Cascades, Leaps, and Strawmen: How Explanations Evolve

Kara Kedrick
Institute for Complex Social Dynamics, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Kevin Zollman
Department of Philosophy and Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Simon DeDeo
Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA
Santa Fe Institute, Santa Fe, NM 87501 USA; sdedeo@andrew.cmu.edu

Abstract

Explanations are social, and when people try to explain something, they usually seek input from others. We present a simple theory of how people use the explanations they encounter as clues to the broader landscape of possible explanations, informing their decision to exploit what has been found or explore new possibilities. The challenge of coming up with novel explanations draws people to exploit or imitate appealing ones (information cascades); this draw increases as less appealing alternatives become more distant (the "strawman" effect). Conversely, pairs of low-quality explanations promote exploratory behavior or long-leaps away from observed attempts, and pairs of divergent high-quality explanations can lead to merging and syncretism. We use a transmission-chain experiment to test, and confirm, these predictions. Intriguingly, we also find that while people imitate good explanations, their imitations often fall short in quality. Our work provides new insight into how collective exploration can be promoted, or stalled, by implicit information about what is yet to be discovered.

Keywords: explanation; cultural evolution; explore-exploit tradeoff; wisdom of the crowds; divergent thinking; creativity

Introduction

Explanations are social objects (van Fraassen, 1980; Mercier & Sperber, 2011). When we satisfy our drive for sense-making (Chater & Loewenstein, 2016) we do so, more often than not, by sharing, talking about, and revising our explanations with others. Little is understood about this process, and how our explanations are affected by the presence of prior explanations that serve as raw material for our own thoughts.

Considerable work, by contrast, has been dedicated to understanding the process by which a single mind, alone, arrives at and evaluates an explanation. Early philosophical theories focused on the relationship between logical deduction and explanation (Hempel & Oppenheim, 1948). Those were replaced by more complicated theories meant to better capture the variety of uses for explanations, especially in scientific contexts (Woodward & Ross, 2021). Psychological research, meanwhile, has provided empirical insights into why we prefer certain explanations over others. Lombrozo (2007), for example, has shown that people prefer explanations with co-explanatory power, one of a number of *explanatory values*. If people, in turn, tend to believe the explanations they value, that can lead to systematic deviations away from Bayes-rational behavior (Wojtowicz & DeDeo, 2020). The feeling of satisfaction that an explanation provides has a complex relationship to actual knowledge-gain (Liquin & Lombrozo, 2022), and work on "explanations in the wild" has uncovered a range of values (Sulik, van Paridon, & Lupyan, 2023)—including teleology, causation, and function—that are as relevant to the progress of science as they are to understanding how people fall victim to conspiracy theories.

No widely accepted explanation, however, emerges fully formed from a single mind. In order to understand how our epistemic drives affect our beliefs—and thus the origin of both our sciences and our cults—we need to understand what happens when explanations are shared and allowed to evolve.

To build this understanding, this work goes beyond prior studies which have focused on *de novo* explanation-making, where an explanation is conceived and developed independently by an individual. We look, instead, at the mechanisms that underlie the variation, selection, and cultural evolution of explanations, focusing on the conditions that promote the exploitation of prior perspectives or exploration of new possibilities. Our work provides a controlled, experimental parallel to data science approaches that look for the proposal and acceptance of explanations in real-world forums (Na & DeDeo, 2022).

We first present a simple framework to show how the existence of prior explanations can promote, or slow, the emergence of new forms of variation, and, in turn, can lead to the improvement, or degeneration, of explanations over time. Drawing on classic transmission-chain paradigms (Mesoudi & Whiten, 2008), we use a simple two-stage experiment to look for these effects in behavior.

Explanatory Landscapes and a Theory of Explanation Evolution

Often, something can be explained in multiple ways. This results in a diverse landscape of potential explanations, varying in both frequency of use and quality. By sampling the explanations offered by others, we gain insight into the structure of this landscape and the ease with which people reach high-quality explanations in different places.

Exposure to other people's explanations ought to influence our subsequent behavior. Most obviously, we can exploit the perspectives of others by simply copying—as best we can—what we find. While copying may be a basic form of cultural life in the steady state, it becomes more interesting when there

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

are multiple models of varying quality. Variation and selection are the cornerstones of evolution, and under the principle of "two heads are better than one" we might expect an evolution of explanations towards higher quality over time.

