Improving the Efficiency of Research and Development Using Belief Networks

by

Keith A. Yost

Submitted to the Department of Nuclear Science and Engineering in partial fulfillment of the requirements for the degrees of Bachelor of Science in Nuclear Science and Engineering

and

Master of Science in Nuclear Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2009

Author ....
Department of Nuclear Science and Engineering
August 7th, 2009

Certified by ....
Michael W. Golay
Professor of Nuclear Science and Engineering
Thesis Supervisor

Certified by ....
Ronald G. Ballinger
Professor of Nuclear Science and Materials Science Engineering
Thesis Reader

Accepted by ....
Jacquelyn C. Yanch
Professor of Nuclear Science and Engineering
Chair, Department Committee on Graduate Students

Improving the Efficiency of Research and Development Using Belief Networks

by

Keith A. Yost

Submitted to the Department of Nuclear Science and Engineering
on August 7th, 2009, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science in Nuclear Science and Engineering
and
Master of Science in Nuclear Engineering

Abstract

Within the past thirty years, the U.S. government has spent over three trillion dollars supporting research and development projects across its various federal agencies. There is a considerable, long-standing need to monitor, justify, and improve this allocation of taxpayer monies. However, oversight and prioritization of research funding has been haphazard, largely because the agencies that administer research funding lack appropriate metrics to measure project success.

We investigate the use of Bayesian belief networks as a means of tracking the success of research and development projects and prioritizing research funding across different experimental efforts. The focus of the thesis is on demonstrating a proof of concept of Bayesian networks by applying the methodology to an alloy research project led by Dr. Ronald Ballinger at the Massachusetts Institute of Technology. We determine the main parameters of interest in the project, establish a network of conditional probability to relate experimental results to alloy viability, perform a Bayesian updating of the research success probability using experimental results, and examine how the optimal choice of experimental design changes as new information is obtained.

We find that belief networks are an appropriate tool for tracking and improving upon the efficiency of research and development. Some potential hurdles are discussed: researcher overconfidence, computational limits of Monte Carlo assessment, and principal-agent games. We reach the conclusion that belief networks are applicable to research and development projects and that their use should be endorsed by the Office of Management and Budget as a means of improving accountability among research-intensive federal agencies.

Acknowledgments

There are four people I must thank for helping me with this thesis.

The first is my advisor, Professor Mike Golay. Besides securing me an internship, setting up my funding, and teaching two of the best classes I've taken at MIT, Professor Golay has been a patient and supportive advisor for the entirety of my thesis work. There is simply no way I could have done this without him.

Next are Professor Ron Ballinger and his graduate student Mike Short. It was very generous on their part to allow me the access that I've had to their project; both have spent several long hours with me, explaining their experiments and sharing the resources that form the core of this thesis. Their helpfulness, candor, and availability were essential to my success.

Lastly, I would like to thank Professor George Apostolakis. Two and a half years ago, in the fall of 2006, I took risk assessment class from him wherein he taught the Bayesian statistics that form the backbone of this thesis. Previously I'd taken frequentist statistics as part of an economics curriculum, but linear regressions and that sort of stuff struck me (and continue to strike me) as a necessary-but-evil form of pseudo-science, one of those types of things you do with data when you can't do *real* science with it. Bayesian statistics, despite its similar roots, feels much more like science, like it's the kind of thing a Karl Popper would approve of. I didn't realize it at the time, but understanding Bayes' Theorem is a life-changing experience— I can't count the number of problems I've encountered that have a Bayesian solution everyone seems to be ignoring. Perhaps it's a curse, perhaps everywhere I go I'll have the misfortune of seeing people trying to fit square pegs into round holes and I'll always be that damned fool who can't stop preaching the gospel of round pegs. Yet... Robert Heinlein, a favorite author of mine, once quipped that it's the knowledge of mathematics that makes a man truly human. If that's the case, then maybe Bayes' Theorem is something akin to the apple in the Eden, and I for one will gladly bear Bayes' burden as the price of my humanity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bayes' Theorem is a simple, straightforward statement of conditional probability. In basic terms, it is a formula that describes a state of knowledge and how that state of knowledge changes when introduced to new evidence. Research and development (R&D) is, at its core, a Bayesian process— faced with uncertainty, scientists gather evidence from experiments and investigations in order to improve their understanding of a system. It is in this manner that a state of knowledge is improved upon; every experiment conducted is, at its most fundamental level, an empirical measurement of some underlying reality, and each measurement of this reality brings a state of knowledge closer to the truth.

If we accept that the purpose of scientific research is to produce knowledge that has utility to mankind, then we are able to show preference between one state of knowledge and another based upon the utility that those states provide. It is then also possible to use Bayes' Theorem to determine the benefit of a given scientific effort. By forming a probabilistic expectation of the effects that an experiment will have on our state of knowledge, it is possible to quantify the benefits of research.

For decades, the U.S. Federal Government has sponsored and directed R&D projects in an effort to capture the benefits of improved knowledge. Federal research is used to inform regulation, improve military capabilities, and develop products of public importance. While the government does not always explicitly quantify the benefits of these public programs, it clearly demonstrates a capacity to prioritize objectives and make trade-offs with scarce resources.

What is lacking in the federal decision making framework is a translation of priorities into efficient funding allocations. It is not enough to assign a value to a new fighter jet, safety regulation, or other research result— until the *probability* of achieving that result is estimated, no true cost-benefit analysis can be performed. Often, the consequence of this disconnect has been the mismanagement of government research: many low-return projects continue to obtain funding long after unfavorable results have been received, leaving more promising endeavors unfunded.

