. Study 2A employed the no-jargon and added-jargon versions of the circular and complete explanations from Studies 1A and C, respectively. In addition, we created no-jargon and added-jargon versions of circular and complete explanations for two new phenomena: the relationship between urban areas and climate change, and the relationship between food waste and greenhouse gas emissions. For example, for food waste, the question was "Why does food waste in landfills increase greenhouse gas emissions?", and the explanations were the following: "Food waste in landfills undergoes chemical transformations that produce greenhouse gases" (circular no jargon), "Food waste in landfills undergoes chemical transformations (specifically, hydrolysis and acidogenesis) that produce greenhouse gases" (circular added jargon), "Food waste in landfills undergoes chemical transformations. The organic matter decomposes into simpler molecules that microorganisms turn into acids. In an environment without oxygen, these acids are transformed by bacteria into methane, a greenhouse gas." (complete no jargon), and "Food waste in landfills undergoes anaerobic decomposition, a kind of chemical transformation. The organic matter decomposes into simpler molecules (through hydrolysis) that undergo acidogenesis, turning into acids with the help of microorganisms. In an anaerobic environment (that is, one without oxygen), these acids are transformed by bacteria called methanogenic archaea into methane, a greenhouse gas." (complete added jargon). Circular explanations were shorter than complete explanations; since the added-jargon texts were created by adding jargon to no-jargon texts, they were not matched in word count (circular no jargon: 10–23 words; circular added jargon: 15–30 words; complete no jargon; 35–51 words; complete added jargon: 55–71 words).

Design and procedure. After providing informed consent, participants read two explanations in random order, one at a time, interleaved with our measures of interest. Jargon was manipulated within participants: One explanation used a no-jargon text and the other an added-jargon text. Explanation completeness was manipulated between participants: Half of the participants read two complete explanations (n = 505), whereas the other half read two circular explanations (n = 509). The two explanations seen by a given participant always involved different phenomena, randomly sampled from our pool of five (wine diamonds, candy triboluminescence, scented candles, urban areas, food waste).

After reading each explanation, participants answered a set of items using 7-point Likert scales, with all items presented on one page. Participants saw the items in the following order: (1) jargon manipulation check, (2) 'perceived gappiness' ("Explanations have 'gaps' when they're missing some of the information you would need for a complete explanation. To what extent does the explanation for [phenomenon] have gaps in it?"; from 1 = No gaps, to 7 = Many and/or large gaps), (3) explanatory satisfaction, (4) comprehensibility, (5) 'truthfulness' ("To what extent do you think that the explanation for [phenomenon] that you just read is true?"; from 1 = Very untrue, to 7 = Very true), (6) confidence, (7) deference to experts, (8) 'source expertise' ("To what extent do you think that the source of the explanation has expertise on the topic?"; from 1 = Very low expertise, to 7 = Very high expertise) and (9) perceived learning. While the measures of perceived gappiness and source expertise were created for the purpose of this study, the truthfulness item was adapted from previous literature on the detrimental effects of jargon[20]; all other items were worded as in previous studies (see Study 1A procedure).

After reading and rating the two explanations, participants provided demographic information and answered our attention checks before being debriefed and dismissed.

Study 2B

This experiment was preregistered on 21 May 2024, and data were collected between 23 and 24 May 2024.

Participants. Our final sample included 509 participants recruited on Prolific (M_{age} = 39.32 years, s.d.$_{age}$ = 12.89 years; 208 men, 293 women, 6 non-binary and 2 did not disclose). An additional 42 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $0.70 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 500 participants. No statistical methods were used to predetermine this study's intended final sample sizes; instead, given that we were interested in one level of the between-participants conditions in Study 2A (that is, circular explanations), we cut Study 2A's sample in half.

Materials. In this study, we only considered circular explanations from previous studies (that is, Studies 1A and 2A). We kept the no-jargon texts as originally included in these studies but adapted the added jargon to employ invented as opposed to actual jargon. This ensured that any boosts to explanatory satisfaction, if found, did not reflect genuine explanatory insight garnered from the content of the jargon terms. To accomplish this, we created pseudowords by rearranging the letters within each jargon word included in the original explanations (for example, in the explanation for food waste: 'hydrolysis and acidogenesis' was replaced with 'dryholysis and egenacidosis'), which we then confirmed to be pseudowords by searching for their meaning online (and failing to find any hits). Thus, invented jargon matched the real jargon used previously in length.