This simple and somewhat Panglossian account, however, neglects the ways in which evolution can turn down blind alleys. Sharing information is commonly seen as a virtue, but a long tradition in the study of the wisdom, and madness, of crowds, suggests that it is a double-edged sword (MacKay, 1852; Galton, 1907). This is made more complicated by the fact that cultural evolution, in contrast to biological evolution, is not so blind. When explanations are of varying quality and type, participants can use those facts, for better or worse, in strategic fashions (Douven & Mirabile, 2018).

In particular, prior explanations can provide not only sources to copy, but also information about the prospective quality of explanations "of that form." Both the apparent quality of the explanation, and the bare fact that it was made at all, can tell us where good ideas might be found and how hard we need to look or explore. We might perceive inadequate explanations as belonging to a cluster of similar unsatisfactory explanations, prompting us to abandon that specific cluster and explore elsewhere. The perceived quality of an explanation can be misleading: I may be dissuaded from a whole family of explanations by seeing a particularly bad example—even if much better explanations of the same type might have been found with a little work.

This process closely resembles how animals forage for food or information across various patches, where an unsatisfactory cluster is akin to a patch offering scant food or information (T. Hills, Todd, & Goldstone, 2008; T. T. Hills, Todd, Lazer, Redish, & Couzin, 2015). When confronted with explanations of a certain form, people have the option to either exploit that form by adopting the same or a similar perspective, or to reject the cluster they have encountered in favor of exploring alternative possibilities. Even within a cluster, a person might opt to (attempt to) wholesale copy a good explanation, or explore narrowly within the cluster for better options.

Deciding when to explore instead of exploit is a separate process from making the explanation itself, and thinking through how resource-limited agents can leverage *social* information—the subject of this work—to help make this decision leads to predictions for how the quality, and diversity, of explanations to hand affects what happens next.

Social Explanation and Navigating the Explore-Exploit Tradeoff. Presented with a prior set of explanations, if the person takes them to be the product of minds similar to their own, who have explored the landscape, the resulting quality of the explanations provides useful information for how to navigate the explore-exploit tradeoff.

If a person is given two similar explanations both of high quality, this provides the location of a potentially fertile part of the explanatory landscape. Conversely, a clearly bad explanation provides evidence that further thinking along those lines may be less likely to bear fruit. Neither inference is *necessarily* true, of course: a good explanation can be very similar in form to ones that are disastrously bad. At the very least, however, such samples provide information about what others have been able to accomplish with different explanatory forms.

If people use that information to navigate the explore-exploit trade-off, a number of predictions follow. Provided with good explanations of a similar form, a person will more frequently adopt similar explanations ("exploitation through imitation"). Conversely, bad explanations of a similar form will lead people to avoid investing time in that area, and to more exploratory behavior ("exploration through rejection").

More complex phenomena can occur in the presence of multiple explanations over very different forms. When there are multiple explanations of similar quality, but distinct form, at least two inferences are possible: (1) good explanations may be found in either location (leading to exploitation of either one), or (2) the overall landscape is relatively fertile (leading to more explorative behavior). While in the first case, individuals are following an exploitation through imitation strategy, the result at the population level for both cases increase the diversity of outcomes overall, relative to the case where everyone imitates explanations of the same form because there are two nearby (good) explanations: a form of "diversity through uniformity"; similar outcomes are expected in the case where both explanations are similarly bad.

Finally, there is the "mixed" case—where one explanation is good and the other bad. In this case, the distance between the two explanations is a critical variable. When the two explanations are very similar in form (the "nearby" case), this provides a signal for the explore-exploit decision that explanations of that form can be produced, but not particularly reliably (*i.e.*, in the "nearby" case, "exploitation through imitation" weakens as the quality of the worse explanation declines). As the distance gets larger, however, a second inference becomes possible: when there is a good explanation at one location and a bad explanation at a very different location, this provides evidence that deviation from the better model is unlikely to be successful. In this "distant" case, in other words, "exploitation through imitation" is expected to strengthen, as participants can use this as information about the danger of exploration.

These considerations lead us to predict an *interaction* between distance and relative quality. We refer to this as the "strawman effect", after the related strawman fallacy in argumentation (Talisse and Aikin (2006)'s "straw man by selection").

