Title

Learning From Surprise: Harnessing a Metacognitive Surprise Signal to Build and Adapt Belief Networks.

Journal

Topics in cognitive science, 11(1)

ISSN

1756-8757

Authors

Munnich, Edward
Ranney, Michael A

Publication Date

2019

DOI

10.1111/tops.12397

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 11 (2019) 164–177

ISSN:1756-8765 online
DOI: 10.1111/tops.12397

This article is part of the topic "The Ubiquity of Surprise: Developments in Theory, Converging Evidence, and Implications for Cognition," Edward Munnich, Meadhbh Foster and Mark Keane (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

Learning From Surprise: Harnessing a Metacognitive Surprise Signal to Build and Adapt Belief Networks

Edward Munnich,[a] Michael A. Ranney[b]

[a]*Department of Psychology, University of San Francisco*
[b]*Graduate School of Education* and *Department of Psychology, University of California, Berkeley*

Received 4 August 2016; received in revised form 1 October 2018; accepted 22 October 2018

Abstract

One's level of surprise can be thought of as a metacognitive signal indicating how well one can explain new information. We discuss literature on how this signal can be used adaptively to build, and, when necessary, reorganize belief networks. We present challenges in the use of a surprise signal, such as hindsight bias and the tendency to equate difficulty with implausibility, and point to evidence suggesting that one can overcome these challenges through consideration of alternative outcomes—especially before receiving feedback on actual outcomes—and by calibrating task difficulty with one's knowledge level. As such, we propose that a major function of education—broadly construed as the work of teachers, journalists, parents, etc.—is to assist learners in using their metacognitive surprise signals to facilitate the building and adaptation of belief networks.

Keywords: Surprise; Belief revision; Metacognition; Learning; Reasoning

1. Introduction

Surprise can play a powerful role in learning, and two broad sets of theories explain how this might happen. One set of theories is based on the extent to which one's expectations are violated—the sense of surprise corresponding to the statement "I didn't *expect*

Correspondence should be sent to Edward Munnich, Department of Psychology, University of San Francisco, San Francisco, CA 94117. E-mail: emunnich@usfca.edu

that to happen." These theories are supported by findings that surprise facilitates learning, for example, by providing an index of how well existing schemas match observed outcomes (e.g., Meyer, Reisenzein, & Schützwohl, 1997), by triggering analysis of an event that leads to other epistemic states that facilitate learning (Valdesolo, Shtulman, & Baron, 2017), or by drawing attention to unexpected events to increase long-term memory storage (e.g, Ranganath & Rainer, 2003). Another set of theories build on Kahneman and Miller's (1986) observation that we try to make sense of events as we observe them, and that we are surprised to the extent that we cannot find an explanation—the sense of surprise corresponding to the statement "I can't *explain* why that happened." These theories are supported by findings that, independently of one's probability estimates, surprise depends on whether one has a ready explanation for an event (e.g., Maguire, Maguire, & Keane, 2011), and that surprise can serve as a metacognitive signal of the difficulty of generating explanations (Foster & Keane, 2015). There is no necessary contradiction between these two sets of theories; indeed, Pezzo's (2003) two-stage model includes both initial surprise, stemming from violation of expectations, and resultant surprise, arising from failure to make sense of an event (see Munnich, Foster, & Keane, this issue, for more detailed discussion of these sets of theories). This article focuses on how surprise might assist a learner in improving social and scientific explanations, so the latter set of theories, emphasizing the role of explanation in surprise, are most directly relevant.

Foster and Keane (2015) reported a series of experiments that manipulated the difficulty of explaining outcomes by varying (a) scenario familiarity (e.g., walking home one day, one realizes one's wallet [familiar] vs. belt [less familiar] is missing); (b) whether a partial explanation was provided to make the task easier; (c) whether one was previously prompted to explain an outcome; (d) whether the event was routine or exceptional for the person involved; (e) how many explanations were solicited; and (f) whether priming cues were helpful or misleading. In all cases, surprise rose with increasing explanatory difficulty, suggesting that the level of surprise one experiences is a metacognitive signal—corresponding to Pezzo's (2003) resultant surprise—that indicates the amount of cognitive work one has done.[1] This metacognitive surprise signal can be represented by a coherence metric in a constraint-satisfaction network. For example, ECHO (Ranney & Thagard, 1988; Thagard, 1989) represents belief as propositional nodes in a network, connected by excitatory and inhibitory connections depending on whether propositions support or contradict each other. Adding a new proposition triggers spreading activation, and a coherence metric is computed, based on the fit between new propositions and existing propositions. As such, a coherence score is low when surprise is high —that is, when new propositions are difficult to explain in light of existing beliefs.