Past efforts by government to improve research and development efficiency have struggled, primarily because the inherent nature of scientific discovery does not lend itself easily to the standard set of measurement techniques used elsewhere. Despite the Office of Management and Budget's (OMB) insistence the efficiency of research should be measured using "outcome-based" metrics, federal agencies have come to rely upon an inadequate set of metrics to define their programmatic efficiency. "Administrative costs as a percentage of total program costs" or "average time to process research grant proposals" are statistics that surely capture *some* aspect of efficiency, but clearly do not fulfill the desire for outcome-based metrics as intended by the OMB. What we are primarily interested in is not the *speed* with which a grant reviewer completes his job, but moreso the *accuracy* with which he completes it— in other words, whether he is identifying the most promising lines of research based upon the priorities that the agency has set.

This thesis examines the possibility of applying Bayesian belief networks— networks of nodes that depict a current state of knowledge and beliefs— to the problem of measuring and improving research and development efficiency. Special attention is given to publicly-led R&D efforts and the difficulty agencies have had in evaluating non-repeated or unique research, multi-year projects, "basic" research, and applied research that has lagging or difficult to measure public benefits. To demonstrate the viability of the Bayesian approach, this thesis works through a prototypical case example using an ongoing publicly-funded research project at the Massachusetts Institute of Technology. The results from this prototypical example are used to create usable metrics, identify possible hurdles to the Bayesian approach and develop solutions to identified hurdles.

The remainder of this introduction details the scope and nature of public research in the United States and makes definitions that will be useful for later analysis.

Chapter two describes the current practices in government research oversight among major U.S. agencies, and details the limitations of these practices.

Chapter three describes Bayes' Theorem and its theoretical application to research and development oversight.

Chapter four describes the research project that was selected as a prototypical case study and gives a brief overview of the scientific concepts necessary to understand the project.

Chapter five describes the analysis of the case study, detailing each step of the process, including defining the success space, determining variables of interest, gathering prior distributions, and conducting the updating of priors using obtained evidence.

Chapter six describes the results of the case study analysis.

Chapter seven discusses the results of the case study, their implication for the use of Bayesian networks as a way of measuring research efficiency, and summarizes the major conclusions of the thesis.

1.1 The Scale of Research and Development in the United States

From 1976 to 2009, the U.S. Federal Government appropriated \$3.46 trillion (2008 dollars) to research and development efforts[11]. These funds amount to more than 1% of U.S. gross domestic product (GDP) during that time period[12], each year accounts for a quarter to a half of all national spending on R&D[8], and internationally, constitutes 10% of all world R&D expenditures[6, 17]. Government-led R&D programs are tasked with making advances in defense, health, space, energy, and other fields of national importance.

In fiscal year 2009, the budget authority for research and development funding, including stimulus spending, was set by Congress at $172 billion. Five agencies accounted for nearly 95% of this funding: the Department of Defense (DoD)— 48.1%, the Department of Health and Human Services (primarily through the National Institutes of Health, NIH)— 24.4%, the Department of Energy (DoE)— 9.5%, the National Aeronautics and Space Administration (NASA)— 8.0%, and the National Science Foundation (NSF)— 4.3%[11].

1.2 The Nature of Public R&D

Publicly funded research and development takes a variety of forms and is directed by a broad range of bodies with diverse and often overlapping programmatic goals. Because Bayesian decision analysis depends on an accurate depiction of current and possible states of knowledge, its application to research and development will vary. In the context of applying Bayesian decision analysis to research and development projects, we conceive of three important features of research that will determine the potential of belief networks to improve outcomes.

1.2.1 Basic vs. Applied Research

One of the most common distinctions made between research programs is that of basic and applied. The OMB defines applied research as "systematic study to gain knowledge or understanding necessary to determine the means by which a *recognized and specific need may be met*[23]" (emphasis added). In other words, applied research is motivated by identifiable needs and is targeted towards the creation of new knowledge that directly improves an agency's capacity to meet those needs. Similar to applied research are development projects, which are more conservative in their generation of new knowledge but are directed towards highly specified ends, such as the creation of a new product or process.

By contrast, the OMB defines basic research as "systematic study directed toward fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts *without specific applications towards processes or products in mind*[23]" (emphasis added). Unlike applied research, basic research is not motivated by unsatisfied needs, but instead by discrepancies or gaps in existing, fundamental knowledge. It is not needs-based, but instead exploratory— as such, it is unclear a priori, whether the new knowledge generated by basic research will fruitfully inform subsequent experiments or provide solutions to immediate needs.

There is also a significant gray area between the realms of basic and applied research. Some research may be theoretical, yet still clearly be motivated by known, specific needs. For example, research into plasma physics and superconductors is primarily motivated by the desire to develop fusion power and improve the efficiency of electrical conductors.

In terms of applying Bayesian decision making to R&D, the most important distinction between basic and applied research is the degree to which the value of different potential research outcomes can be known prior to receiving the results. One example of research that meets almost every pre-requisite for successful Bayesian decision making EXCEPT for the ability to assign value to research outcomes is the Large Hadron Collider (LHC), a 27km long particle accelerator designed to run experiments that will improve our understanding of particle physics. While the LHC provides quantifiable data (through a massive array of detectors and sensors), will explore questions whose answer sets can be fully characterized (the Higgs Boson either exists or does not), and can report prior beliefs on the answer set (scientists at a rival collider, the Tevatron, recently projected the chances of their facility gathering enough evidence to confirm the existence of the Higgs Boson as a function of the particle's mass and the length of facility operating period, see Figure D in the appendices), it is effectively impossible to estimate the societal value of one research result versus another, and thus there is no rigorous way to form an expectation of the benefits of the project and compare it to other potential research efforts.