Varieties of Exploration. When people *do* choose to explore—*i.e.*, to produce explanations that are distant from all the options they're presented with—what happens next? One possibility is, of course, "pure" exploration; creating explanations that make no use of the social information except, perhaps, to rule out forms to include.

More complex forms of behavior are possible, including either to (1) merging, *e.g.*, “both A & B”, or (2) syncretism, an explanation that sits between A and B, drawing on some features of both. While syncretism may be more cognitively demanding, it may be superior to simple merging, particularly when A and B are partially contradictory, or when syncretism can produce something more simple, and thus more appealing, than the raw conjunct.

Methods

To test these predictions, we conduct a simple transmission-chain experiment with two stages: *de novo* generation and social iteration.

Participants are assigned to one (and only one) of the two stages; in either case, they are presented with a description of one of three events (a scenario), and asked “why” it happened. In order to encourage a diversity of explanations, the scenarios are deliberately constructed to be puzzling and to rule out obvious solutions. Here is one of our scenarios:

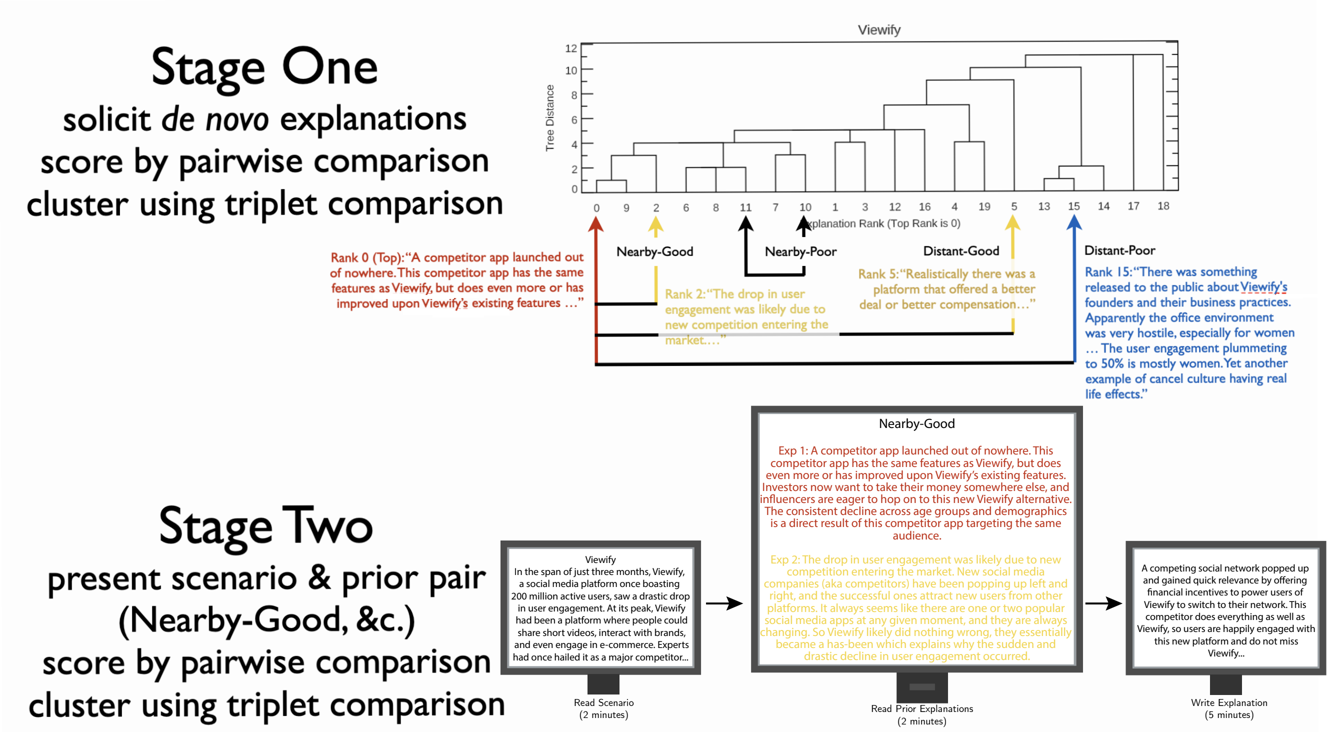
In the span of just three months, Viewify, a social media platform once boasting 200 million active users, saw a drastic drop in user engagement. At its peak, Viewify had been a platform where people could share short videos, interact with brands, and even engage in e-commerce. Experts had once hailed it as a major competitor to other big names in social media.