2. Overview of learning scenarios, starting from Scenario A

Fig. 1 illustrates how people might use a surprise signal in different learning scenarios. On the left are scenarios in which one experiences a weak surprise signal, and on the right are scenarios in which the surprise signal is strong. In Scenario A, surprise is low because

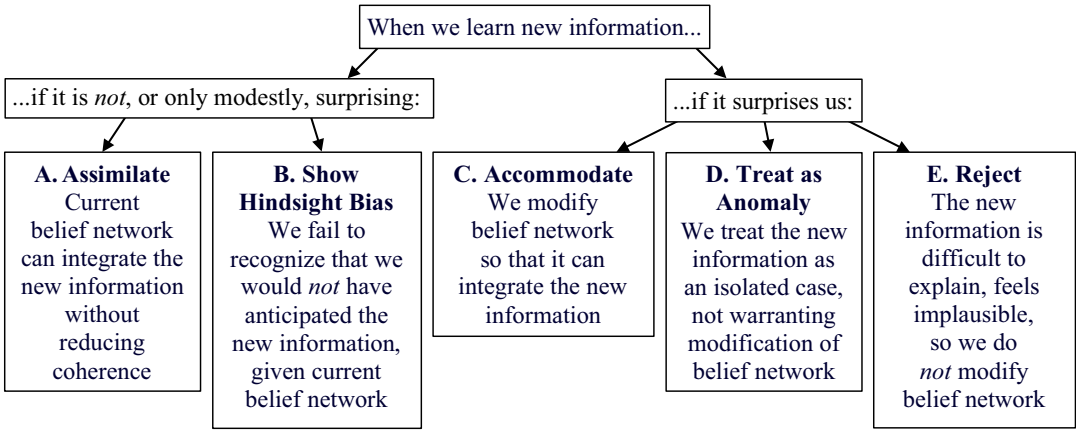

Fig. 1. Scenarios in which people might use, or fail to optimally use, a metacognitive surprise signal in learning.

one's belief network aligns reasonably well with the causal structure of events in the world, and new information is readily assimilated (Piaget, 1977), so coherence is high and surprise is low. Since Scenario A involves cases in which a learner is neither surprised by new information, nor should be, we mention it only as a reference point for other scenarios and will not explore it in depth. By contrast, Scenario B illustrates the case of hindsight bias, in which surprise is relatively low, which is potentially misleading. Turning to the right side of Fig. 1, Scenario C illustrates the case of new information that conflicts with existing beliefs, is markedly surprising, and so triggers accommodation (Piaget, 1977; Ranney & Clark, 2016; Ranney, Munnich, & Lamprey, 2016; Ranney, Shonman, Fricke, Lamprey, & Kumar, in press). In contrast to this, Scenarios D and E represent ways in which high surprise could be misleading: On the one hand, information is surprising in Scenario D because it represents an anomalous case, a misunderstanding of information, or a deception that should be regarded with skepticism (as with "fake news," one reviewer pointed out), and does not warrant accommodation. On the other hand, in Scenario E, people sometimes reject surprising information that could be helpful in reorganizing their belief networks, because it is difficult to make sense of it. We posit that a major function of education (broadly including the roles of teachers, journalists, parents, mentors, etc.) is to teach strategies to avoid Scenario B and to find the right balance between Scenarios D and E, so learners can harness their surprise signals to achieve Scenarios A and C, as appropriate. Let us now explore evidence regarding the scenarios in depth.

3. How to learn when we are not as surprised as we should be (Scenario B)

In contrast to the adaptive use of a low level of surprise in Scenario A, Scenario B represents hindsight bias, in which one learns new information and, unjustifiably, believes

that one would have "known it all along" (e.g., Ash, 2009; Fischhoff, 1975; Pezzo, 2003; Roese & Olson, 1996; Roese & Vohs, 2012; Schkade & Kilbourne, 1991). In a comprehensive review of the hindsight bias literature, Roese and Vohs (2012) discussed three broad sources of the phenomenon: (a) selective memory of what one has learned to be true, (b) misattributing one's ease of understanding of a known outcome as likelihood that one would have anticipated the outcome, and (c) motivations to avoid blame and see the world as orderly. Notably, hindsight bias from any of these three sources would correspond to a having a relatively sparse belief network, composed mostly or entirely of propositions that support a known outcome. In this case, a constraint-satisfaction model would yield a relatively high coherence score, corresponding to lower surprise than in Scenario C, but only because Scenario B excludes nodes that might raise conflicts.