The funding allocation between basic research, applied research, and development varies from one government body to another. In the DoD for example, more than 92% of all R&D funds are spent on development efforts, while only 2% goes towards basic research[9]. By contrast, outside of the DoD, most R&D funding is split evenly between basic and applied research, with only a small fraction going towards development projects. In total, recent history has seen 60% of public R&D funding gone towards development, with the remaining 40% split evenly between basic and applied research[10].

Importantly for Bayesian applications, most public R&D projects are organized around clear missions with well-defined programmatic needs. The NIH, which supports over half of all public basic research, targets its funding toward specific needs: cancer, allergies, infectious disease, cardiopulmonary health, and other medical areas in which improvements can lead to easily estimable benefits in terms of lives saved. As such, Bayesian decision analysis can be applied to the large majority of federal R&D.

1.2.2 "Hard" vs. "Soft" Science

Another important distinction is between so-called hard and soft science. The term "hard science" is used to describe research that provides objective, quantifiable data through experiment. "Soft science," by contrast, yields qualitative data that requires subjective interpretation. Soft science does not lend itself easily to a Bayesian treatment because unlike quantified data, which must take some value on a number line, qualitative data may take unpredictable forms. If the possible results of an experiment cannot be characterized prior to the experiment, then not all nodes in the Bayesian network can be depicted accurately. Qualitative data can always be organized into categories that are mutually exclusive and collectively exhaustive through the use of a lexicographic set (such as "high," "medium," "low") that (as a last resort) includes "Other" as an answer. However, to the extent that lexicographic sets lump together dissimilar data, Bayesian networks are unlikely to be useful— by and large, belief networks do not adapt well to qualitative or subjective inputs.

Furthermore, hard science has another advantage over soft science: as part of sound experimental design, researchers in hard science are likely to have already given thought to the implications of different potential research results. In medical trials for example, considerable forethought is given to the sample population sizes to be tested in order to yield statistically significant results.

1.2.3 Intramural vs. Extramural Research

Finally, there are considerable differences in funding mechanisms across agencies. For the NIH, typically 90% of the research budget is spent on extramural research, that is, research conducted by outside institutions through grants and contracts, and consistently more than half of the budget is spent on research grant awards[7]. Other agencies rely more on intramural research to satisfy their mission goals, delegating research tasks to in-house laboratories and facilities. There is also significant variation between agencies in how much of their extramural research budget goes towards projects that were solicited by the agency and how much goes towards research topics that were unsolicited by the agency but instead were suggested by outside individuals and institutions.

Extramural research presents an interesting challenge to the application of belief networks. The decision-making strength of a belief network relies strongly on the honesty with which initial beliefs and conditional probability relationships are put into the network. If an experimenter lies about the expected accuracy of an experiment or submits overly optimistic beliefs on the probability of obtaining favorable results, the value of his experiment will be exaggerated.

This creates what is commonly referred to as an "agency problem[16]." If the interests of the principal (the government / funding body) do not align with those of its agent (the researcher who conducts work on the principal's behalf), there is a motivation for the agent to "game" the system in a way that runs counter to the principal's interest. To the extent that extramural input is used to inform decision-making, funding bodies will need to set up an incentive structure that rewards honesty and punishes misrepresentation.

1.3 Chapter Summary

Public research and development is a major enterprise within the United States, with many federal agencies spending billions and tens of billions of dollars annually on research and development[11]. Scientific inquiry is crucial to achieving missions of great public importance, and federal agencies have struggled to ensure that research monies are spent in the most effective way possible. Bayesian decision analysis holds promise in improving the results of research and better allocating the public's investment in new knowledge and products. There are some limitations to the Bayesian method, and as such, the method shows the greatest promise when research motives are clearly defined, the range of potential results can be characterized prior to conducting the research, and the decision-making inputs are solicited from honest, informed experts. This thesis examines an ongoing publicly funded research project and uses it as a prototypical case study to demonstrate the use of Bayesian methods.

Chapter 2

Past and Current Efforts to Improve R&D Efficiency

2.1 Defining Efficiency

It is worthwhile to ask what exactly we mean when we say that we would like to improve the "efficiency" of research and development.

Conceptually, there are two different kinds of efficiency: productive efficiency and allocative efficiency. Productive efficiency, or "process efficiency" is concerned with how much research can be performed for a given set of resources. If two researchers can perform the same experiment to the same standard, but one can perform it with fewer resources than the other, we would say that the researcher who can complete the task with fewer resources is more efficient. Allocative efficiency, or "portfolio efficiency" is concerned with how well funding is allocated across different research options. It is primarily concerned with how well research funding is invested, whether it is directed to projects that offer the greatest societal return per unit of resources invested.

Efficiency is most meaningfully defined as the ratio of benefits to costs[18]. What makes this definition difficult to apply in the context of R&D is that the ultimate outcomes of research are both probabilistic and difficult to objectively measure. The purpose of tracking efficiency is to improve it— even if the ultimate outcomes of research could be objectively and promptly measured, the element of chance would obscure the relationship between measured "efficiency" and true allocative and productive efficiency. Changing the definition of efficiency to the *expected* ratio of benefits to costs solves the problem but it raises a new concern: "How do agencies form their expectations of results?" Under such a system, an agency would be asked to form two expectations: an expecation on the probability of a project's success (overcoming the problem of probabilistic results), and an expectation on the ultimate value of that success (overcoming the problem of unmeasurable ultimate outcomes).

Thus, what is needed is not just a new metric for efficiency, but also (as this new metric will be one level removed from the meaningful definition of efficiency) a method of confirming the accuracy and relevance of this new metric. It is of added benefit if the process of measurement highlights areas for improvement or suggests a course of action.