The user interface was lauded for its ease of use, the recommendation algorithms were spot-on, and the platform had successfully forged partnerships with big names in the entertainment industry. Advertising revenue was at an all-time high, and Viewify seemed poised for a golden era.

Then, almost overnight, things changed dramatically. User engagement plummeted by 50%. Many users started deactivating their accounts, while others became inactive. Brands began pulling their advertising, citing low return on investment. Several influencers publicly announced their move to other platforms.

Panic set in among the investors and stock prices tumbled. Oddly, there were no changes in the platform’s UI/UX, no major outages, and the company hadn’t made any significant controversial decisions recently. Moreover, the decline was uniform across different age groups and demographics, which stumped analysts even further. Surveys conducted by third parties indicated user satisfaction but did not offer any clear insight into the sudden drop.

After reading the scenario, participants answer the question: *Why did Viewify experience such a sudden and drastic decline in user engagement despite no apparent missteps?* A



Stage One

solicit *de novo* explanations

score by pairwise comparison

cluster using triplet comparison

Rank 0 (Top):“A competitor app launched out of nowhere. This competitor app has the same features as Viewify, but does even more or has improved upon Viewify’s existing features …”

Rank 2:“The drop in user engagement was likely due to new competition entering the market.…”

Rank 5:“Realistically there was a platform that offered a better deal or better compensation…”

Rank 15:“There was something released to the public about Viewify's founders and their business practices. Apparently the office environment was very hostile, especially for women … The user engagement plummeting to 50% is mostly women. Yet another example of cancel culture having real life effects.”

Stage Two

present scenario & prior pair

(Nearby-Good, &c.)

score by pairwise comparison

cluster using triplet comparison

Viewify

In the span of just three months, Viewify, a social media platform once boasting 200 million active users, saw a drastic drop in user engagement. At its peak, Viewify had been a platform where people could share short videos, interact with brands, and even engage in e-commerce. Experts had once hailed it as a major competitor...

Nearby-Good

Exp 1: A competitor app launched out of nowhere. This competitor app has the same features as Viewify, but does even more or has improved upon Viewify’s existing features. Investors now want to take their money somewhere else, and influencers are eager to hop on to this new Viewify alternative. The consistent decline across age groups and demographics is a direct result of this competitor app targeting the same audience.

Exp 2: The drop in user engagement was likely due to new competition entering the market. New social media companies (aka competitors) have been popping up left and right, and the successful ones attract new users from other platforms. It always seems like there are one or two popular social media apps at any given moment, and they are always changing. So Viewify likely did nothing wrong, they essentially became a has-been which explains why the sudden and drastic decline in user engagement occurred.

A competing social network popped up and gained quick relevance by offering financial incentives to power users of Viewify to switch to their network. This competitor does everything as well as Viewify, so users are happily engaged with this new platform and do not miss Viewify...

Figure 1: Illustration of Stage One Analysis and Stage Two Presentation. After soliciting *de novo* explanations in Stage One, and pairwise ratings from a separate set of participants, we clustered the explanations based on triplet judgements of “more similar” using GPT. We then selected explanation pairs with different quality ratings, quality rating differences, and distances. In Stage Two, participants (1) read the scenario, as in Stage One, (2) read one of the explanation pairs, and (3) provided the explanation they think best explained the scenario. Stage Two participants were told that they could reuse all or part of the explanations they saw.

brief description of our other scenarios is as follows: Scenario *SensAI* asked why a Generative AI system "paused" before answering a question about the nature of happiness; Scenario *School* asked why a school reform, which was initially very successful, began to fail for some of the students and not others.

Stage One: *de novo* generation. In the first stage of the experiment, we solicited explanations from 20 participants, recruited through Prolific. Participants were presented the three event scenarios, and asked to explain why they happened. These explanations were independently generated, and were shown to other participants in Stage Two. Explanations were limited to a maximum of one hundred words, and participants were incentivized: they would receive an additional bonus if their explanations were rated in the top ten-percent of all those provided.

Next, we recruited 103 participants who judged the quality of the explanations in a series of pairwise comparisons that asked which of the pair was "more satisfying". Each participant made ten pairwise comparisons for two scenarios; overall, we had 561 comparisons for each scenario, an average of three per pair, and each of the 190 possible pairs for each scenario measured at least once. These revealed preferences enabled us to construct an overall quality rating for each explanation, using a simple logit model.

We then measured the diversity of forms or the semantic distance between explanations. To do this, we used GPT-3.5-Turbo to make triplet comparisons to cluster explanations into a tree-like hierarchy: for each of the 1140 triplets of explanations, we asked GPT which pair are the "most similar" (see Figure 1). The distance between two explanations is then defined as the percentage of the time that the two were chosen as more similar than any third.

Stage Two: social iteration. In Stage Two, 119 participants were presented the three original scenarios (see Figure 1). Following each scenario, they were then shown, on the following screen, a *pair* of explanations drawn from participants in Stage One, described as "explanations that other participants provided" (referred to as *inherited explanations*). Participants saw one of six pairs of explanations for each stimulus, that differed in quality (*e.g.*, both good, both bad, or mixed good and bad) and semantic distance (*e.g.*, similar or different). Finally, the explanations are removed, they are shown the original scenario again, and given a text box to provide their own explanation (referred to as *derived explanations*).

They were incentivized as in Stage One, receiving a bonus if their explanations were rated in the top ten-percent of all those provided that round. They were told they could reuse all, or part, of an explanation they had seen before, and that their explanation would not be in competition with the ones they were shown. In our final sample of 357 explanations, we found only two cases where one of the stimuli was copied and pasted exactly. In two cases, we found language that showed a participant used a ChatGPT-like system to answer, and this was excluded. The average length of an explanation in Stage One was 62 ± 6 words, and in Stage Two was 62 ± 1 words (all ± reports are standard errors). Afterward, we evaluated the quality and diversity similarly to the methods used in Stage One, additionally calculating the semantic distance between the explanations from Stage One and Stage Two.

(DV)	Q_1	D_{12}	$\mathrm{D}_{12}(Q_1 - Q_2)$
	a	$-b\Delta_\star$	b
$\mathrm{D}_{1,ans}$	$\mathbf{-0.22 \pm 0.05}$	$\mathbf{0.51 \pm 0.08}$	$\mathbf{-0.45 \pm 0.08}$
(t)	-4.1	$+6.1$	-5.5

Table 1: Exploitation is moderated by metainformation about the underlying landscape of explanations; as distance between the inherited explanations (Dist_{12}) increases, the effect of the lower-quality explanation depends upon its relative score ($Q_1 - Q_2$).

We observed high inter-rater reliability when comparing human pairwise rankings of Stage One results to judgements made by GPT-3.5-Turbo with a crafted prompt; the Pearson correlation of human- and GPT-derived logit scores was 0.86 (for the prompt "which explanation is better"), 0.78 (for "which explanation is more likely"), and 0.66 (for "which explanation is more satisfying"), comparable to the correlation when the human-derived set was split in two halves (0.86). This validated our use of GPT-3.5-Turbo to make the additional pairwise judgements needed to produce scores for the derived explanations. Triplet comparisons by GPT were less reliable; overall, hand-checks of 60 triplets by a human rater found that machine judgements matched 55% of the time, compared to 33% at chance ($\kappa = 0.33$, "fair"). This is better than it seems, for two reasons: (1) our final distance judgements average over 19 alternatives, and (2) many errors are due to the fact that some triplets are, indeed, very close. Human and machine judgement match better when one pair is close compared to other combinations; for example, for triplets where the minimum distance is less than half that of the maximum distance ($N = 36$), humans and machines agree 65% of the time ($\kappa = 0.52$, "moderate").

Results

The explore-exploit tradeoff. We first examine the factors that predicted a Stage Two participant's deviation from the prior explanations. Following the discussion above, we focus on the semantic distance between the explanation a participant generated and the two they were presented with (the "inhereted" explanations), and how this is predicted by the quality ratings of those inherited explanations, and the distance between them.

Formally, we model the semantic distance from the better of the two inherited explanations to the derived explanation, $D_{1,ans}$. A decrease in distance is indicative of more exploitation, where the participants' responses differ less from the best inherited explanation. Alternatively, an increase in distance signals more exploratory strategies. We expect that the distance is driven by the quality of the better explanation, Q_1. It is also affected by the difference between the qualities of the explanations, $Q_1 - Q_2$ where Q_2 is the quality of the less-good explanation.