As a possible antidote, several lines of research have used explanatory "foresight" activities that yield diverse hypotheses *before* one learns new information. In a constraint-satisfaction model, this would reflect a richer network of propositions with greater possibilities for conflict than in Scenario B, so that lower coherence scores and reorganization of belief networks are possible where appropriate. Slovic and Fischhoff (1977; Experiment 3) found that participants showed hindsight bias regarding how surprised they were about experimental results reported in a research paper. For each paper, *hindsight* participants read the introduction, method, and results sections, and rated (a) how surprising each outcome was and (b) how likely it was to replicate; *foresight* participants read only the introduction and method, and rated the projected surprisingness and likelihood of replication of two possible outcomes. For five of the six experiments, foresight participants indicated reliably more surprise than hindsight participants for the outcomes that hindsight participants read, and, for a different five of the six experiments, foresight participants indicated a lower likelihood of replication than hindsight participants indicated for those outcomes. Given recent failures of some widely cited psychological studies to replicate (e.g., Open Science Collaboration, 2015), we are reminded of the importance of keeping an open mind regarding a replication's likelihood; Slovic and Fischhoff's findings suggest that considering alternative outcomes before readings the results of a study might help in this regard.

Further evidence regarding hindsight bias comes from research on clinical diagnosis. Arkes, Faust, Guilmette, and Hart (1988) presented neuropsychologists with a case history, for which a foresight group estimated the likelihoods of three different diagnoses. Their responses were compared to those of three hindsight groups, each of which were told the "correct" diagnosis (one of the three diagnoses the foresight group saw per hindsight group), and then indicated the likelihoods they would have predicted for each of the three diagnoses, had they not been told which was correct. First, Arkes et al. replicated hindsight bias—hindsight participants systematically indicated higher likelihoods of making correct diagnoses than foresight participants—suggesting that when clinicians learn diagnoses without considering alternatives, they miss an opportunity to learn and to become better diagnosticians. Interestingly, when new groups of hindsight and foresight participants were prompted for one reason for each of the three diagnoses, they showed considerably less bias, presumably because they were reminded why they would have

considered each of the diagnoses. However, even after providing explanations for each alternative diagnosis, hindsight participants favored the given diagnosis for two out of three conditions and did not converge with the results of foresight participants. Therefore, although bias can be reduced in hindsight, considering alternatives in foresight appears to be more effective in preserving the possibility to be surprised by that which one would not have predicted.

4. Accommodation (Scenario C)

Before we discuss Scenario C, we ask you to consider the following: In 2005, there were 145 traffic fatalities per million U.S. residents. With this statistic in mind, please make a note of your answers to the following questions: (a) What is your best estimate of the U.S. traffic fatality rate in *2010*? (We provide feedback below.) (b) What factors do you believe caused traffic fatalities to increase or decrease between 2005 and 2010? In our laboratories, we have considered how base rate statistics—such as traffic fatality rates—can shape beliefs and preferences. Base rates provide concise generalizations across many outcomes (Ranney, Schank, Hoadley, & Neff, 1996), and they can be used to make inferences about sets of outcomes, provided that they are presented in an understandable format (e.g., frequencies; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). In parallel to studies that explicitly ask participants to consider alternative outcomes, asking participants to estimate a base rate implicitly requires them to consider alternative outcomes (e.g., the base rate for traffic fatalities depends not only on outcomes in which people died in car crashes, but also on outcomes that do not lead to traffic fatalities). Of greatest relevance to this paper, surprising base rates can serve as catalysts for accommodative learning (e.g., through the data priority principle in the Theory of Explanatory Coherence; Thagard, 1989) and can also transform people's preferences and policy decisions (e.g., Garcia de Osuna, Ranney, & Nelson, 2004; Munnich, Ranney, Nelson, Garcia de Osuna, & Brazil, 2003; Ranney, Cheng, Nelson, & Garcia de Osuna, 2001). Within a constraint-satisfaction network, this process could be represented as adding new propositional nodes, changing relationships among such nodes, and/or reweighting pre-existing links among nodes—all to produce a new equilibrium (e.g., Ranney & Thagard, 1988; Thagard, 1989).