We employed a total of ten circular explanations, five without jargon and five with invented jargon, spanning five different phenomena (scented candles, wine diamonds, candy triboluminescence, food waste and urban areas). Since explanations were either those used previously or the result of simply replacing jargon words with rearrangements of the same letters, their length matched that of circular explanations used in other studies (no jargon: 10–23 words, vs added jargon: 15–30 words).

Design and procedure. The design and procedure followed Study 2A, except that the explanations used were circular with invented jargon (see 'Materials').

Study 2C

This experiment was preregistered on 21 November 2024, and data were collected between 21 and 22 November 2024.

Participants. Our final sample included 1,012 participants recruited on Prolific (M_{age} = 39.05 years, s.d.$_{age}$ = 13.32 years; 399 men, 597 women and 16 non-binary). An additional 90 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $0.70 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 1,000 participants. This number was again informed by power simulations conducted using the Superpower package (v.0.2.0)[85], matching target sample to that collected in studies with similar designs (that is, with between-participants factors, Study 2A) and based on pilot data. Based on these power simulations, we would have 99.85%, 95.60% and 67.30% power to detect an effect size of 100%, 75% and 50% (respectively) of that obtained with pilot data for the considered design and sample size.

Materials. In this study, we adapted the explanations from Study 2B to reduce the total amount of jargon and eliminate any words that could offer participants some understanding (for example, we removed the word 'repair' from 'base excision repair,' which became the single pseudoword 'bensatonicity'). We also relaxed the constraint that our pseudowords had to contain the same letters as the original jargon to create more natural-sounding pseudowords. These explanations were then incorporated into the wording and visual layout of a tweet (see Fig. 3), a social media post from the platform called X (previously called Twitter). The alleged poster username was generated using a tweet handle randomizer based on keywords for each explanation (that is, 'medicine,' 'oenology,' 'chemistry,' 'environmentalist' and 'architect'). The metadata (that is, posting date, likes, quotes and retweets) were randomized using online tools.

In addition, we added an introductory sentence or question to each post to more realistically replicate the properties of similar posts; hence the texts were longer than in Study 2B (although the explanations themselves had approximately the same length; no jargon: 26–44 words, vs added jargon: 26–45 words).

Design and procedure. Participants were first presented two tweets, one with and one without jargon, in random order, and asked, for each, whether they would share them (that is, "Would you retweet or otherwise share this post?"). These were presented to familiarize participants with the visual format of a tweet and to increase participants' sensitivity to the role of jargon (that is, establishing a baseline for what explanations could look like). They then read a third tweet about a different topic that, crucially, included either an explanation with (n = 510) or without (n = 502) jargon. Jargon was thus manipulated between participants. The topics were randomly allocated to the three tweets that participants were presented with. After reading the third tweet (twice, the second time being asked to pay close attention to it), participants answered the measures collected in Study 2B (minus the jargon manipulation check) in the same order as in previous studies.

Study 3A

This experiment was preregistered on 14 March 2024, and data were collected between 14 and 27 March 2024.

Participants. Our final sample included 1,026 participants recruited on Prolific (M_{age} = 39.40 years, s.d.$_{age}$ = 13.46 years; 429 men, 559 women, 35 non-binary and 3 did not disclose). An additional 84 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $1.05 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 1,000 participants. This number was informed by power simulations conducted using the Superpower package (v.0.2.0)[85]. We focused on the target interaction between jargon and timing observed in pilot data, and for circular explanations specifically. For N = 500, we would have 95.9% power to detect an interaction (that is, Jargon × Timing) effect matching that found in our data, 87.65% power to detect one with 3/4 of its magnitude and 69.15% to detect one with half the magnitude of the original effect we obtained. We doubled this sample to accommodate a second level in a between-participants condition—explanation completeness: complete explanations, which we did not consider in our power simulations because we did not anticipate any effect to emerge for this level of the condition.

Materials. We used the same 20 explanations from Study 2A, reflecting explanations that did or did not contain jargon, were circular or complete, and pertained to one of five phenomena (scented candles, wine diamonds, candy triboluminescence, urban areas, food waste). In addition, and for each phenomenon, we created one open-ended question. The questions were constructed to meet two criteria: (1) they

requested information contained within the complete explanations and (2) they requested information that was not contained within the circular explanation. For example, the follow-up question corresponding to the no-jargon text for scented candles was, "How do dangerous chemicals induce changes in cell replication?". For participants who received added-jargon texts, we used a version of the question that added the corresponding jargon, for example, "How do dangerous chemicals induce changes in cell replication (specifically, in base excision repair)?". This was done to better match the questions to the complete explanations from which they were extracted and to make it less likely that participants would answer the question by simply restating the jargon from their explanation when it was provided.

Design and procedure. Participants first read an explanation and then responded to the following measures in the following order: (1) perceived gappiness, (2) explanatory satisfaction, (3) perceived learning and (4) comprehensibility. Participants then had to provide an answer for a question about the phenomenon the explanation was about. They were told to think carefully and required to spend at least 30 s completing this task, although there was no maximum time limit. After answering the question, participants were again asked to rate the original explanation using the four measures listed above (presented in the same order).

Participants repeated this procedure twice, once for a no-jargon text and once for an added-jargon text. The follow-up question contained no jargon for no-jargon texts and jargon for added-jargon texts. Thus, jargon was manipulated within participants, and explanation completeness between (circular explanations: $n = 508$, complete explanations: $n = 518$). The two explanations always pertained to two distinct, randomly assigned phenomena. Finally, participants provided demographic information, answered attention checks, and were debriefed and dismissed.

Study 3B

This experiment was preregistered on 28 March 2024, and data were collected between 4 and 10 April 2024.

Participants. Our final sample included 512 participants recruited on Prolific ($M_{age} = 40.65$ years, $s.d._{age} = 13.62$ years; 195 men, 303 women, 10 non-binary and 4 did not disclose). An additional 53 participants were excluded for failing to complete the experiment or failing attention checks, as preregistered. Participants were paid $1.40 for their participation (that is, at a $10.50 hourly rate).

We preregistered a final sample of 500 participants. This number was informed by power simulations conducted using the Superpower package (v.0.2.0)[85]. We based the simulation on pilot data, considering the primary effect of interest (that is, interaction between jargon and timing). For $N = 500$, we would have 96.55% power to detect an interaction (that is, Jargon × Timing) effect matching that found in our data, 88.15% power to detect one with 3/4 of its magnitude and 46.85% to detect one with half the magnitude of the original effect we obtained.

In a second set of analyses that were not preregistered (see main text) but emerged after reviewing generated explanations, we considered only participants who generated explanations that were of high quality ($N = 374$). Participants were included in this subsample if none of their generated explanations: (1) were obviously Googled (for example, copied and pasted from an online source) or generated by a large language model (for example, contained LLM speech markers; $N = 22$), (2) mentioned an explanation having been provided previously ($N = 18$), or (3) simply refused to answer or claimed ignorance ($N = 50$); 48 participants met more than one criterion.

Materials. We used no-jargon and added-jargon texts for circular explanations from previous studies (that is, Studies 1A, 2A and B, and 3A). There were 10 explanations in total, spanning five phenomena (scented candles, wine diamonds, candy triboluminescence, urban areas and food waste) across two levels of jargon (no jargon and added jargon).

Design and procedure. In an initial block, participants read explanations, one at a time, and assessed them on the following measures in the following order: perceived gappiness, explanatory satisfaction, 'generation quality' ("If you were to write an explanation for [phenomenon] based on what you've learned, how good do you think another person would think it is?"; from 1 = Very poor, to 7 = Very good), comprehensibility and perceived learning. This procedure was repeated four times for different explanations about different phenomena presented in a random order and assigned at random (from our pool of five phenomena). Two of the explanations used no-jargon texts and two used added-jargon texts, such that jargon was manipulated within participants.

In the second block, participants were instructed to generate their own explanations. They received the same why-questions corresponding to the explanations they had just read but were instructed to use their own words to explain the phenomena they had just learned about. This generation task was self-paced, but participants were told that they always had to write something before proceeding. They wrote explanations one at a time, after which they again responded to measures assessing 'perceived gappiness' ("Explanations have 'gaps' when they're missing some of the information you would need for a complete explanation. To what extent do you now think that the explanation for [phenomenon] that you read at the beginning (that is, before you wrote an explanation for it yourself) had gaps in it?"; from 1 = No gaps, to 7 = Many and/or large gaps), 'explanatory satisfaction' ("How satisfying do you now find the explanation for [phenomenon] that you read at the beginning (that is, before you wrote an explanation for it yourself)?"; from 1 = Not at all satisfying, to 7 = Very satisfying) and 'generation quality' ("If another person evaluated your explanation for [phenomenon], how good do you think that other person would think it is?"; from 1 = Very poor, to 7 = Very good). These items were worded somewhat differently from their versions in the first block to make it clear that the evaluation pertained to the original explanation (for perceived gappiness and explanatory satisfaction) or to their generated explanations (for generation quality). Participants generated explanations and provided ratings for the four phenomena they read about initially, although their order was again randomized (hence could be different from the order in the first block).

Study 4

This experiment was preregistered on 18 April 2024, and data were collected between 19 and 22 April 2024.

Participants. Our goal was to pair each participant from Study 3B with (at least) one participant in this study. To ensure that Study 4's participants were presented with adequate explanations, we only considered Study 3B participants who provided high-quality explanations ($N = 374$; see Study 3B 'Participants' subsection for criteria). We collected data in waves, such that we recruited a new participant for each participant in Study 3B who still had not been paired with a valid participant in previous waves after exclusions; all waves were recruited via Prolific. Therefore, in Study 4, no statistical methods were used to predetermine sample sizes.

Our final sample consisted of 421 participants ($M_{age} = 37.37$ years, $s.d._{age} = 12.48$ years; 196 men, 213 women, 11 non-binary and 1 did not disclose). This was achieved after an initial sample of 374 participants, an additional 63 in wave two (due to 48 missing pairs, because some Study 3B participants were paired more than once, and 15 exclusions in wave one), an additional 6 participants in wave three (four missing pairs and two exclusions in wave two) and a final 2 participants in wave four (one exclusion in wave three and a buffer to account for possible exclusions). Participants were paid $0.88 for their participation (that is, at a $10.50 hourly rate).

Materials. The explanations used in this study were those generated by participants in Study 3B who passed a quality check (see Study 3B 'Participants' subsection). Each participant in Study 4 received all four explanations from their yoked participant in Study 3B, such that two had been generated after reading no-jargon texts and two after reading added-jargon texts; all four pertained to different phenomena.

Design and procedure. The procedure resembled that of Studies 1A–2B, except that participants read explanations generated by participants in a previous study, instead of explanations created by the research team. Each participant in this study was paired with a participant from Study 3B, such that the explanations they were presented were those generated by the corresponding participant from Study 3B; conversely, each participant in Study 3B's explanations were matched with at least one participant in Study 4.

Participants read several explanations, one at a time, in random order. After reading each explanation, they responded, on a separate page, to the following measures in the following order: 'explanation quality' (mirroring Study 3B's wording for generation quality; "How good do you think that this explanation for [phenomenon] is?"; from 1 = Very poor, to 7 = Very good), perceived gappiness, explanatory satisfaction, comprehensibility and perceived learning. Finally, we debriefed and dismissed participants after they provided demographic information and answered two attention checks.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data for all studies are publicly available in a dedicated OSF folder at https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591 (ref. 86).

Code availability

Code for all analyses, figures and tables in both the main text and Supplementary Information is publicly available in a dedicated OSF folder at https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591 (ref. 86).