The effect of this difference is expected to depend on the distance between the two inherited explanations, D_{12}; when the semantic distance is large and the alternative is weak, we expect $D_{1,ans}$ to decrease with distance—people use the presence of the "strawman" as information that exploration is unlikely to be successful. Conversely, when two low-quality explanations have a small semantic distance, and the second-best alternative is recognizably at par with the optimal—we expect the opposite. This leads to an interaction, which we model using a linear regression as

$$D_{1,ans} = aQ_1 + bD_{12}((Q_1 - Q_2) - \Delta_\star) + \text{Const}. \qquad (1)$$

The first term characterizes the basic "exploitation through imitation" and "exploration through rejection" effects from our framework; we expect a to be negative.

The second term characterizes the more complex interaction effects of the "strawman" effect predicted by our framework. The b term is expected to be negative (the strawman draws participants towards the better explanation), and the $\Delta_\star$ term is expected to be a (positive) critical distance, beyond which that strawman effect kicks in. This means that the coefficient on D_{12} alone, equal to $-b\Delta_\star$, is expected to be positive).

As shown in Table 1, all three of the predictions of our framework are borne out by the data, with both large effect sizes and significance.

Our findings also imply that when both explanations are good (Q_1 is high, and $Q_1 - Q_2$ is small), responses tend to be more distant from the best explanation ($D_{1,ans}$ rises) as the semantic distance D_{12} increases. This could be due to two effects: (1) participants can increase $D_{1,ans}$ if they occasionally choose the second alternative to imitate, (2) they could be engaging in more exploratory forms, such as merging or syncretism.

It is difficult to disentangle these two effects using only regressions on $D_{1,ans}$ (or the distance between the derived explanation and the second lower-quality explanation, $D_{2,ans}$). This is because they require us to distinguish Effect 1: "an (imperfect) imitation of the second explanation (E2)" from Effect 2: "a exporatory combination of first explanation (E1) and second explanation (E2)", but the emergence of explanations of either kind will show both increased distance from E1 and decreased distance from E2.

One difference that is potentially detectable is in the distribution of distances to the *nearest* of the two stimuli: as the E2-near explanations increase in quality, the histogram of distance to the nearer of E1 and E2 would either have a peak concentrated at zero (as people choose one or the other to imitate, Effect (1)), or would peak at some intermediate distance (exploration, syncretism and merging, Effect (2)).

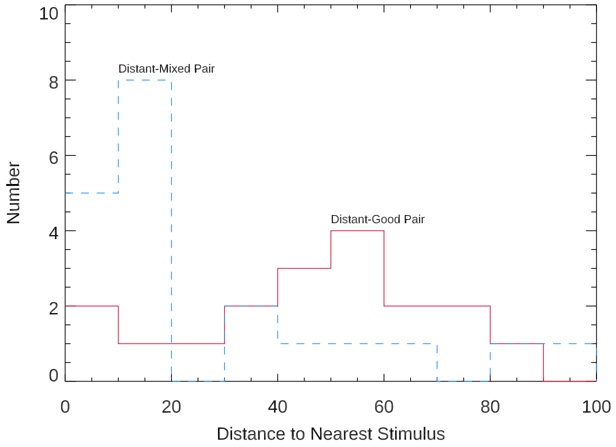


Figure 2: Histogram of the distance from participant explanation to the nearest of the two provided stimuli, for two particular explanation pairs ("Distant-Good", provided to 18 participants, and "Distant-Mixed", provided to 20 participants). In this example, the two Distant-Good explanations are both equally high-quality, and far from each other (Distant-Good; $S_1 = 52$, $Q_1 - Q_2 = 1.7$, $D_{12} = 62$), and participants tend to provide answers that deviate from both. This can be compared to a case where one explanation is clearly better (Distant-Mixed; $S_1 = 62$, $Q_1 - Q_2 = 55$, $D_{12} = 89$).

Fig. 2 shows this histogram for a "Distant-Good" Pair where E1 has a reasonably high quality score, the difference in quality small, and the semantic distance is reasonably large; the peak in the middle shows that Effect (2) is more likely to be in play in this example. We can contrast this distribution with a "Distant-Mixed" Pair, where we can see what high concentration looks like when the difference is clear. Examples such as these provide suggestive evidence; a more stringent comparison of Effects (1) and (2), that would account for the *a priori* structure of the space, would fix E1, and find a range of explanations as close as possible to E2, of varying quality. We leave this to future work.

Quality of explanation. The previous section confirmed our framework's major hypotheses about how people use social information to navigate the explore-exploit tradeoff, and then explored in more detail the mechanisms behind of the "diversity through uniformity" hypothesis.

In this section, we conduct a more exploratory investigation to evaluate whether providing better explanations produces better outcomes. We consider the following model for Q_{ans}, the quality of the derived explanation, as a function of the quality scores of the two provided explanations, Q_1 and Q_2, the distance between them, D_{12}, and how that is affected by difference in quality $Q_1 - Q_2$:

$$Q_{ans} = aQ_1 + bQ_2 + cD_{12}((Q_1 - Q_2) - \Delta_\star) + \text{Const}. \qquad (2)$$

In general, we expect better inputs to provide better outputs: we expect both a and b in Eq. 2 to be positive. When distance is small, we expect a similar effect for the quality of the second explanation, so that b is also positive. As the distance increases, however, and participants are presented with multiple, incompatible models, the results of the previous section lead us to expect that merging effects will kick in. If merging leads to better performance at fixed distance, we expect the coefficient c, on $D_{12}(Q_1 - Q_2)$ to be negative, and the coefficient on D_{12} alone to be positive.

The results, however, are surprising. We do find the expected positive effect of Q_1 ($\beta = 0.46 \pm 0.17$, $t = +2.6$), but the effect of Q_2 is in the *opposite* direction ($\beta = -0.53 \pm 0.22$, $t = -2.3$). We do find an effect of distance: c is positive ($\beta = -0.75 \pm 0.22, t = -2.9$), and $c\Delta_\star$ is positive ($\beta = 0.36 \pm 0.12$, $t = +3.1$).

Taken at face value, these suggest a somewhat surprising "competition" story. The most significant improvements come from having one explanation be clearly better than another; indeed, regardless of distance, increasing the quality of the second explanation *decreases* the outcome quality (happily, the raw coefficients balance in favor of Q_1; increasing the quality of both increases the outcome quality).

Notably, the overall r^2 of the prediction is small (0.03, $N = 351$); although we see a general increase in quality—the average score in Stage One is 50 ± 3, and the average score in Stage Two is 56 ± 1—the change in quality is not strongly predicted by the quality and diversity of the inherited explanations. While people are certainly capable of being drawn—quite strongly—towards better explanations (see previous section), and they can leverage good explanations to produce better ones in turn, the fitness improvements are not particularly strong.

Group-Level Diversity. The previous sections considered the effect of inherited explanations on individual participants. We can also ask about the effect at the group level: under what conditions are the derived explanations more distinct from each other?

In particular, for each of our stimulus pairs, we can measure the average pairwise distance between the explanations produced by participants who received that stimulus, Div_{ans}. This is a group-level measurement, that tells us how much particular pairs of explanations can send participants off in diverse directions.

We consider the following model for Div_{ans},

$$\mathrm{Div}_{ans} = a(Q_1 - Q_2) + bD_{12} + \mathrm{Const}. \qquad (3)$$

We expect higher diversity when there is no clear winner (*i.e.*, $Q_1 - Q_2$ is small); the "diversity through uniformity" predicts a is negative; we also expect higher diversity when the distance between the inherited explanations is large, and thus expect b to be positive.

We expect that when there is a clear winner ($Q_1 - Q_2$ is large), increasing distance should further harm diversity (ruling out more of the space as "infertile"), so that b is negative. Conversely, we expect that when there is not a clear winner, increasing distance should help, so we expect $b\Delta_\star$ (the coefficient on D_{12} alone) to be positive.

Our expectations are directionally correct: a is negative (-0.32 ± 0.28, $t = -1.2$), and b is positive (0.35 ± 0.28, $t = 1.3$); r^2=0.12. However, neither effect is significant; in part because we only have a small number of distinct pairs.

Discussion

Work on how we share and are influenced by the ideas of others has tended to focus on relatively simple questions: when Galton (1907) introduced the idea of wisdom of the crowds, it was in the context of guessing the weight of an ox, and classic work on information cascades includes tasks such as predicting the color of balls in an urn (Anderson & Holt, 1997), rating a movie (Lee, Hosanagar, & Tan, 2015), or choosing which URL to mention on Twitter (Galuba, Aberer, Chakraborty, Despotovic, & Kellerer, 2010). Much less is understood about what happens in more complex tasks such as explanation-making, where the space is exponentially vast and even, potentially, unprestatable (Longo, Montévil, & Kauffman, 2012).

Drawing on the classic explore–exploit tradeoff, this work has presented a framework that predicts how we choose between exploiting explanations from others or exploring new possibilities in the process of constructing our own. We can see the emergence of information cascades, as good explanations promote exploitation. We also observe the more complex strawman effect. Our results suggest, for example, that people are drawn closer to a (good) explanation by being presented with a poor, distant alternative. The strawman effect illustrates why a particular (fallacious) argumentative strategy is especially effective: by discouraging foraging in an area that may contain stronger explanations.

Our second finding is that there is some improvement to be gained by seeing better explanations, but the gains to be found are small. Those who teach may be less surprised by this second finding. It is certainly possible to draw students closer to a good explanation, but few students are able to reproduce that explanation at the same level of quality. The insight that the teacher provides, and the student experiences, does not always lead to learning. It is not enough to be drawn to an explanation—one must also, perhaps, adapt to it on a deeper level (Nersessian, 1989). Presented with good models, students may become more passive observers, and forfeit the benefits of actively engaging in the learning process (Gureckis & Markant, 2012) or experiencing "the generation effect" (Bertsch, Pesta, Wiscott, & McDaniel, 2007; Rosner, Elman, & Shimamura, 2013; Slamecka & Graf, 1978). The explanations they produce may be in the right ballpark, but miss the point in decisive ways.

Acknowledgments

We thank Dillon Guludec for RA assistance. SD was supported in part by the Survival and Flourishing Fund.

References

Anderson, L. R., & Holt, C. A. (1997). Information cascades in the laboratory. *The American Economic Review*, *87*(5), 847–862. Retrieved 2024-02-01, from `http://www.jstor.org/stable/2951328`

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*, 201–210. doi: 10.3758/BF03193441

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, *126*, 137–154.

Douven, I., & Mirabile, P. (2018, 02). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology Learning Memory and Cognition*, *44*. doi: 10.1037/xlm0000545

Galton, F. (1907). Vox Populi. *Nature*, *1949*(75), 450–451.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., & Kellerer, W. (2010). Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd wonference on online social networks* (p. 3). USA: USENIX Association.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*, 464–481. doi: doi.org/10.1177/1745691612454304

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Hills, T., Todd, P., & Goldstone, R. (2008, 09). Search in external and internal spaces evidence for generalized cognitive search processes. *Psychological science*, *19*, 802-8. doi: 10.1111/j.1467-9280.2008.02160.x

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46-54. doi: https://doi.org/10.1016/j.tics.2014.10.004

Lee, Y.-J., Hosanagar, K., & Tan, Y. (2015). Do i follow my friends or the crowd? information cascades in online movie ratings. *Management Science*, *61*(9), 2241-2258. doi: 10.1287/mnsc.2014.2082

Liquin, E. G., & Lombrozo, T. (2022). Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology*, *132*, 101453. doi: https://doi.org/10.1016/j.cogpsych.2021.101453

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232-257. doi: https://doi.org/10.1016/j.cogpsych.2006.09.006

Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. In *Proceedings of the 14th annual conference companion on genetic and evolutionary computation* (p. 1379–1392). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2330784.2330946

MacKay, C. (1852). *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds*. London: Office of the National Illustrated Library.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3489–3501.

Na, R. W., & DeDeo, S. (2022). The Diversity of Argument-Making in the Wild: from Assumptions and Definitions to Causation and Anecdote in Reddit's "Change My View"'. In H. R. . V. R. J. Culbertson A. Perfors (Ed.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 969–975).

Nersessian, N. J. (1989). Conceptual change in science and in science education. *Synthese*, *80*(1), 163–183.

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex*, *49*, 1901–1909. doi: doi:10.1016/j.cortex.2012.09.009

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604. doi: doi: 10.1037/0278-7393.4.6.592

Sulik, J., van Paridon, J., & Lupyan, G. (2023). Explanations in the wild. *Cognition*, *237*, 105464.

Talisse, R., & Aikin, S. F. (2006). Two forms of the straw man. *Argumentation*, *20*, 345–352.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon Press.

Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences*, *24*(12), 981–993.

Woodward, J., & Ross, L. (2021). Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University.