Now, please return to your 2010 traffic-fatality estimate. The true value was 106 traffic fatalities per million U.S. residents—as compared to 145 per million in 2005. This amounts to a 27% decrease in traffic fatalities, representing 12,000 lives annually. Are you surprised by this statistic? Munnich, Milazzo, Stannard, and Rainford (2014) found that 79% of an Amazon Mechanical Turk sample of U.S. residents believed that traffic fatalities had *risen* from 2005 to 2010, and their median estimate was 190, which would amount to a 36% increase. When asked to explain their reasons for their estimates, the foresight group's most common responses were increases in drunk driving, cell phones, and texting while driving. Upon learning that the fatality rate had actually fallen sharply, their median level of surprise was a "3" on a 5-Point Likert Scale—corresponding to

"very surprised." By contrast, a hindsight group, who saw the actual 2010 statistic at the outset of the experiment, without having estimated or considered reasons for surprise, were reliably less surprised by the true statistics, indicating a median surprise of "2" corresponding to "slightly surprised." To the extent that the foresight group was surprised by the feedback they received, they were confronted with the fact that they had *not* "known it all along." This echoes the findings of Arkes et al. (1988) regarding clinicians making predictions before they were told that one diagnosis was correct.

Given evidence that the possibility of being surprised can be preserved with foresight activities focused on explanation, such activities might provide learners with an opportunity to undergo an accommodative change in their belief networks regarding what contributes to, or mitigates, a phenomenon like traffic fatalities. To test this, Munnich, Ranney, and Song (2007) asked 95 eighth-grade Algebra I students to estimate widely-ranging statistics (e.g., voter registration, immigration, incarceration, athletes' salaries). These topics were familiar and interesting to students, and they were selected to vary in how surprising the true statistics would likely be. Participants (a) estimated two statistics per day over a 4-day period while indicating what they would prefer the values to be before receiving feedback, (b) received feedback (i.e., actual values), (c) indicated how surprised they were on a 5-point Likert scale, and (d) indicated their (possibly-changed) preferences. Either 8 days or 12 weeks later, as a delayed posttest, participants attempted to recall the statistics and indicate their current (post-delay) preferences. Echoing other findings from our labs (e.g., Clark & Ranney, 2010; Garcia de Osuna et al., 2004; Munnich et al., 2003; Ranney et al., 2001), providing true statistics as feedback immediately influenced participants' preferences. Although a plurality of participants' responses (44.5%; Munnich et al., 2007) maintained the preference they held before feedback, when participants changed their numerical preferences, they did so reliably more often in the direction of the feedback value than away from it. Over both 8 days and 12 weeks, participants' recall of the statistics and their preferences for those quantities moved reliably in the direction of the feedback they received, reflecting long-term shifts in their beliefs and preferences. Most important regarding surprise and learning, students' surprise ratings reliably predicted recall accuracy over 8 days and was a marginally significant predictor of preferences changing in the direction of the statistics recalled after the 8-day period. Notably, this intervention took roughly 5 min. per day for 4 days, indicating that even a small amount of time devoted to foresight activities can have relatively long-term learning benefits (Munnich et al., 2007). However, we note that surprise only predicted recall and preference over an 8-day period, so it seems important for learners to subsequently build upon such activities to capitalize on the utility of surprise.

Munnich et al. (2007) showed that accommodative learning is often accompanied by a change in one's preferences (e.g., one might realize that fewer people vote than one thought, and would favor more drastic action to increase voting than one originally indicated). Although the point of these tasks was not to change people's preferences in a given way, when we observe changes, it provides converging evidence that accommodative learning has taken place: People have adjusted their belief networks to more closely reflect outcomes in the world, and correspondingly, their preferences and/or the policies

they support may shift to cohere with their new understandings. In other studies (Garcia de Osuna et al., 2004; Munnich et al., 2003; Ranney et al., 2001), we typically probed topics such as abortion and capital punishment, for which a minority of participants initially preferred the rates to be zero; that is, some people's preferences (at least initially) reflected their *ideal* world, whereas others' preferences reflected a *better* world that they thought was achievable. "Better-world" participants often changed their preferences to reflect a realistic improvement on what they learned the statistics to be, and even "ideal-world" participants showed changes in the policies they would support (e.g., many who supported legal abortion but preferred zero abortions, called for improvements in birth control technology after learning an abortion statistic that was considerably higher than they expected; Garcia de Osuna et al., 2004). Though they shifted in different ways, both ideal-world and better-world participants show, through these shifts, that they have accommodated new information in their belief networks. The consequences of accommodation were further illustrated by Rinne, Ranney, and Lurie (2006), who found shifts in funding allocations for various diseases (e.g., whether to spend more money on preventing breast cancer or heart disease in women), after participants received surprising feedback on the diseases' relative occurrence.