References

1. Keil, F. C. Running on empty? How folk science gets by with less. *Curr. Dir. Psychol. Sci.* **21**, 329–334 (2012).
2. Sloman, S. & Fernbach, P. *The Knowledge Illusion: Why We Never Think Alone* (Penguin, 2018).
3. Ballantyne, N. *Knowing Our Limits* (Oxford Univ. Press, 2019).
4. DiPaolo, J. What's wrong with epistemic trespassing? *Philos. Stud.* **179**, 223–243 (2022).
5. DiPaolo, J. Who knows what? Epistemic dependence, inquiry, and function-first epistemology. *Inquiry* **67**, 670–687 (2024).
6. Peltokorpi, V. Transactive memory systems. *Rev. Gen. Psychol.* **12**, 378–394 (2008).
7. Ren, Y. & Argote, L. Transactive memory systems 1985–2010: an integrative framework of key dimensions, antecedents, and consequences. *Acad. Manage. Ann.* **5**, 189–229 (2011).
8. Wegner, D. M. in *Theories of Group Behavior* (eds Mullen, B. & Goethals, G. R.) 185–208 (Springer, 1987).
9. Bromme, R. & Thomm, E. Knowing who knows: laypersons' capabilities to judge experts' pertinence for science topics. *Cogn. Sci.* **40**, 241–252 (2016).
10. Wilkenfeld, D. A., Plunkett, D. & Lombrozo, T. Depth and deference: when and why we attribute understanding. *Philos. Stud.* **173**, 373–393 (2016).
11. Sparrow, B., Liu, J. & Wegner, D. M. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778 (2011).
12. Bromme, R., Rambow, R. & Nückles, M. Expertise and estimating what other people know: the influence of professional experience and type of knowledge. *J. Exp. Psychol. Appl.* **7**, 317–330 (2001).
13. Watson, J. C. Epistemic neighbors: trespassing and the range of expert authority. *Synthese* **200**, 408 (2022).
14. Lilienfeld, S. O. Can psychology become a science? *Pers. Individ. Dif.* **49**, 281–288 (2010).
15. Boyd, K. Trusting scientific experts in an online world. *Synthese* **200**, 14 (2022).
16. Lilienfeld, S. O. Public skepticism of psychology: why many people perceive the study of human behavior as unscientific. *Am. Psychol.* **67**, 111–129 (2012).
17. Aslanov, I. & Guerra, E. Tautological formal explanations: does prior knowledge affect their satisfiability? *Front. Psychol.* **14**, 1258985 (2023).
18. Aslanov, I. A., Sudorgina, Y. V. & Kotov, A. A. The explanatory effect of a label: its influence on a category persists even if we forget the label. *Front. Psychol.* **12**, 745586 (2022).
19. Bennett, E. M. & McLaughlin, P. J. Neuroscience explanations really do satisfy: a systematic review and meta-analysis of the seductive allure of neuroscience. *Public Underst. Sci.* **33**, 290–307 (2024).
20. Bulut, N. S., Gürsoy, S. C., Yorguner, N., Çarkaxhiu Bulut, G. & Sayar, K. The seductive allure effect extends from neuroscientific to psychoanalytic explanations among Turkish medical students: preliminary implications of biased scientific reasoning within the context of medical and psychiatric training. *Think. Reason.* **28**, 625–644 (2022).
21. Eriksson, K. The nonsense math effect. *Judgm. Decis. Mak.* **7**, 746–749 (2012).
22. Fernandez-Duque, D., Evans, J., Christian, C. & Hodges, S. D. Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *J. Cogn. Neurosci.* **27**, 926–944 (2015).
23. Giffin, C., Wilkenfeld, D. & Lombrozo, T. The explanatory effect of a label: explanations with named categories are more satisfying. *Cognition* **168**, 357–369 (2017).
24. Hemmatian, B. & Sloman, S. A. Community appeal: explanation without information. *J. Exp. Psychol. Gen.* **147**, 1677–1712 (2018).
25. Hopkins, E. J., Weisberg, D. S. & Taylor, J. C. V. The seductive allure is a reductive allure: people prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* **155**, 67–76 (2016).
26. Hopkins, E. J., Weisberg, D. S. & Taylor, J. C. V. Does expertise moderate the seductive allure of reductive explanations? *Acta Psychol.* **198**, 102890 (2019).
27. Liquin, E. G. & Lombrozo, T. Motivated to learn: an account of explanatory satisfaction. *Cogn. Psychol.* **132**, 101453 (2022).
28. Minahan, J. & Siedlecki, K. L. Individual differences in need for cognition influence the evaluation of circular scientific explanations. *Pers. Individ. Dif.* **99**, 113–117 (2016).
29. Rhodes, R. E., Rodriguez, F. & Shah, P. Explaining the alluring influence of neuroscience information on scientific reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 1432–1440 (2014).
30. Weisberg, D. S., Hopkins, E. J. & Taylor, J. C. V. People's explanatory preferences for scientific phenomena. *Cogn. Res.* **3**, 44 (2018).
31. Weisberg, D. S., Taylor, J. C. V. & Hopkins, E. J. Deconstructing the seductive allure of neuroscience explanations. *Judgm. Decis. Mak.* **10**, 429–441 (2015).
32. Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E. & Gray, J. R. The seductive allure of neuroscience explanations. *J. Cogn. Neurosci.* **20**, 470–477 (2008).

33. Bullock, O. M., Colón Amill, D., Shulman, H. C. & Dixon, G. N. Jargon as a barrier to effective science communication: evidence from metacognition. *Public Underst. Sci.* **28**, 845–853 (2019).
34. Scharrer, L., Bromme, R. & Stadtler, M. Information easiness affects non-experts' evaluation of scientific claims about which they hold prior beliefs. *Front. Psychol.* **12**, 678313 (2021).
35. Scharrer, L., Pape, V. & Stadtler, M. Watch out: fake! How warning labels affect laypeople's evaluation of simplified scientific misinformation. *Discourse Process.* **59**, 575–590 (2022).
36. Scharrer, L., Stadtler, M. & Bromme, R. You'd better ask an expert: mitigating the comprehensibility effect on laypeople's decisions about science-based knowledge claims. *Appl. Cogn. Psychol.* **28**, 465–471 (2014).
37. Scharrer, L., Stadtler, M. & Bromme, R. Judging scientific information: does source evaluation prevent the seductive effect of text easiness? *Learn. Instr.* **63**, 101215 (2019).
38. Scharrer, L., Britt, M. A., Stadtler, M. & Bromme, R. Easy to understand but difficult to decide: information comprehensibility and controversiality affect laypeople's science-based decisions. *Discourse Process.* **50**, 361–387 (2013).
39. Scharrer, L., Bromme, R., Britt, M. A. & Stadtler, M. The seduction of easiness: how science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learn. Instr.* **22**, 231–243 (2012).
40. Rozenblit, L. & Keil, F. The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* **26**, 521–562 (2002).
41. Bromme, R., Thomm, E. & Ratermann, K. Who knows? Explaining impacts on the assessment of our own knowledge and of the knowledge of experts. *Z. Pädagog. Psychol.* **30**, 97–108 (2016).
42. Meyers, E. A., Gretton, J. D., Budge, J. R. C., Fugelsang, J. A. & Koehler, D. J. Broad effects of shallow understanding: explaining an unrelated phenomenon exposes the illusion of explanatory depth. *Judgm. Decis. Mak.* **18**, e24 (2023).
43. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
44. Montoya, A. K. Probing moderation analysis in two-instance repeated-measures designs. *Multivar. Behav. Res.* **53**, 140–141 (2018).
45. Diedenhofen, B. & Musch, J. cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE* **10**, e0121945 (2015).
46. Zou, G. Y. Toward using confidence intervals to compare correlations. *Psychol. Methods* **12**, 399–413 (2007).
47. Bromme, R. & Goldman, S. R. The public's bounded understanding of science. *Educ. Psychol.* **49**, 59–69 (2014).
48. Kelemen, D., Rottman, J. & Seston, R. Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *J. Exp. Psychol. Gen.* **142**, 1074–1083 (2013).
49. Goldberg, R. F. & Thompson-Schill, S. L. Developmental 'roots' in mature biological knowledge. *Psychol. Sci.* **20**, 480–487 (2009).
50. Shtulman, A. Qualitative differences between naïve and scientific theories of evolution. *Cogn. Psychol.* **52**, 170–194 (2006).
51. Shtulman, A. & Valcarcel, J. Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition* **124**, 209–215 (2012).
52. Kovaka, K. Climate change denial and beliefs about science. *Synthese* **198**, 2355–2374 (2021).
53. Anderson, C., Brion, S., Moore, D. A. & Kennedy, J. A. A status-enhancement account of overconfidence. *J. Pers Soc. Psychol.* **103**, 718–735 (2012).
54. Cheever, N. A. & Rokkum, J. in *The Wiley Handbook of Psychology, Technology, and Society* (eds Rosen, L. D. et al.) 56–73 (Wiley, 2015).
55. Hofer, B. K. Epistemological understanding as a metacognitive process: thinking aloud during online searching. *Educ. Psychol.* **39**, 43–55 (2004).
56. Fonseca, B. & Chi, M. in *Handbook of Research on Learning and Instruction* (eds Mayer, R. E. & Alexander, P. A.) 296–312 (Routledge, 2010).
57. Kruger, J. & Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* **77**, 1121–1134 (1999).
58. Sanchez, C. & Dunning, D. Intermediate science knowledge predicts overconfidence. *Trends Cogn. Sci.* **28**, 284–285 (2024).
59. Lackner, S., Francisco, F., Mendonça, C., Mata, A. & Gonçalves-Sá, J. Intermediate levels of scientific knowledge are associated with overconfidence and negative attitudes towards science. *Nat. Hum. Behav.* **7**, 1490–1501 (2023).
60. Light, N., Fernbach, P. M., Rabb, N., Geana, M. V. & Sloman, S. A. Knowledge overconfidence is associated with anti-consensus views on controversial scientific issues. *Sci. Adv.* **8**, eabo0038 (2022).
61. Sanchez, C. & Dunning, D. Overconfidence among beginners: is a little learning a dangerous thing? *J. Pers. Soc. Psychol.* **114**, 10–28 (2018).
62. Sanchez, C. & Dunning, D. Decision fluency and overconfidence among beginners. *Decision* **7**, 225–237 (2020).
63. Hoogeveen, S. et al. The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nat. Hum. Behav.* **6**, 523–535 (2022).
64. Kominsky, J. F., Zamm, A. P. & Keil, F. C. Knowing when help is needed: a developing sense of causal complexity. *Cogn. Sci.* **42**, 491–523 (2018).
65. Simis, M. J., Madden, H., Cacciatore, M. A. & Yeo, S. K. The lure of rationality: why does the deficit model persist in science communication? *Public Underst. Sci.* **25**, 400–414 (2016).
66. Trench, B. in *Communicating Science in Social Contexts: New Models, New Practices* (eds Cheng, D. et al.) 119–135 (Springer, 2008).
67. Hendriks, F., Kienhues, D. & Bromme, R. in *Trust and Communication in a Digitized World: Models and Concepts of Trust Research* (ed. Blöbaum, B.) 143–159 (Springer, 2016).
68. Kaden, T., Jones, S., Catto, R. & Elsdon-Baker, F. Knowledge as explanandum: disentangling lay and professional perspectives on science and religion. *Stud. Relig.* **47**, 500–521 (2018).
69. Scheufele, D. A. Science communication as political communication. *Proc. Natl Acad. Sci. USA* **111**, 13585–13592 (2014).
70. Cruz, F. & Mata, A. Self-serving beliefs about science: science justifies my weaknesses (but not other people's). *Public Underst. Sci.* **34**, 172–187 (2025).
71. Ditto, P. H. & Lopez, D. F. Motivated skepticism: use of differential decision criteria for preferred and nonpreferred conclusions. *J. Pers. Soc. Psychol.* **63**, 568–584 (1992).
72. Munro, G. D. & Ditto, P. H. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Pers. Soc. Psychol. Bull.* **23**, 636–653 (1997).
73. Rutjens, B. T., Sutton, R. M. & van der Lee, R. Not all skepticism is equal: exploring the ideological antecedents of science acceptance and rejection. *Pers. Soc. Psychol. Bull.* **44**, 384–405 (2018).
74. Scharrer, L., Rupieper, Y., Stadtler, M. & Bromme, R. When science becomes too easy: science popularization inclines laypeople to underrate their dependence on experts. *Public Underst. Sci.* **26**, 1003–1018 (2017).
75. Sloman, S. A. & Rabb, N. Your understanding is my understanding: evidence for a community of knowledge. *Psychol. Sci.* **27**, 1451–1460 (2016).