2.2 The Government Performance and Results Act

For decades, the U.S. Federal Government has recognized the importance of performance evaluation as a means of ensuring the judicious use of taxpayer monies. Early efforts were focused on simple forms of accountability, such as identifying project goals and improving transparency in budgeting[19].

As time went on, these efforts became more comprehensive and attempted to more closely track the relationship between program resources and outcomes. The current system of oversight has its roots in the 1990's when, in response to perceived waste and inefficiency in publicly-funded programs, Congress and the executive branch initiated a reform of the statutory requirements and management practices of federal agencies. The linchpin of this reform effort came in 1993, when Congress passed the Government Performance and Results Act (GRPA). This act requires that all government agencies develop strategic plans for program activities, establish performance goals, and submit to Congress an annual report on the progress made by each program activity in achieving its set goals[3].

2.3 The Program Assessment Rating Tool

In 2002, the Office of Management and Budget (OMB) developed the Program Assessment Rating Tool (PART), a diagnostic tool designed to improve compliance with GRPA requirements. In its first year, the OMB used PART to assess 234 programs covering roughly 20% of the federal budget, and each year following, the OMB expanded evaluations to an additional 20% of budget coverage. As part of compliance with PART, each program explicitly identifies performance metrics and uses them to justify public funding[19].

PART is a questionnaire intended to ensure consistency of performance evaluation across programs and assist agencies in identifying ways to improve program performance. It is made up of 25 questions divided into four sections, with each section weighted to form an overall score (a summary of PART sections is given in Table 2.1). Each question is, by default, given equal weighting within its section, although the weighting can be altered on a case by case basis. Additionally, research and development programs are given three extra questions, two in the Strategic Planning Section and one in the Program Management section.

Table 2.1: PART Question Sections

Section	Weight	Questions	Answer Types
Program Purpose and Design	20%	5	Yes/No
Strategic Planning	10%	8	Yes/No
Program Management	20%	7	Yes/No
Program Results/Accountability	50%	5	Yes/Large Extent/Small Extent/No

The full list of questions can be found in Appendix A. If the OMB judges that a program has not developed acceptable performance metrics or does not have sufficient data upon which to make a judgement, it is given a "Results Not Demonstrated" rating. Otherwise, the OMB converts the numeric score that results from the PART questionaire into a semantic rating of "effective," "moderately effective," "adequate," or "ineffective."

Table 2.2: PART Question Scoring

Rating	Score Range
Effective	85–100
Moderately Effective	70–84
Adequate	50–69
Ineffective	0-49

For a public program, there are high stakes attached to the results of the PART questionaire. Programs that receive a rating of "Ineffective" are often targeted for cuts by Congress. Over time, there has been a general improvement in PART rating assessments— it is unclear whether this effect is due to actual improvement among programs, bureaucratic adaptation to OMB oversight, or is a statistical artifact of the different program populations that each year samples (the 2002 assessment only looks at 20% of applicable programs whereas later years look at nearly all federal programs)[21].

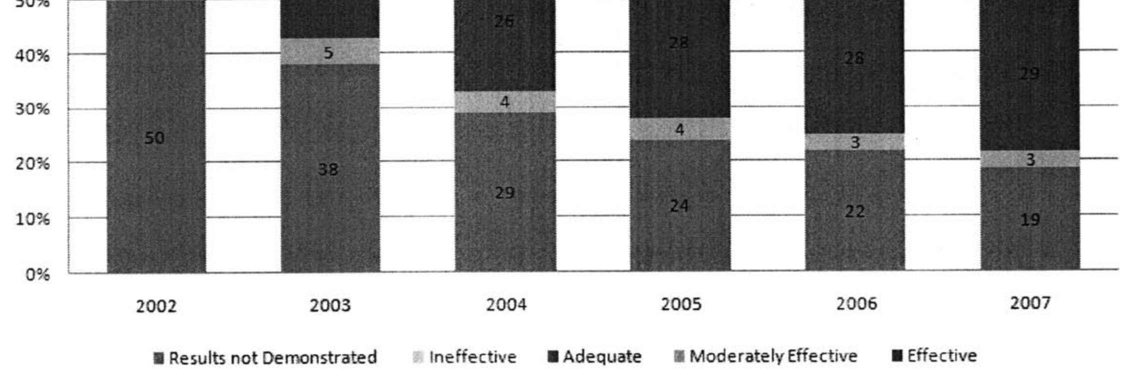


Figure 2-1: PART Assessment Ratings, 2002-2007

The GRPA applies to all public programs, but research programs in particular have found it challenging to comply with GRPA/PART requirements to measure efficiency. Unlike other public efforts, where there are clear benchmarks and metrics available with which to fulfill the PART requirements, research is a difficult field to measure for three reasons:

1. Research consists largely of unique, one-off efforts that defy easy standardization or comparison.
2. Research outcomes are probabilistic, not deterministic.
3. The benefits of research are often delayed or dependent on follow-up actions.

In 2007, to help federal agencies complete PART assessments for research projects, the OMB released Research and Development Program Investment Criteria[20]. In their guidance, the OMB emphasized three criteria: Relevance, Quality, and Performance, reflecting the relevance of research investments to national priorities, a defensible method of quality assessment, and performance assurance through the development of clear objectives, metrics, and identifiable results. The full guidance memorandum is included in Appendix C

Research-intensive agencies have expressed frustration with PART requirements, particularly question 3.4, "Does the program have procedures to measure and achieve efficiencies and cost effectiveness in program execution?" There is a widely held perception that the OMB does not apply the metric requirement evenly across all projects, that there is considerable variation between OMB reviewers as to what types of metrics get accepted or rejected. Indeed, nearly identical metrics have been accepted for some research projects and rejected for others, without clear reason as to what differentiates the two[24].