A subsequent study by Clark and Ranney (2010) pointed to dual mechanisms of surprise in learning, depending on whether participants had an episodic memory of having learned a statistic. They asked participants to estimate a statistic, then provided participants with the statistic's true value and solicited participants' surprise levels. One day later, participants recalled/estimated statistics from the day before as well as novel statistics and indicated how sure they were that they had seen the statistic the day before. Those who were more surprised by feedback were more likely to report episodic memories of learning statistics, and those who reported episodic memories remembered the numbers better. In addition, surprise was an independent predictor of semantic memory—even those who had little recollection of having seen the statistic remembered it better when they were more surprised. Building on the results discussed above, this study provides the beginnings of an understanding of the mechanisms by which surprise can lead to persisting accommodative shifts in knowledge.

5. When surprising information should *not* lead to belief revision (Scenario D)

Although surprise often leads to adaptive belief revision, in some cases, a surprising outcome is anomalous, misleading, or deceiving, and it does not warrant belief revision. In such cases, it is wise to have a certain amount of skepticism about information that surprises us. Dawson et al. (1988) found considerable hindsight bias among both expert and novice physicians who learned of a diagnosis and stated how likely they would have been to have generated that diagnosis. However, there was one notable exception: *Expert* physicians did *not* show hindsight bias when they learned that the true disease was a rare one: They recognized that they would not likely have made such a diagnosis; as experts, even if they are surprised by a particular outcome, their prior experience prevented them

from believing that they could have predicted that outcome. This awareness might also shield experts from overreacting with an accommodative change in response to a single anomalous case.

Ironically, if novice physicians' hindsight bias keeps them from being surprised by anomalies, it might at least prevent them from unjustified accommodative change. This should be a warning to educators who seek to minimize learners' hindsight bias by challenging them to explain alternative outcomes, at a point when learners' expertise is not sufficient to recognize anomalies. In constraint satisfaction terms, novice learners have fewer other node-link complexes than experts, making it more difficult to inhibit an anomalous proposition from triggering reorganization of a network. One solution to this challenge in modeling terms would be to assign lower weight to single propositions that are not supported by other propositions; the parallel for educators would be to encourage learners not to give too much weight to outcomes that seem implausible. However, down-weighting propositions based on implausibility points to a different set of challenges, corresponding to Scenario E.

6. How to learn from surprising information that is difficult to grasp (Scenario E)

To the extent that a metacognitive surprise signal tracks cognitive difficulty, high surprise should correspond to deeper, more systematic, processing (e.g., Diemand-Yauman, Oppenheimer, & Vaughan, 2011). This is consistent with Tiedens and Linton's (2001) finding that participants primed with uncertain moods—including surprise—showed more systematic processing than those primed with certain moods—such as anger. As more systematic processing would likely lead one to be critical of one's assumptions, these findings echo the role of surprise in accommodative learning described earlier. As such, surprise could signal to learners that they are experiencing a "desirable difficulty" (Bjork, 1994) that would be helpful for revising beliefs. Unfortunately, when an outcome is very difficult to explain, people tend to think of it as less plausible (e.g., Hirt & Markman, 1995; Sanna, Schwarz, & Small, 2002). In a constraint-satisfaction model, decreasing weights of seemingly implausible propositions—which would play a helpful role in Scenario D—could also lead the system to reject valuable propositions.

Like Foster and Keane (2015), Sanna et al. (2002) manipulated the number of explanations participants were asked to provide. Participants read a synopsis of the British-Gurkha War, including advantages each side had, and learned that the British actually won the war. A control group provided baseline estimates of how likely each side was to have won the war, and they estimated that the British were slightly more likely to have won. One experimental group was asked to provide two thoughts on why the British were more likely to win the war, and a second group provided two reasons why the Gurkha could have won the war. When asked to judge the probability of each side's winning, each group judged its respective side to have had a higher likelihood of winning than was indicated by controls. However, when two additional experimental groups provided ten such thoughts for the alternate sides, they both rated the task to be

more difficult than the two-thoughts groups had indicated, and they judged the probability of its respective side winning to be *lower* than the estimate provided by controls. These findings indicated that judged implausibility of an outcome reflects the task's perceived difficulty. In a second experiment, Sanna et al. found converging evidence that difficulty had suggested implausibility; they asked one group of participants to contract their brows—an expression associated with a difficult task—while considering one of the possible outcomes in the British–Gurkha War: Those who contracted their brows found the task more difficult, and they considered the outcome to be *less* likely than did controls who did not contract their brows. This suggests that a strong surprise signal arising from cognitive difficulty might inhibit even appropriate, feedback-driven belief revision.