76. Fisher, M., Goddu, M. K. & Keil, F. C. Searching for explanations: how the Internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* **144**, 674–687 (2015).
77. Rabb, N., Fernbach, P. M. & Sloman, S. A. Individual representation in a community of knowledge. *Trends Cogn. Sci.* **23**, 891–902 (2019).
78. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
79. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
80. Ostinelli, M., Bonezzi, A. & Lisjak, M. Unintended effects of algorithmic transparency: the mere prospect of an explanation can foster the illusion of understanding how an algorithm works. *J. Consum. Psychol.* **35**, 203–219 (2025).
81. Dung, L. Current cases of AI misalignment and their implications for future risks. *Synthese* **202**, 138 (2023).
82. Kidd, C. & Birhane, A. How AI can distort human beliefs. *Science* **380**, 1222–1223 (2023).
83. Celiktutan, B., Cadario, R. & Morewedge, C. K. People see more of their biases in algorithms. *Proc. Natl Acad. Sci. USA* **121**, e2317602121 (2024).
84. Erdfelder, E., Faul, F. & Buchner, A. GPower: a general power analysis program. *Behav. Res. Methods Instrum. Comput.* **28**, 1–11 (1996).
85. Lakens, D. & Caldwell, A. R. Simulation-based power analysis for factorial analysis of variance designs. *Adv. Methods Pract. Psychol. Sci.* https://doi.org/10.1177/2515245920951503 (2021).
86. Cruz, F., & Lombrozo, T. How laypeople evaluate scientific explanations containing jargon. *OSF* https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591 (2023).

Acknowledgements

We thank Fundação para a Ciência e Tecnologia (Doctoral Fellowship 2022.13009.BD) and Fulbright Portugal (Fulbright Research Fellowship 2023–2024) for their support in sponsoring F.C.'s visit to Princeton University, as well as the Concepts and Cognition Lab for feedback on subsets of this work. This work was not supported by any external funding, and the entities mentioned above had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. A subset of these experiments was presented at the 51st Annual Meeting of the Society for Philosophy and Psychology and at the 46th Annual Meeting of the Cognitive Science Society.

Author contributions

F.C. and T.L. contributed equally to this work. F.C. and T.L. conceptualized the studies and research designs, and developed the relevant stimuli. F.C. carried out experiment programming and data analysis, and wrote the original draft. T.L. provided funding and supervision. F.C. and T.L. read, reviewed and agreed to the published version of the Article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02227-0.

Correspondence and requests for materials should be addressed to Francisco Cruz.

Peer review information *Nature Human Behaviour* thanks Steven Sloman, Marc Stadtler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Francisco Cruz

Last updated by author(s): Mar 24, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected using Princeton University's institutional Qualtrics license
Data analysis	Data was analyzed using R (v.4.3.2) and openly available R packages (namely and primarily lmerTest v.3.1.3); we also used SPSS (v.4.2) and SPSS' PROCESS (v.) and MEMORE (v.3.0) macros

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data for all studies is publicly available in a dedicated OSF folder (https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We provide descriptive statistics on gender distribution as self-reported by participants, who have consented to share this information. No analyses were conducted taking into account gender, as its effects were not effects of interest in the present work.
Reporting on race, ethnicity, or other socially relevant groupings	We collected information on participants' educational background, but do not provide analyses considering it for parsimony (though descriptive information is available in Supplementary Table 31). All participants have consented to share this information. This information is provided in the data available online (https://osf.io/ytakw/?view_only=f3c34c42f79d4ecca2ab5502c35c0591) and can be used for further analyses. There is no information on race/ethnicity.
Population characteristics	See above. Additionally, descriptive information pertaining to participants' age is available in the Methods section (with means ranging 37.37-41.07 years, and standard deviations ranging 12.48-13.95 years, across studies); we additionally report educational attainment (see Supplementary Table 31). None of these were treated as covariate-relevant (e.g., included as coariates in the models).
Recruitment	Participants were recruited online via the Prolific Academic platform, a commonly-used platform for online data collection. No specific self-selection effects are expected.
Ethics oversight	Princeton University's Institutional Review Board (IRB Protocol - 10662 IAA: Explanations and Concepts). Participants gave informed consent before initiating the experiments.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitive data collected using within-subjects or mixed designs. Participants read several explanations with differing characteristics (representing the manipulation of interest) and rated them on several dimensions.
Research sample	Participants were recruited online via the Prolific platform and the sample is not representative. This platform was considered due to the need to recruit a substantial number of participants per study. Descriptive statistics pertaining age and (self-identified) gender are available in the Methods section; descriptive statistics pertaining to educational attainment are available in the Supplementary Information (Supplementary Table S31).
Sampling strategy	Sample size was determined a priori based on: a) power analysis considering previous effects in the literature (Studies 1A, 1B, and 1C), b) power simulations considering previous studies (Studies 2A and 2C), c) power simulations considering pilot data (Studies 3A and 3B), or d) previous studies' samples with no power analysis/simulations (Study 2B and 4); for a detailed justification for each study, see Methods subsection. Final samples ranged from 421 to 1026 participants. Sampling strategy was non-probabilistic by convenience sampling.
Data collection	Data were collected online and participants used their personal devices (i.e., only computers were allowed); only a web browser was required and no video or audio input was provided. Therefore, no experimenter was present when participants conducted the experiments, which were self-paced. Participants were blind to the experimental conditions and data are stored in Qualtrics' servers associated with Princeton University's institutional account.
Timing	Study 1A: December 14-20, 2023; Study 1B: December 6-14, 2023; Study 1C: December 20, 2023 - January 2, 2024; Study 2A: February 28 - March 3, 2024; Study 2B: May 23-24 2024; Study 3A: March 14-27, 2024; Study 3B: April 4-10, 2024; Study 4: April 19-22, 2024
Data exclusions	Data exclusions for each study are reported in the Methods section. Data exclusion criteria were preregistered and exploratory analyses with partial samples are indicated as such in the main text, and accompanied by their underlying rationale.
Non-participation	Study 1A: 0 declined, 26 dropped out; Study 1B: 0 declined, 19 dropped out; Study 1C: 0 declined, 18 dropped out; Study 2A: 0 declined, 41 dropped out; Study 2B: 0 declined, 0 dropped out; Study 2C: 0 declined, 0 dropped out; Study 3A: 1 declined, 128 dropped out; Study 3B: 0 declined, 41 dropped out; Study 4: 1 declined, 0 dropped out.

Randomization	For mixed designs, participants were allocated to the between-subjects in a counterbalanced manner (i.e., using Qualtrics' Randomizer paired with its Evenly present elements option).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*