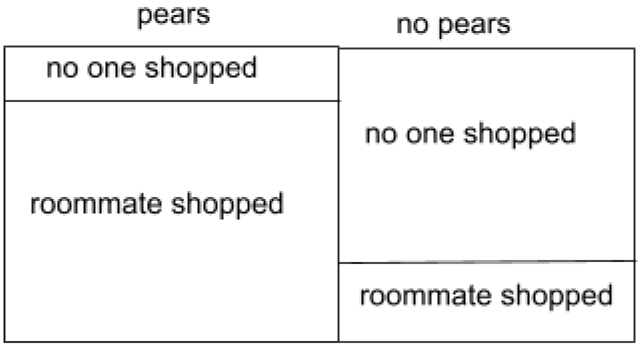
The Problem of New Theories

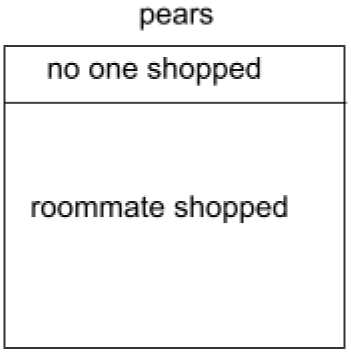
Sara Aronowitz

1. The Problem

The problem of new theories, also known as the problem of awareness growth or new hypotheses, is primarily known as a challenge for subjective Bayesian theories of scientific and individual learning. Secondarily, it is a wider set of challenges for many other theories of learning. Subjective Bayesian theories explain how rational agents take probabilistic expectations and update them by preserving existing relationships while eliminating possibilities through learning. For instance, I might think it's about equally likely to have pears in the fridge as not, and that if I have pears in the fridge, it's highly likely that my roommate went shopping. In the diagram below, the area corresponds to my degree of credence in each hypothesis, such that the total area of the rectangle sums to 1:



My initial credence in the proposition that my roommate went shopping is about .5. Learning that there are pears removes the entire right side of the diagram, but scales my remaining credence such that the entire area still sums to 1.



That is, my new credence that my roommate shopped, which is around .8, is equal to my old credence that she shopped given that there were pears. This form of learning involves removing and redistributing my degree of credence after I learn that something previously considered possible is now impossible. The Bayesian view that we will now challenge is just the above idea, formalized, along with the interpretation that these credences express subjective uncertainty rather than, for instance, the objective likelihood of events.

We can now formulate the first version of the problem. Imagine that I see the pears in the fridge, and notice that they have leaves still attached and no stickers, and have another thought: these could be pears from a farm rather than a store, so maybe my friend Nina, who grows pears, has made a surprise visit. This observation cannot be captured by the operation of removing a part of possible space and redistributing credence. But then how should I update my credences? Call this question of accommodating a previously-unconsidered possibility the **accommodation problem.** Further, we might ask: what exactly happened to me when the sight of the leafy pears raised a new possibility—is that process learning, and what would it mean to rationally consider relevant new possibilities? This is the **learning problem**.

In answering either of these questions, we will need to ask: are there meaningfully different types of cases where I acquire new theories? We can already see the issue of types of new theories in the pear example. When I moved from dividing the world into pears/no pears to store pears/farm pears/no pears, I adopted a strictly more fine-grained view of the world. Call this *refinement*, following Steele & Stefansson (2021). But when I moved from roommate/no one shopped to roommate/no one/Nina shopped, I added a new possibility that falls outside those that had been considered rather than inside. That is, my possibility space expanded. This difference has ramifications for both the accommodation and learning problems, as we'll see.[1]

2. A Brief History

In *The Foundations of Statistics* (1954), Leonard Savage introduced his approach to decision-making by contrasting the saying "look before you leap" with "cross that bridge when you come to it":

> Though the 'Look before you leap" principle is preposterous if carried to extremes, I would none the less argue that it is the proper subject of our further discussion, because to cross one's bridges when one comes to them means to attack relatively simple problems of decision by artificially confining attention to so small a world that the "Look before you leap" principle can be applied there. I am unable to formulate criteria for

[1] The distinction goes back at least to Earman (1992), and usually is treated as a continuum rather than a categorical distinction.

> selecting these small worlds and indeed believe that their selection may be a matter of judgment and experience about which it is impossible to enunciate complete and sharply defined general principles.

Savage's approach, in other words, was to build a decision theory around the assumption that a person plans over the entire world and decides only once. In translating from this highly idealized context to ordinary human circumstances, he thought of us as encountering (and perhaps selecting) many smaller but complete decision problems. Savage's strategy represents a thread in philosophical reasoning that leads to the problem of new theories. His bet, in effect, is that if we are careful enough about constructing decision problems, we do not have to revise the core principles of decision-making in a grand world. The problem of new theories, like its twin problem of old evidence as well as problems with sequential choice, arises when we want to model an episode that crosses between these neat, complete small worlds.

In the tradition of rational choice in economics and other behavioral sciences, the problem of new theories is usually discussed under the heading of unawareness and in tight connection with decision-making. This includes changing awareness of not just features of the world like the varieties of pears but of potential acts of the agent. See for example Halpern & Rego (2006) and Ozbey (2007); Schipper (2012) provides an extensive bibliography.

Philosophers, on the other hand, originally broached the issue of learning new theories in a purely epistemic context, without direct concern for action or specific issues around learning new possibilities about one's own acts. In this form, the problem of new theories was first raised by Glymour (1980) and later picked up by Earman (1992), both of whom are mostly interested in Bayesianism as a theory of scientific confirmation. Glymour's discussion pairs the new theories problem with the *problem of old evidence*, which raises the question of how evidence can be said to confirm a hypothesis if the evidence was acquired before the hypothesis was formulated. This is a problem for Bayesians because, just as in the example above, fitting an old observation to a new theory is not learning in the sense of removing possibilities. Glymour anticipates the response that "an ideal Bayesian would never suffer the embarrassment of a novel theory", but contends that this conflicts with the project of using Bayesian theory to explain even ideal scientific argument.

At the root of Glymour's objection to Bayesianism is the idea that theories are not just distributions of likelihoods over observations, or in his terms, collections of their consequences. Theories are (or provide) explanations. This entails that the relationship between theory and evidence is structural, which is just to say there is an objective sense of fit between the two that goes beyond mere predictive validity. The theory of relativity and a complete table of all its consequences are the same for the Bayesian, but cannot be the same from the scientific perspective. This idea, of a structural, and on his view, objective relationship between theory and evidence is suggestive of a positive answer to the learning problem: shouldn't we be able to construct, or at least locate, a new theory based on the evidence in front of us rather than demand a full arsenal of theories in advance?

Earman (1992) sees the problem of new theories at the heart of a set of difficulties around Bayesianism, where the Bayesian can only save convergence to the truth by baking in a massive amount of non-rationally acquired knowledge. In this vein, he considers the proposal that Bayesianism just applies to normal science, in the Kuhnian sense. His retort: “If a redistribution of probabilities from the catchall is taken to be a definition of a scientific revolution, then such revolutions occur with monotonous frequency and the applicability of Bayesianism threatens to shrink to the vanishing point.”

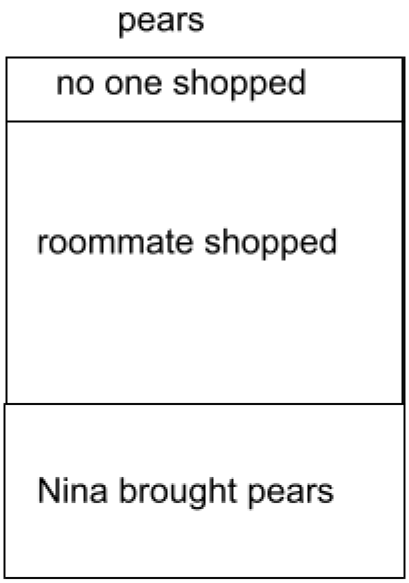
As I’ll now discuss, recent (partial) solutions have been proposed by Wenmackers & Romeijn 2016, Bradley 2017, and Steele & Stefansson (2021). While these explicitly address the accommodation problem, the learning problem seems to be lurking in the background.

3. The Accommodation Problem

If we follow Glymour and Earman in accepting that the problem of new theories cannot be sidestepped by the claim that the ideal agent always already knows all theories, then we are left with the accommodation problem: how should she respond to a new theory when it occurs to her?

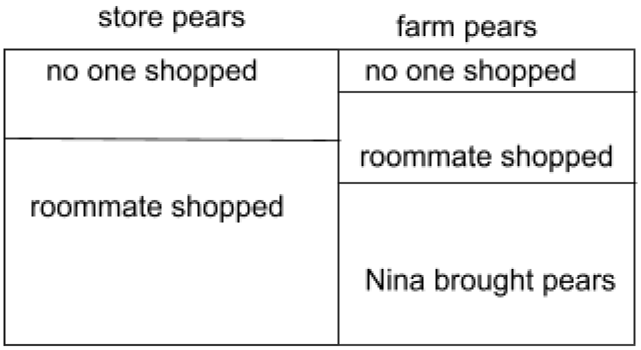
One solution, however incomplete, is the catchall hypothesis. Here we insist that something was missing from my initial table. If I was rational, even though I need not know every hypothesis, I always leave a little room for a catchall, something like “or something else”. Adding the catchall does not fully answer the question, but it does turn expansion into refinement. In our example, my initial space should have included roommate shopped / no one shopped / something else, and then I would adjust to a 4-fold partition roommate shopped / no one shopped / Nina is visiting / something else. The catchall description faces major issues when stretched to cover all cases of expansion, however: it’s hard to say what probability the catchall should have, the mechanism is somewhat ad hoc, and in many cases it seems psychologically inaccurate.

Probably the most prominent answer to the accommodation problem is what is called *reverse Bayesianism* (Karni & Vierø 2013, 2015). Reverse Bayesianism is an attempt to accommodate new theories under the maxim of minimal change. In our first case of regular conditionalization, learning there were pears changed the probability of my roommate having shopped but preserved the ratio between all of the partitions of the left side of the original space. Likewise, the reverse Bayesian intuition is that the addition of a new theory should preserve ratios of existing credences. In our case, let’s imagine that you first considered the possibility of your friend Nina having brought the pears, an expansion. According to reverse Bayesianism, that would produce something like the following update:



The ratio between no one shopping and roommate shopping is conserved even though the new possibility is added. Related accounts are offered by Wenmackers & Romeijn (2016) and Bradley (2017), where the key similarity is adherence to a conservative principle: create as little change as possible by the addition of the new theory.

But as Steele & Steffanson (2021) argue, this answer will not cover all cases because often the introduction of a new hypothesis does intuitively change existing ratios. For instance, in our case, the refinement of farm pears causes a re-weighting between no one shopped/roommate shopped. Originally I described an expansion, but for now, let’s imagine once you think about the possibility of farm pears, it seems to you like your roommate, who always shops at a big supermarket, would not have bought them. So you might have a different ratio on the right side, leading to an overall decreased credence that your roommate shopped just because of the addition of the new hypothesis. Steele & Steffanson propose their own solution, a reframing of the probability assignments to treat the possibility space you consider as potentially fully altered as a result of new hypotheses being added. This fits our case well, since with the addition of the Nina hypothesis, the “no one” hypothesis does seem to mean something new. The final update could be visualized as follows on their view, where any ratio could shift or remain the same:



Allowing for this less conservative change grants a lot of resources in describing cases, since in principle the new hypothesis space can be composed of entirely new components. Conversely, they have so many degrees of freedom in their account that it is hard to get traction in falsifying it.

4. The Learning Problem

So far, we've seen that standard forms of Bayesianism struggle to even accommodate the addition of new theories. But this problem is derivative of a deeper problem. It is intuitively plausible that some new theories are learned, such as in an analysis given by Gentner et al. (1997) where Kepler built up a physical theory in stages through analogy. Other new theories are arrived at on a whim, or given to us by others, or may be arrived at entirely by chance. If this is true, then how we should accommodate a new theory should depend in part on where it came from. Thus, since Bayesians do not in general even aspire to treat the acquisition of theories as part of rationality, it may be that no solution can be given to the accommodation question from within that framework.

What is the alternative? While this entry has discussed the problem for Bayesians, the issue of learning theories is problematic for many frameworks and often the problem is put aside in epistemology following the tradition of treating the "context of discovery" as distinct from confirmation.

Other attempts to provide a solution take inspiration from how humans actually do learn hypotheses in addressing the normative learning problem. Carey (2009) argues that children start with a repertoire of core conceptual systems, such as object recognition and the theory of mind. While her primary aim is to intervene in the nativism vs empiricism debate, her framework also provides a potential solution to the problem of learning new concepts. Using the concepts and structures from these core conceptual systems, she describes a process of bootstrapping to new theories. Bootstrapping here means a kind of development that uses the previous functional parts as a starting point not because of a unique fit with the evidence of the new domain but instead as a necessary first attempt which is then improved and transformed as more evidence accrues about the new domain. Thus on her view, conceptual learning is a combination of strategic use of existing resources and a process of refinement and correction that can start to take place once a potential theory has some substance.

Tenenbaum et al. (2011) also put structure at the core of theory learning, but instead of semi-lateral structural learning through bootstrapping, they appeal to hierarchical structures. On this view, while I might not have a theory for a new domain, I do have knowledge of the abstract features of theories. For instance, the objects in this domain might be grouped by causal origin, by feature-based trees, or something else. I can then use these abstract features to generate likelihoods for the various theories under them, exploring as I go. This approach, like Carey's, only addresses the learning problem by assuming that the thinker has key prior knowledge.

A different, but complementary, approach seeks to understand the role of chance in theory discovery. Wilson et al. (2014) divide exploration, often understood as pure noise, into strategic and random exploration which they suggest might have an optimal balance given parameters of the situation. In philosophy, Thoma (2015) puts rational bounds on random exploration given features of a scientific community and division between roles. Aronowitz (2021) argues in the individual case for randomness in proportion to features of the epistemic situation, and driven to some extent by exploration of nearby possibilities.

Thus while no full solution to the learning problem has been proposed, in thinking about limited agents we have begun to partially distinguish rational from irrational theory discovery. This fact alone suggests a connection between the accommodation and learning questions.

5. Conclusion

The problem of new theories arises when we attempt to use a Bayesian framework to analyze decision contexts where previously-unnoticed possibilities get put on the table. These cases are common in science as well as everyday life, since limited reasoners almost never operate with a complete understanding of the possibilities. Whether to view this problem as limiting Bayesianism to a proper domain or a true challenge is up for debate, but rests in part on the question of whether the acquisition of pertinent new theories is itself a kind of learning.

Bibliography

Aronowitz, S. (2021). Exploring by Believing. *Philosophical Review*, *130*(3), 339-383.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Carey, S. (2000). The origin of concepts. Oxford University Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Fagin, R., & Halpern, J. Y. (1988, March). Reasoning about Knowledge and Probability. In *TARK* (Vol. 88, pp. 277-293).

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The journal of the learning sciences*, *6*(1), 3-40.

Glymour, C. (2016). Why I am not a Bayesian. In *Readings in Formal Epistemology* (pp. 131-151). Springer, Cham.

Halpern, J. Y., & Rego, L. C. (2006, May). Extensive games with possibly unaware players. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 744-751).

Karni, E., & Vierø, M. L. (2013). " Reverse Bayesianism": A choice-based theory of growing awareness. *American Economic Review*, *103*(7), 2790-2810.

(2015). Probabilistic sophistication and reverse Bayesianism. *Journal of Risk and Uncertainty*, *50*(3), 189-208.

Ozbay, E. Y. (2007, June). Unawareness and strategic announcements in games with uncertainty. In *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge* (pp. 231-238).

Savage, L. J. (1954). Foundations of statistics. Wiley Publications in Statistics.

Schipper, B. (2012). The unawareness bibliography. *URL=< http://www.econ.ucdavis.edu/faculty/schipper/unaw.htm>*

Steele, K., & Stefánsson, H. O. (2021a). *Beyond uncertainty: Reasoning with unknown possibilities*. Cambridge University Press.

(2021b). Belief revision for growing awareness. *Mind*, *130*(520), 1207-1232.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Thoma, J. (2015). The epistemic division of labor revisited. *Philosophy of Science*, *82*(3), 454-472.

Wenmackers, S., & Romeijn, J. W. (2016). New theory about old evidence. *Synthese, 193*(4), 1225-1250.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.