As illustrated by Scenario D above, healthy skepticism is useful and we should not revise our beliefs every time we are surprised, so connecting difficulty with implausibility is clearly adaptive in some cases. However, for cases in which one can benefit from engaging with difficult tasks, following Sanna et al. (2002), it might be helpful for educators to ask novice learners to explain just one or two (rather than many) alternative outcomes at first. Moreover, following Foster and Keane (2015), it might be helpful to provide hints or partial explanations, in order to keep difficulty and surprise within ranges that promote learning and gradually increase wisdom (Ranney et al., 2016; Ranney, Shonman, Fricke, Kumar, & Lamprey, in press; also see especially Ranney & Clark, 2016, regarding global warming's surprising mechanism-explanation).

7. Internalizing use of the surprise signal

So far, we have considered ways in which learners can use a metacognitive surprise signal productively when prompted with alternative outcomes. Going a step further, Hirt, Kardes, and Markman (2004) found evidence of transfer: Considering alternatives in one domain (football or sitcom rankings) prompted participants to consider alternatives in a different domain (basketball rankings). The duration of this effect is unclear, and one possibility is that participants were only primed to think of alternatives for the duration of the experiment. However, another possibility is that such activities lead participants to adopt strategies of considering alternatives, which would help them to use a surprise signal more productively at a later date, even when they are not explicitly prompted to consider alternative outcomes.

Several studies in our labs have focused on students' recruitment of surprising information in their reasoning (Ranney et al., 2016) and considered whether there is a long-term effect on learners' openness to alternative perspectives. For example, Munnich, Ranney, and Appel (2004) asked students to engage with the alternative perspectives of classmates regarding statistics relevant to career choices (e.g., college vs. high school graduates' incomes) and public policy issues (e.g., the U.S. poverty line and oil imports). Each night, as homework, students individually estimated and provided preferences for one statistic. In class the next day, they explained their estimates and preferences, and listened

to those of classmates—first in small groups, and later with the whole class, when each group presented collective estimates and preferences they had agreed on. We hypothesized that repeated discussion of a variety of alternative possible outcomes suggested by classmates, would lead students to spontaneously consider alternatives, and therefore to show superior estimation for novel items.

As expected, the intervention class's estimation accuracy improved from pre- to posttest, whereas a control class that did not receive the intervention showed no improvement, indicating that the intervention increased students' estimation accuracy. (For the intervention's full question-list, as well as pre- and posttests, see table 1 of Munnich et al., 2004.) Estimation improvements extended not only to near transfer items (e.g., an intervention item on U.S. population may have improved posttest estimates of California's population), but also to far transfer items not mentioned in the intervention (e.g., average sleep-hours). To assess the mechanism behind this improvement, a researcher who was blind to each student's class (i.e., control vs. intervention) interviewed volunteers from each group several months after the curriculum and found greater richness in the strategies employed by the intervention class on novel items (Ganpule, 2005). Collectively, these findings, and similar findings with graduate journalism students (Ranney et al., 2008; Yarnall & Ranney, 2017), indicate that learners can internalize the consideration of alternatives, and learn to process social and personal issues more systematically, in the course of a relatively brief intervention (e.g., 12% of a 10-week class time for high school students). Once one has internalized this process, one might spontaneously invoke alternative possible outcomes to avoid hindsight bias, embrace cognitive difficulty as a chance to enrich one's belief network, and, as one's network grows in richness, develop the ability to appropriately reject anomalies. In short, it seems that this type of intervention moves one toward an adaptive use of the metacognitive surprise signal, as expressed in Scenarios A and C.

8. Conclusion

We have reviewed evidence that the metacognitive surprise signal can lead to restructuring of belief networks when necessary to (a) generate superior predictions, and (b) form preferences that are more closely aligned with true information. Alternatively, when one experiences little or no surprise because one actually would have anticipated an outcome, this indicates that one is converging on a well-adapted belief network in a particular domain, with no need for reorganization. In both cases, the surprise signal might lead learners to form more accurate representations of mechanisms underlying decisions they make as voters, consumers, etc., and arrive at better-calibrated preferences. We have considered scenarios in which one should or should not be surprised, reviewed techniques that facilitate use of the metacognitive surprise signal, and suggested how use of this signal might be internalized. As an index of cognitive work, this surprise signal could be used by learners in the way that a pain signal is used by athletes as an index of physical work. Seasoned athletes understand that pain is indicative of building muscle strength,

and, by analogy, expert learners might welcome surprises as indicators that they have a chance to make a breakthrough in understanding a new subject area. Moreover, just as athletes can gauge their progress toward workout goals by the decrease in pain after successive workouts, learners can observe that they are getting closer to mastery of a topic as they progressively encounter fewer surprises.