2.4 Current Efficiency Metrics

The OMB has strongly emphasized that efficiency metrics used by agencies in fulfillment of their PART requirement should be "outcome-based." As an example of what is meant by "outcome," the OMB gives the example of a reduction of HIV infections—reducing the transmission of a disease is a clear, quantifiable public benefit[20]. What the OMB intends by emphasizing outcome-based metrics is to create a linkage between resources invested in R&D and societal benefits returned.

Unfortunately, it is impractical to measure the ultimate outcomes of research projects using any objective system of accountancy[3]. The outcomes generated by research and development are not easily attributable to individual projects or efforts for a variety of reasons: the effects of research are difficult to isolate from the effects of other factors, research does not directly create a usable good for society, the realization of research outcomes often lags the research itself, research outcomes are probabilistic and metrics based upon them will not distinguish between "lucky" results and those created by well-designed projects, and often, the mechanism by which research improves societal welfare (especially in the case of basic research) is the enabling of subsequent research efforts.

Unable to develop "outcome"-based metrics, federal research programs have settled for a series of substitutes which could, at best, be characterized as "output"-based metrics. "Outputs," as defined by the OMB, are the internal, intermediate results of a program. They are products that are not of direct benefit to the public, but are correlated in some way with ultimate benefits. Rather than attempt to track the final benefits of research, current metrics instead track the intermediate results of research. As a proxy measure for ultimate outcomes, these metrics are poor: the relationship between the outputs tracked by an agency and outcomes is weak, and in most cases the use of such metrics may be counter-productive[3]. A common saying in productivity tracking is "You get what you measure." The consequence of attaching high stakes to outputs that are only tangentially related to ultimate success is that those outputs will be increased even when it is detrimental to the agency's mission[13].

The metrics most commonly used by the largest five research agencies include[21]:

- Average cost per standard measurement or analysis.

- Time, in days, between receipt of grant proposal and award.

- Administrative costs as a percentage of total costs.

- Peer-reviewed publications per full-time employee (or per dollar).

- Percentage of projects that are peer or merit-reviewed.

- Variance from schedule and cost.

The most meaningful definition of efficiency is the ratio of benefits to costs. None of the above metrics provide a reliable measurement of benefits and costs except in rare cases. Three of the metrics only address one half of the cost-benefit ratio (administrative costs as a percentage of total costs, percentage of projects that are peer or merit-reviewed, and variance from schedule and cost). The remaining three metrics make unreliable assumptions about how output or cost correlates with the chosen tracked data. Cost per measurement is a sufficient metric if and only if each measurement is conducted with similar quality standards and provides roughly constant expectation of benefits to society— most research efforts are sufficiently unique to defy such easy collective treatment[3]. Time in days between receipt and award of a grant proposal assumes both that time is a good proxy for total cost (it is unclear why this should be the case), and that the award of a grant proposal returns some constant benefit (highly dubious). The last of the three, peer-reviewed publications per dollar, may be an appropriate metric if the primary activity of the program is the production of publications and peer review is sufficient to guarantee a particular level of social benefit per publication. However, one simple reason to suspect that not all publications have the same social value is that publication rates vary by discipline[13]— an interdisciplinary agency that tracked efficiency with this metric might find itself becoming more "efficient" simply by shifting emphasis between disciplines.

The shortfalls of these efficiency metrics give strong reason to believe that among research projects, programmatic goals are being poorly served (if not actively harmed) by existing regulatory requirements. Until new metrics can be provided, there is little reason to believe that either PART or the GRPA are improving research outcomes.

2.5 Chapter Summary

Outcome-based metrics have been long sought-after by research intensive federal agencies, as well as the Office of Management and Budget, as a means of justifying and improving the efficiency of public research funding. Current metrics and other quantitative tools of oversight are insufficient: they do not capture the most important measure of research efficiency, the ratio of a project's benefits to costs, and may, if relied upon too strongly, have contrary effects on agency performance. Bayesian belief networks may offer a method of developing meaningful, outcome-based metrics of efficiency, leading to improved research decision-making and a stronger capability by agencies to carry out their mission.

Chapter 3

Bayes' Theorem and Its Application to Research Projects

3.1 Bayes' Theorem

Bayes theorem is a simple, yet powerful statement of conditional probability[1]

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{B})} \tag{3.1}$$

where A and B represent hypotheses, evidence, latent variables, or unknown parameters. Pr(A) is called the prior probability, or, more simply, the prior. It represents a current belief, prior to new evidence, concerning the probability of a hypothesis, A, being true. The prior can be informed by past evidence and measurements, or it could be the purely subjective assessment of an expert. Pr(A|B) is called the posterior probability, or, more simply, the posterior. It represents a new belief, updated by evidence, concerning the probability of a hypothesis, A. $\frac{\Pr(\mathrm{B}|\mathrm{A})}{\Pr(\mathrm{B})}$ is the probability of observing B if A were true, divided by the probability of observing B.

Most commonly, Bayes' Theorem is used to describe a hypothesis, A, and evidence that relates to that hypothesis, B. For example, A might be the event that it has rained earlier today, and B might be the observation that nearby grass is wet.

[1] A quick derivation of Bayes' Theorem is included in Appendix B

More interestingly, A might be the event/hypothesis that a parameter is equal to a given value, and B is an experimental measurement of that parameter. In that case, Bayes' Theorem can be used to derive Equation 3.2, which describes how a continuous probabilistic distribution would change in response to evidence

$$f_X(x|Y=y) = \frac{f_Y(y|X=x)\, f_X(x)}{f_Y(y)} \tag{3.2}$$

where the terms parallel those of the discrete form of Bayes' Theorem, with $f_Y(y|X=x)$ called a likelihood function.

This process of using evidence to change a belief in a hypothesis is referred to as Bayesian inference or updating. One unique property of the process is that it is path-independent; if two relevant observations are made, B and C, it does not matter in which order they are used to update A, the posterior will be the same regardless.

Bayesian updating can also be used to define a new class of probability distribution functions whose prior and posterior are of the same family of functions when updated by a given likelihood function. When this happens, the prior is said to be "conjugate" to the likelihood function, and the prior function itself is called a "conjugate prior." When updating a conjugate prior, Bayes' Theorem is mathematically simple to apply— the parameters that define the prior distribution (called hyperparameters to distinguish them from the underlying parameter that the prior distribution tracks), are modified by the evidence in a simple, formulaic fashion.

It is also possible to link together several random variables to form a Bayesian network or "belief network." This network will usually have a small set of parent nodes, which represent the evidential results of experiments, and a larger set of child nodes, which represent the parameters or hypotheses of interest in an experiment. Relationships of conditional probabilities link the nodes together. The central idea of this thesis is to make experimental "success" (defined either as a simple hypothesis or a variable whose value is dependent on the parameters represented by other nodes) one of the defined nodes in a Bayesian network, so that predictions can be made on the likelihood or degree of success, and these predictions can be updated by new evidence.

3.1.1 An Example of Bayes' Theorem

Paul is given a coin which he knows to be biased such that a coin flip returns the biased side with $\frac{3}{4}$ probability. While Paul knows the coin is biased, he does not know which side the coin is biased towards. Not having seen evidence suggesting either heads or tails, nor having any subjective reason to believe one bias more likely than the other, Paul begins with as naive a prior as possible: he assigns equal probability to each possibility.

Seeking to discover whether the coin is heads-biased or tails-biased, Paul flips the coin and obtains a tails result. Using Bayes' Theorem, he would like to know the probability that the coin is heads-biased.

If we define A as the hypothesis that the coin is heads biased, and B as the observation of a tails result from a coin flip, then the prior, Pr(A), is equal to $\frac{1}{2}$. Pr(B), is equal to the probability that the coin is heads-biased multiplied by the probability of receiving a tails with a heads-biased coin, plus the probability that the coin is tails biased multiplied by the probability of receiving a tails with a tails-biased coin; $\frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{3}{4} = \frac{1}{2}$. And finally, Pr(B|A) is the probability of obtaining a tails with a heads-biased coin, $\frac{1}{4}$. The probability that the coin is heads biased, given a tails flip, Pr(A|B), is equal to $\frac{\frac{1}{4}*\frac{1}{2}}{\frac{1}{2}} = \frac{1}{4}$.

Paul then decides to flip the coin again and receives another tails result. This time, instead of using a naive prior, he uses the latest posterior as his prior: $\Pr(A) = \frac{1}{4}$. Pr(B), again, is equal to the probability that the coin is heads-biased multiplied by the probability of receiving a tails with a heads-biased coin, plus the probability that the coin is tails biased multiplied by the probability of receiving a tails with a tails-biased coin; $\frac{1}{4} * \frac{1}{4} + \frac{3}{4} * \frac{3}{4} = \frac{5}{8}$. And finally, Pr(B|A) is the probability of obtaining a tails with a heads-biased coin, $\frac{1}{4}$. After the second tails result, the new posterior, Pr(A|B), is equal to $\frac{\frac{1}{4}*\frac{1}{4}}{\frac{5}{8}} = \frac{1}{10}$.

3.2 Bayes' Theorem's Relation to Research

The relevance of Bayesian updating to research and development is clear if the above example is reframed. Suppose that Paul is not flipping coins, but instead studying alloys for use in a new product. He is looking for an alloy that meets an array of measureable material requirements, such as tensile strength, resistance to corrosion, thermal conductivity, and so on. Paul has a candidate alloy that he believes has a 50% chance of meeting the property requirements. He also has a scientific test that he can run to determine whether the alloy meets these requirements. If the alloy truly is viable, the test will return a positive result 75% of the time, but if it's an inviable alloy, the test will return a positive result with only 25% probability.

The mathematics of the alloy example are identical to those of the previous coin flip example, but instead of the posterior tracking the probability that a coin is heads biased, it tracks the probability that a given alloy meets a set of physical requirements.

The most important use of this approach is to track the top-level probability of success in an R&D program— in other words, let Pr(A) be the probability of success and let B be the results of project experiments. This will enable research funding agencies to make midcourse corrections and actively manage their portfolio of research investments. For example, suppose that Paul has two candidate alloys that he can perform tests on. His prior beliefs are that each alloy has a 50% chance of being viable. Paul's cost of testing an alloy is $1 for the first test, $0.60 for the second test, $0.36 for the third test, and so on, such that the cost of each test is a geometric series, with the cost of determining the viability of the alloy with absolute certainty (the cost of conducting infinite tests) equal to $2.50.

Suppose, as a simplifying assumption, that Paul assigns value to only one state of knowledge: knowing that an alloy is viable with absolute certainty. He assigns no inherent value to any other state of knowledge, including knowing that an alloy is inviable with absolute certainty, or knowing the viability of an alloy without absolute certainty— these states of knowledge are valued only insofar as they help Paul reach the state of knowledge that an alloy is viable with absolute certainty.

Paul, having no initial reason to believe one alloy is more likely to be viable than the other, runs a test on the first alloy (the two alloys will be called Alloy I and Alloy II for convenience) and receives a negative result. He is now given the choice between continuing his testing of Alloy I to completion at a cost of \$1.50, or switching to Alloy II and testing it to completion at a cost of \$2.50.

After the first test, there is only a 25% chance of Alloy I being a viable alloy, and a 50% chance of Alloy II being viable. Because the cost of fully testing Alloy I is \$1.50 and the cost of fully testing Alloy II is \$2.50, Alloy II offers the greater chance of success per dollar. Given only these two choices, Paul decides to switch to Alloy II.

Now suppose that Paul has a less restricting set of choices; his choices are switching to Alloy II and continuing testing of it to completion, or conducting one more test on Alloy I before having to decide whether to continue testing Alloy I to completion or switching to Alloy II. Should he end the testing of Alloy I immediately, or conduct one more test?

If Paul conducts one more test on Alloy I, there is a $\frac{3}{8}$ chance that the result will be positive (there is a $\frac{1}{4}$ chance that the alloy is viable, and a $\frac{3}{4}$ chance that it is inviable, with a $\frac{3}{4}$ and $\frac{1}{4}$ chance of getting a positive in each scenario respectively). If the result is positive, the updated probability on Alloy I's viability will be 50%, and Paul's choices will be between continuing testing on Alloy I with a probability of finding a viable alloy of 50% and cost of \$0.90, or beginning work on Alloy II with a success probability of 50% and a cost of \$2.50. If the result is negative, Paul's choices will be between continuing work on Alloy I with a success probability of 10% and a cost of \$0.90, or switching to Alloy II with a success probability of 50% and a cost of \$2.50.

Thus, if Paul chooses to run one more test on Alloy I, he will have a $\frac{3}{8}$ chance (the probability of a positive result) of paying \$1.50 total (the cost of testing Alloy I to completion), and a $\frac{5}{8}$ chance of paying \$3.10 total (the cost of testing Alloy II to completion plus the cost of having run the additional test on Alloy I), and will have a 50% chance of achieving success in both cases. His total expected cost is therefore

\$2.50. If Paul chooses not to conduct the extra test on Alloy I and instead switch to Alloy II immediately, he would have the same expected cost and success. In fact, for any common ratio less than 0.6 in the geometric series used to describe the succession of costs in alloy testing, the option to conduct one more test on Alloy I will be more attractive, and for a common ratio greater than 0.6, switching to Alloy II will be more attractive.

In this manner, Bayesian updating can be used to make decisions regarding when to stop a low-performing research project and begin another. More generally however, Bayes' Theorem enables researchers to make a broader range of decisions than simply when to end one series of tests and begin another. As long as the cost of each test can be estimated, the range of possible test outcomes can be characterized, the conditional probability on each test result is known, and a value can be assigned to each possible state of knowledge, then it is possible to use Bayes' Theorem to decide what experiment to run next among any number of experiments and perform constrained optimizations, either maximizing research benefits for a given set of resources, or minimizing cost to achieve a given level of research benefits. To do this, one would make expected benefits or expected costs the top-level measure of a Bayesian network and evaluate each possible decision pathway computationally to find the path that maximizes the expected benefit to costs ratio.

Most of the conditions needed to make Bayesian networks work are usually present in a research program. The costs of experiments are routinely estimated by researchers and funding agencies. Except in very basic research, the range of potential outcomes can usually be characterized— experiments are frequently designed to produced quantified evidence that must fall within a given numeric range, and even experiments that look for qualitative evidence can often describe the results within predetermined lexicographic bounds. Knowing the conditional probability on each test result is a pre-requisite of good experimental design— without any knowledge of the uncertainty on an experimental result, it's nearly impossible to know the informative value of an experiment, and typically, rigorous uncertainty analysis is built into the experiment itself. The last requirement, assigning a value to each state of knowledge, is generally done implicitly rather than explicitly and quantitatively. A discussion of the success space is discussed further in Section 3.2.1.

3.2.1 Defining Success

One of the main difficulties cited in evaluating the value of research outcomes is that research discoveries themselves provide no direct benefits— it's only when the new knowledge is applied through the creation of new processes, products, or regulations that ultimate benefits are created in the form of reduced costs, new goods, or saved lives.

This disconnect creates problems for a number of reasons:

1. There are often significant delays between when research is performed and when ultimate outcomes are realized.

2. Research outcomes are frequently used by actors other than the agency that initiated the research.

3. Sometimes the purpose of research is only to inform new research pathways, creating even more disconnect between the research activity and ultimate outcomes.

The inability to *directly* measure success is not an insurmountable barrier. Expert and peer review panels routinely evaluate the importance of research based upon agency priorities and research relevance. So long as it is possible to *estimate* the relative value of outcomes, it is possible to give weight to different regions of success space and perform optimizations.

Ultimately, the value of research outcomes are tied to the actions taken with those outcomes. As an illustrative example, let's examine the following case:

An entrepreneur has rented a booth at a county fair and is deciding what to sell at the booth. The entrepreneur has two options: a lemonade stand or a hot cocoa stand. His action set is mutually exclusive— he cannot sell both types of drinks, only one. Prior to making his decision however, the entrepreneur does not know what the weather will be for the day. For simplicity, let's suppose that the weather is a binary variable with two possible values: hot or cold.

The entrepreneur estimates that if he sets up a lemonade stand and the weather is hot, he will receive a payout of V_1, but if the weather is cold, he will receive a lower payout of V_2. If instead the entrepreneur sets up a hot cocoa stand, he will receive a low payout, V_3 in the event of hot weather, but a higher payout of V_4 in the event of cold weather. Prior to making the decision of what type of stand to set up, the entrepreneur has imperfect knowledge of what the weather will be.

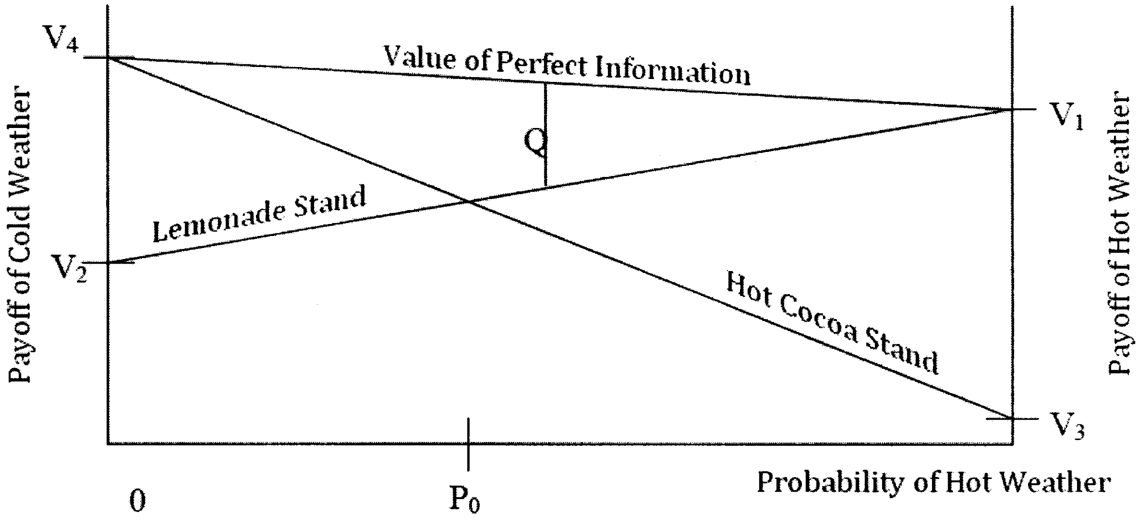


Figure 3-1: The Value of Information

The lines connecting V_1 to V_2 and V_3 to V_4 represent the expected value of setting up a lemonade stand and a hot cocoa stand respectively. When the probability of of hot weather is high, the lemonade stand has a higher expected value than the cocoa stand, and vice versa— when the probability of cold weather (the complement of the probability of hot weather) is high, the cocoa stand has a higher expected payout. At some probability, $P_0 = \frac{V_2+V_4}{V_1-V_2-V_3+V_4}$, the expected value of both actions are equal.

The line between V_1 and V_4 represents the expected value of acting with perfect information. When the weather is known with absolute certainty, as is the case when $P_0 = 0$ and $P_0 = 1$, there is no difference between the value of acting with perfect information and the value of acting with current information (current information is already perfect). When there is uncertainty however, a difference, Q appears. For example, if the probability of hot weather is 50% (roughly where the Q line is drawn in Figure 3.2.1), the expected value with imperfect knowledge is the average of V_1 and V_2 (the entrepreneur chooses his best option, the lemonade stand, and receives V_1 with probability $\frac{1}{2}$ and V_2 with probability $\frac{1}{2}$), while the expected value with perfect information is the average of V_1 and V_4 (with probability $\frac{1}{2}$ the entrepreneur will be informed that the weather will be hot and he will set up a lemonade stand and receive V_1, and with complementary probability $\frac{1}{2}$ the entrepreneur will be informed that the weather will be cold and he will set up a lemonade stand and receive V_4). The difference between these two expected values is Q, the amount the entrepreneur should be willing to pay to obtain perfect information.

There are three major differences between this simplified example and real-world applications.

1. In research, perfect information is never obtained, only sample information.

2. The action set (i.e. lemonade stand vs cocoa stand) is larger, and in many applications, is parameterized such that there are an infinite number of possible actions over a range of actions.

3. The state of knowledge (i.e.weather is a binary variable and the state of knowledge is a simple statement of the probability of one outcome) is more complex. Almost all real-world applications will involve a state of knowledge that has two parameters (typically mean and variance) or more. For example, rather than represent weather as "Hot" or "Cold" it would be more accurate to describe it as a degree temperature, and the state of knowledge would be a probability distribution function across the range of temperatures.

Perfect Information vs. Sample Information

To illustrate the example of the difference between perfect information and sample information, let's return to the coin-flip example with Paul. Suppose again that a coin is either heads or tails biased such that the biased side comes up 75% of the time.

Paul is asked to predict which side of the coin it is biased towards. If he predicts that the coin is heads-biased and he is correct, he will receive \$8. If he predicts that the coin is tails-biased and is correct, he will receive \$4. If he predicts incorrectly, either heads or tails, he receives nothing. In other words, $V_1 = 8$, $V_2 = 0$, $V_3 = 0$, and $V_4 = 4$.

Paul initially believes that there is a 50% chance the coin is heads-biased, and 50% tails-biased. He asked how much he would be willing to pay to be able to flip the coin once prior to making his prediction.

Paul's highest valued action without new information is to predict heads, which offers an expected value of 4. After the flip, there is a 50% chance probability that the outcome will have been tails, in which case Paul's highest valued action is to predict tails with an expected value of 3 (the probability that the coin is tails-biased is 75%). There is also a 50% probability after the flip that the outcome will have been heads, in which case Paul's highest valued action is to predict heads with an expected value of 6. Paul's expected value given sample information is $\frac{1}{2} * 3 + \frac{1}{2} * 6 = 4.5$. