Understanding *Aha!* Moments

Hermish Mehta[*]
Department of Electrical Engineering & Computer Sciences
University of California, Berkeley
Berkeley, CA 94720
hermish@berkeley.edu

Abstract

In this paper, we explore a preliminary computational model that seeks to explain *Aha!* moments. Under a Bayesian framework, we argue that these occur when we are exposed to evidence that significantly alters the probability distribution over the hypothesis space, converging towards one or a select few hypotheses. This sudden change in understanding, spurred by a single Bayesian update or hypothesis space shift, is characteristic of our colloquial interpretation of *Aha!* moments.

1 Introduction

Inspiration still maintains some of its original historical associations, conjuring vivid images of artists or scientists receiving divine guidance to produce sudden and unexpected valuable insight. This interpretation is reflected in the word's etymology, coming from the Latin verb *inspiro*, meaning to excite, inflame or inspire, representing an introduction of something new or revolutionary to a system.

Yet as human beings we experience this routinely, a phenomenon we call *insight*, the smooth process by which an individual goes from a state of not knowing, to solving a problem [9]. Psychology has been fascinated with understanding the nature of such insight—colloquially called *Aha!* moments—for nearly a century [11]. However, more rigorous treatments of this phenomenon have recently begun to appear, including precise hardware-level explanations of insight.

For example, a recent cognitive neuroscience paper argues that similar problems solved with insight versus analytically corresponded to distinct patterns of brain patterns [4]. This builds on previous work that sought to understand *Aha!* moments through its hardware-level manifestation, attempting to link the subjective experience to some objective measure, such as semantic priming [1]. Insight, then, might be characterized in the brain by the culmination of many brain processes operating at different time scales [4].

These explanations offer progress on a mechanistic and hardware-level account of *Aha!* moments, yet, in the terminology of Marr's four levels of analysis, do not completely address the computational and algorithmic problem from which this phenomenon emerges [8]. In particular it leaves a computational-level model to be desired, since understanding how to influence insight could yield better pedagogical practices—van Hiele [12] directly examines the role of insight in mathematics education.

Simply put, then, how exactly does our understanding of the world change as a product of *Aha!* moments, and how could this be leveraged to improve cognition? In this paper, we explore a preliminary computational model with an underlying Bayesian framework, to characterize the cognitive *Aha!* moments we experience while solving problems or learning something new. Under this model, we present two possible mechanisms through which this can occur and test these with human subjects on realistic problems.

[*] Based on work with Rachit Dubey and Mark Ho.

Preprint. Work in progress.

2 Bayesian Cognitive Models

Systematic scientific change progresses quite cautiously and methodically, expanding the realm of scientific knowledge slowly, but predictably, outward. Yet this machinery of science is very different then our own mysterious internal processes which govern belief change and learning, processes which are well-documented to be far from perfect.

While we have an understanding of how scientific change takes place, to model cognitive phenomenon, we must start with some fundamental assumptions about our beliefs. In particular, we assume the same language of mathematics and probability theory that scientists use to formally reason about hypotheses, captures something about our beliefs. Namely, the idea that degrees of belief can be expressed as probabilities between 0 and 1 [10].

Under this assumption, we can construct probabilistic models which provide a computation-level account of cognition, reducing inductive problems to those of probabilistic inference [5]. The powerful assumption is justified by the success of this top-down approach in explaining phenomena from acquisition in children [3] to decision making [6] to visual perception [7].

The core advantage of treating degrees of belief as probabilities is that probabilistic inference can naturally be expressed as an application of Bayes' rule. Given some hypothesis or belief h, we denote its initial degree of belief, $P(h)$, as the *prior*. The problem of inference involves computing a *posterior* probability of h given some observation d, which can be expressed as a conditional probability $P(h|d)$.

Clearly this posterior should depend on our original degree of belief, $P(h)$, and some measure of how consistent h is with d. The latter is called the *likelihood*, which is the probability of observing d given the particular hypothesis h: a conditional probability $P(d|h)$. This qualitative relationship between the prior, likelihood and posterior can be expressed as

$$P(h|d) \propto P(h)P(d|h),$$

which implies higher priors and higher likelihoods should encourage a higher posterior probabilities. More precisely, to ensure the posterior remains between 0 and 1, we normalize by dividing the right-hand side by the probability we observe d; we can derive the resulting formula using the axioms of probability.

$$\begin{aligned} P(h|d) &= \frac{P(h,d)}{P(d)} \\ &= P(h)\frac{P(d|h)}{P(d)} \end{aligned} \tag{1}$$

Therefore, when give an observation d, we can accordingly update our prior probabilities, accounting for the additional data. If d is inconsistent with the hypothesis h_i, then its respective likelihood will be zero, driving the posterior probability to zero in the computation. This is compatible with intuitive notions of rejecting a hypothesis given contradictory evidence.

Naturally, in building cognitive models, we must not only understand the rules of inference, but also the hypothesis space where these beliefs come from. In a neural network, for example, a characterization of the hypothesis space is rather simple: the universal approximation theorem tells us this is the set of continuous functions on $\mathbb{R}^n$ [2]. Yet, this structure is less clear in humans with limited cognitive resources. To account for these natural limitations, we distinguish between the *latent hypothesis space* and the *explicit hypothesis space*.

The latent spaces refers to the set of all hypotheses capable of being represented, while the explicit space is the set of hypotheses currently, actively being represented and manipulated [10]. Hypothesis generation, then, involves moving hypothesis from the latent to explicit space. For clarity, we use the following notation for each of these:

$$\mathcal{L} \equiv \text{latent hypothesis space}$$
$$\mathcal{H} \equiv \text{explicit hypothesis space}$$

Notice that the space of all possible hypothesis in the latent space can be infinite, juxtaposed against the reality of finite cognitive resources. As a result, only a small subset $\mathcal{H} \subseteq \mathcal{L}$ is actually considered. Overall, however, since the latent space contains an exhaustive, mutually-exclusive list of all hypotheses, assuming a hypothesis capable of being represented is true, we should have

$$\sum_{h \in \mathcal{L}} P(h) = 1 \tag{2}$$

Notice that this property is conserved while updating probabilities with respect to some observation, since the the set of all hypotheses is exhaustive. This can readily be proved by computing the sum of the posterior probabilities of all hypotheses after being condition on an observation d.

$$\begin{aligned}
\sum_{h \in \mathcal{L}} P(h|d) &= \sum_{h \in \mathcal{L}} P(h) \frac{P(d|h)}{P(d)} \\
&= \frac{1}{P(d)} \sum_{h \in \mathcal{L}} P(h)P(d|h) \\
&= \frac{1}{P(d)} P(d) \\
&= 1
\end{aligned}$$

All of this machinery establishes a consistent way to study beliefs over hypothesis spaces, and describe optimal inference, even if working over the latent space is cognitively intractable.

3 Modelling *Aha!* Moments

3.1 Beliefs & Hypothesis Spaces

Reasoning occurs over the explicit hypothesis space, and not the latent space, so when modeling cognition, we need to restrict our attention to only those hypotheses actively being represented by the conceptual system. However, we also have some error estimate, namely a confidence associated with the belief that the true hypothesis lies outside the explicit hypothesis space. Denote the set of all such external hypotheses with epsilon.

$$\varepsilon = \mathcal{L} \setminus \mathcal{H}$$

Now let the hypothesis h_ε represents the belief that the correct hypothesis is not in $\mathcal{H}$. Constructively, this definition reads that exactly one of the hypotheses in ε is true.

$$h_\varepsilon = \bigvee_{h \in \varepsilon} h \tag{3}$$

Of course, by definition, we have no knowledge of the structure or contents of ε while reasoning about our beliefs; otherwise, those hypotheses would lie in the explicit space instead. Note that h_ϵ is not strictly in $\mathcal{H}$ since it does not represent a single hypothesis, but rather is an acknowledgment of the incompleteness of the hypothesis space. Optimally, our degree of belief in h_ϵ should reflect exactly the shortcoming of our explicit hypothesis space.

$$P(h_\varepsilon) = \sum_{h \in \varepsilon} P(h) \tag{4}$$

Notice that $\mathcal{H}$ and ε partition that the latent hypothesis space, since all representable hypotheses lie either within the explicit space or just outside it. Therefore, we once again have a mutually-exclusive, exhaustive set of beliefs, so it follows that the total degree of belief among them must be one.

$$\begin{aligned} P(h_\varepsilon) + \sum_{h \in \mathcal{H}} P(h) &= \sum_{h' \in \varepsilon} P(h') + \sum_{h \in \mathcal{H}} P(h) \\ &= \sum_{h \in \mathcal{L}} P(h) \\ &= 1 \end{aligned} \tag{5}$$

While these probabilities represent the ideal degrees of beliefs of to each of these hypotheses, in reality we have some approximation to this, assigning a probability to every element of

$$\mathcal{H}' = \mathcal{H} \cup \{h_e\}.$$

Any particular explicit hypothesis space, $\mathcal{H}$, induces a particular distribution over the possible beliefs in $\mathcal{H}'$. This distribution fully characterizes our belief over all hypotheses actively being considering. In this setting, learning can occur in one of two ways, either through observation and testing, or by hypothesis generation.

3.2 Ambiguity & Completeness

For a particular distribution $\mathcal{H}'$ based on an explicit hypothesis space $\mathcal{H}$, we can define a measure of ambiguity. First, note that with full information about the distribution of $\mathcal{L}$, the entropy of the distribution is a good metric of overall uncertainty. However, this metric has two important shortcomings:

1. it is reflective of the unpredictability of the latent space, which does not correspond to our more limited knowledge; and
2. it cannot be computed without knowing the distribution of hypotheses in ε.

Rather, we are interested in the entropy of the distribution over hypotheses in our explicit hypothesis space, excluding our error hypothesis. This is the same as computing the entropy of the distribution over $\mathcal{H}$, conditioned on the true hypothesis lying within. To account for the effect of cardinality, we instead compute the normalized conditional Shannon entropy.

$$H_{\text{norm}}(\mathcal{H}|h^* \in \mathcal{H}) = -\frac{1}{\lg|\mathcal{H}|} \sum_{h \in \mathcal{H}} P(h|h^* \in \mathcal{H}) \lg P(h|h^* \in \mathcal{H}) \tag{6}$$

$$= -\frac{1}{\lg|\mathcal{H}|} \sum_{h \in \mathcal{H}} \frac{P(h)}{P(h^* \in \mathcal{H})} \lg \frac{P(h)}{P(h^* \in \mathcal{H})}$$

$$= -\frac{1}{\lg|\mathcal{H}|} \sum_{h \in \mathcal{H}} \frac{P(h)}{1 - P(h_\varepsilon)} \lg \frac{P(h)}{1 - P(h_\varepsilon)} \tag{7}$$

This gives a reasonable explanation of the unpredictability our current set of hypotheses—how clearly we can discriminate between possible explanations given our fixed current set of hypotheses. We can define this quantity as the *ambiguity* of our explicit hypothesis space. However, clearly this metric is meaningless if our absolute degree of belief in these ideas is low. Therefore, we introduce a second metric to characterize our overall confidence that we have the correct explanation, which we denote *completeness*.

$$\text{ambiguity}(\mathcal{H}') = H_{\text{norm}}(\mathcal{H}|h^* \in \mathcal{H}) \tag{8}$$
$$\text{completeness}(\mathcal{H}') = 1 - P(h_\varepsilon) \tag{9}$$

Both measures lie in the interval $[0, 1]$. Under this model, we can characterize an *Aha!* moment as a sudden decrease in ambiguity, given high completeness. Below we examine and detail two possible mechanisms through which this could occur.

3.3 Observation

An *Aha!* moment corresponds to a change in distribution over all hypotheses, so this may occur while updating probabilities given an observation. Remember, given an observation d, we compute posteriors for each hypothesis in $\mathcal{H} \cup \{h_\varepsilon\}$ by applying Bayes' rule, inducing a new distribution over the same hypotheses. This new distribution will have different measures of ambiguity and completeness.

$$\text{ambiguity}(\mathcal{H}') \mapsto \text{ambiguity}(\mathcal{H}'|d)$$
$$\text{completeness}(\mathcal{H}') \mapsto \text{completeness}(\mathcal{H}'|d)$$

Given the resulting the completeness of $\mathcal{H}$ given d is sufficiently high, we are interested in the change in ambiguity as a product of observing d. We can denote this difference the *power* of this observation.

$$\text{power}(d; \mathcal{H}') = \text{ambiguity}(\mathcal{H}'|d) - \text{ambiguity}(\mathcal{H}') \tag{10}$$

Qualitatively, this represents the change in unpredictability over our hypothesis space. More negative values correspond to greater drops in entropy, which imply the observation is better at discriminating between hypotheses in the explicit hypothesis space. To illustrate this, consider a example, where there are a small number of hypotheses which can exhaustively be maintained in the explicit hypothesis space.

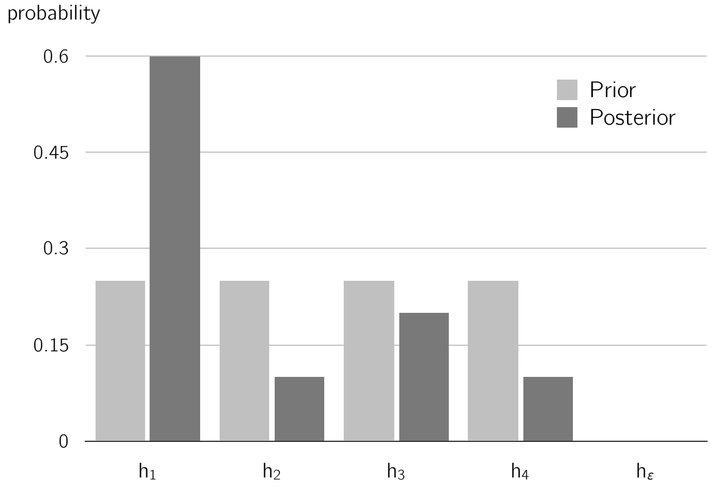

Figure 1. An example of a size four hypothesis space with no error probability, changing with respect to an observation. While the prior distribution is roughly uniform across all hypotheses, the posteriors yield a significantly more uneven distribution.

Notice that the *completeness* of this distribution, which we denote $\mathcal{H}_4$, is one before and after being updated. In this example, the *power* of the observation can directly be computed as the difference in *ambiguities* of the prior and posterior distributions.

$$\begin{aligned}
\text{power}(d; \mathcal{H}'_4) &= \text{ambiguity}(\mathcal{H}'_4|d) - \text{ambiguity}(\mathcal{H}'_4) \\
&= -(0.6 \lg 0.6 + 2 \cdot 0.1 \lg 0.1 + 0.2 \lg 0.2) + (4 \cdot 0.25 \lg 0.25) \\
&\approx -0.129
\end{aligned}$$

The sign of the tells us the observation does reduce the entropy of the conditional hypothesis distribution—its magnitude gives us a relative estimate of how strong this effect is. The more negative this value is, the more likely we would expect it to cause an *Aha!* Moment.

3.4 Hypothesis Generation

Hypothesis generation is the process of moving hypotheses from the latent hypothesis space to the explicit hypothesis space [10]. Clearly, during hypothesis generation, barring noisy approximations, we should expect our error estimate to decrease, since the explicit hypothesis space grows with new beliefs. If hypotheses $S \subseteq \varepsilon$, move to the explicit space, the change in the completeness will be non-negative.

$$\begin{aligned}
\Delta\,\text{completeness}(\mathcal{H}') &= (1 - P(h^* \in \mathcal{H})) - (1 - P(h^* \in \mathcal{H} \cup S)) \\
&= P(h \in \mathcal{H} \cup S) - P(h \in \mathcal{H}) \\
&= \sum_{h \in S} P(h) \\
&\geq 0
\end{aligned}$$

Under our definition, this could result in an *Aha!* Moment if this increases completeness beyond the threshold required, assuming the resulting ambiguity is low. More interesting, however, is examining the change in ambiguity as a product of hypothesis generation. Although the probabilities of hypotheses do not change, their relative values can be affected.

If a hypothesis with high probability is introduced into the explicit hypothesis space, for example, the conditional probabilities of all current hypothesis would markedly decrease, which could cause a sharp decrease in ambiguity, leading to an *Aha!* Moment.

$$\text{ambiguity}(\mathcal{H}' \leftarrow S) = -\frac{1}{\lg(|\mathcal{H} \cup S|)} \sum_{h \in \mathcal{H} \cup S} \frac{P(h)}{P(h^* \in \mathcal{H} \cup S)} \lg \frac{P(h)}{P(h^* \in \mathcal{H} \cup S)} \qquad (11)$$
$$\text{where } P(h^* \in \mathcal{H} \cup S) \geq P(h^* \in \mathcal{H})$$

This is because we normalize the distribution over hypotheses with respect to the completeness of the hypothesis space, so hypothesis generation has the effect of potentially reducing the conditional probabilities associated with already present hypotheses. Once again, if the ambiguity is sufficiently reduced, our model predicts a resulting *Aha!* Moment.

4 Future Work

Future work on this project could involve refining the model further, and testing its predictions and consequences on real-world settings.

One specific application of this model could be in pedagogical setting. In particular, this model seems to suggest that the nature of the problem, the hypothesis space, and the prior belief distribution are all responsible for producing an this kind of realization, or *Aha!* Moment. This is relevant to the problem of picking appropriate examples or problems to encourage learning; namely, it predicts that examples chosen strategically to differentiate between current hypotheses should be more effective in reducing uncertainty.

Experimental evidence for (or against) this model could come from simple games or puzzles, where the hypotheses space is either explicitly defined or restricted. Here, optimal inference can be directly modeled, along with completeness and ambiguity of current knowledge. Directly collecting information about participants distribution of beliefs across all hypotheses could also be used to test whether *Aha!* Moments do indeed tend to correlate with drops in entropy.

Acknowledgments

This work was based on an ongoing research project with Rachit Dubey and Mark Ho at the University of California, Berkeley.

References

[1] Bowden, E. M. and M. Jung-Beeman (2003). Aha! insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin & Review 10*(3), 730–737.

[2] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems 2*(4), 303–314.

[3] Gopnik, A. and H. M. Wellman (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin 138*(6), 1085.

[4] Kounios, J. and M. Beeman (2009). The aha! moment. *Current Directions in Psychological Science 18*(4), 210–216.

[5] L Griffiths, T., N. Chater, C. Kemp, A. Perfors, and J. B Tenenbaum (2010, 08). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences 14*, 357–364.

[6] Lee, M. D., I. G. Fuss, and D. J. Navarro (2007). A bayesian approach to diffusion models of decision-making and response time. In *Advances in neural information processing systems*, pp. 809–816.

[7] Lee, T. S. and D. Mumford (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A 20*(7), 1434–1448.

[8] Marr, D. and T. Poggio (1976). From understanding computation to understanding neural circuitry. Technical report, Cambridge, MA.

[9] Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition* (2 ed.). New York, NY: W. H. Freeman.

[10] Perfors, A., J. B. Tenenbaum, T. L. Griffiths, and F. Xu (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition 120*(3), 302–321.

[11] Sternberg, R. and J. Davidson (2002). *The Nature of Insight*. Cambridge, MA: MIT Press.

[12] Van Hiele, P. M. (1986). *Structure and insight : a theory of mathematics education*. Orlando, FL: Academic Press.