As indicated at various points in this paper, predictions of how the surprise signal functions could be modeled by a coherence measure in a constraint-satisfaction network (e.g., ECHO; Ranney & Thagard, 1988; Thagard, 1989). Moreover, developments in metacognitive signal detection (Barrett, Dienes, & Seth, 2013) could provide techniques to measure the surprise signal, and thereby allow us to more precisely specify its role in learning. For now, we conclude that one can learn to use the surprise signal adaptively in building belief networks, through activities that (a) involve consideration of alternative outcomes, especially in foresight, (b) are at an appropriate level of cognitive difficulty to trigger systematic processing without necessarily suggesting implausibility, and (c) draw on a learner's sense of implausibility to avoid accommodation when an outcome is anomalous.

Note

1. Here, we distinguish a *metacognitive* surprise signal from various *neural* surprise signals discussed elsewhere (e.g., Alexander & Brown, 2011; this volume; Hayden, Heilbronner, Pearson, & Platt, 2011; Kawaguchi et al., 2015).

References

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921.

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*(2), 305–307. https://doi.org/10.1037/0021-9010.73.2.305.

Ash, I. K. (2009). Surprise, memory, and retrospective judgment making: Testing cognitive reconstruction theories of the hindsight bias effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 916–933. https://doi.org/10.1037/a0015504.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Clark, D., & Ranney, M. A. (2010). Known knowns and unknown knowns: Multiple memory routes to improved numerical estimation. In K. Gomez, L. Lyons & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the Ninth International Conference of the Learning Sciences, Volume 1-Full Papers* (pp. 460–467). Chicago, IL: International Society of the Learning Sciences, Inc.

Dawson, N., Arkes, H., Siciliano, C., Blinkhorn, R., Lakshmanan, M., & Petrelli, M. (1988). Hindsight bias: An impediment to accurate probability estimation in clinicopathologic conferences. *Medical Decision Making*, *8*(4), 259–264. https://doi.org/10.1177/0272989X8800800406.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 114–118. https://doi.org/10.1016/j.cognition.2010.09.012.

Fischhoff, B. (1975). Hindsight-foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288–299. https://doi.org/10.1136/qhc.12.4.304.

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116. https://doi.org/10.1016/j.cogpsych.2015.08.004.

Ganpule, S. (2005). *Strategy use in numerical estimation: Investigating the effects of an EPIC curriculum*. Unpublished Master's Project, University of California, Berkeley.

Garcia de Osuna, J., Ranney, M., & Nelson, J. (2004). Qualitative and quantitative effects of surprise: (Mis) estimates, rationales, and feedback-induced preference changes while considering abortion. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 422–427). Mahwah, NJ: Erlbaum.

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

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of Neuroscience*, *31*(11), 4178–4187. https://doi.org/10.1523/JNEUROSCI.4652-10.2011.

Hirt, E. R., Kardes, F. R., & Markman, K. D. (2004). Activating a mental simulation mind-set through generation of alternatives: Implications for debiasing in related and unrelated domains. *Journal of Experimental Social Psychology*, *40*(3), 374–383. https://doi.org/10.1016/j.jesp.2003.07.009.

Hirt, E., & Markman, K. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*, 1069–1086.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136.

Kawaguchi, N., Sakamoto, K., Saito, N., Furusawa, Y., Tanji, J., Aoki, M., & Mushiake, H. (2015). Surprise signals in the supplementary eye field: Rectified prediction errors drive exploration-exploitation transitions. *Journal of Neurophysiology*, *113*(3), 1001–1014. https://doi.org/10.1152/jn.00128.2014.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: LMC*, *37*, 176–186. https://doi.org/10.1037/a0021609.

Meyer, W., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251–274. https://doi.org/10.1023/A:1024422330338.

Munnich, E., Milazzo, J., Stannard, J., & Rainford, K. (2014, July). Can causal sense-making benefit foresight, rather than biasing hindsight? In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Munnich, E., Ranney, M., & Appel, D. (2004). Numerically-driven inferencing in instruction: The relatively broad transfer of estimation skills. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum and Association.

Munnich, E., Ranney, M., Nelson, J., Garcia de Osuna, J., & Brazil, N. (2003). Policy shift through Numerically-Driven Inferencing: An EPIC experiment about when base rates matter. In R. Alterman, & D. Kirsch (Eds.), *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society* (pp. 834–839) Mahwah, NJ: Erlbaum.

Munnich, E., Ranney, M., & Song, M. (2007). Surprise, surprise: The role of surprising numerical feedback in belief change. In D. S. McNamara & G. Trafton (Eds.) *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 503–508). Mahwah, NJ: Erlbaum.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), 1–8. https://doi.org/10.1126/science.aac4716.

Pezzo, M. (2003). Surprise, defence, or making sense: What removes hindsight bias? *Memory*, *11*(4/5), 421–441. https://doi.org/10.1080/09658210244000603.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures* (A. Rosin, Trans.). New York: Viking (Original work published 1975).

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*(3), 193–202. https://doi.org/10.1038/nrn1052.

Ranney, M., Cheng, F., Nelson, J., & Garcia de Osuna, J. (2001). *Numerically driven inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates*. Paper presented at the Annual Meeting of the Society for Judgment & Decision Making.

Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75. https://doi.org/10.1111/tops.12187.

Ranney, M., Munnich, E., & Lamprey, L. (2016). Increased wisdom from the ashes of ignorance and surprise: Numerically-driven inferencing, global warming, and other exemplar realms. In B. H. Ross (Ed.), *The psychology of learning and motivation*, *65*, 129–182. New York: Elsevier. https://doi.org/10.1016/bs.plm.2016.03.005

Ranney, M., Rinne, L. F., Yarnall, L., Munnich, E., Miratirx, L., & Schank, P. (2008). Designing and assessing numeracy training for journalists: Toward improving quantitative reasoning among media consumers. In P. A. Kirschner, F. Prins, V. Jonker & G. Kanselaar (Eds.), *International Perspectives in the Learning Sciences: Proceedings of the Eighth International Conference for the Learning Sciences* (pp. 2-246–2-253). Utrecht, The Netherlands: International Society of the Learning Sciences, Inc.

Ranney, M., Schank, P., Hoadley, C., & Neff, J. (1996). "I know one when I see one": How (much) do hypotheses differ from evidence? In R. Fidel, B. H. Kwasnik, C. Beghtol & P. J. Smith (Eds.) *Advances in classification research: Vol. 5*. (ASIS Monograph Series; pp. 141–158, etc.) Medford, NJ: Learned Information. https://doi.org/10.7152/acro.v5i1.13783

Ranney, M. A., Shonman, M., Fricke, K., Kumar, P., & Lamprey, L. N. (in press). Information that boosts normative global warming acceptance without polarization: Toward J. S. Mill's political ethology of national character. In R. Samuels & D. Wilkenfeld (Eds.) *Anthology on the experimental philosophy of science* New York: Bloomsbury (In Bloomsbury's Advances in Experimental Philosophy series.).

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In L. Patel, & G. J. Groen (Eds.), *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426–432). Hillsdale, NJ: Erlbaum.

Rinne, L., Ranney, M. A., & Lurie, N. (2006). Estimation as a catalyst for numeracy: Micro-interventions that increase the use of numerical information in decision-making. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *Proceedings of the 7th International Conference on Learning Sciences* (pp. 571–577). Mahwah, NJ: Erlbaum.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology*, *32*, 197–227.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303.

Sanna, L. J., Schwarz, N., & Small, E. M. (2002). Accessibility experiences and the hindsight bias: I knew it all along versus it could never have happened. *Memory & Cognition*, *30*(8), 1288–1296. https://doi.org/10.3758/BF03213410.

Schkade, D. A., & Kilbourne, L. M. (1991). Expectation-outcome consistency and hindsight bias. *Organizational Behavior and Human Decision Processes*, *49*, 105–123. https://doi.org/10.1016/0749-5978(91)90044-T.

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: HPP*, *3*, 544–551. https://doi.org/10.1037/0096-1523.3.4.544.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–502. https://doi.org/10.1017/S0140525X00057046.

Tiedens, L. Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology*, *81*(6), 973–988. https://doi.org/10.1037/0022-3514.81.6.973.

Valdesolo, P., Shtulman, A., & Baron, A. S. (2017). Science is awe-some: The emotional antecedents of science learning. *Emotion Review*, *9*(3), 215–221. https://doi.org/10.1177/1754073916673212.

Yarnall, L., & Ranney, M. A. (2017). Fostering scientific and numerate practices in journalism to support rapid public learning. *Numeracy*, *10* (1), article 3 [30 pages]. Available at: https://doi.org/10.5038/1936-4660.10.1.3 Also Available at http://scholarcommons.usf.edu/numeracy/vol10/iss1/art3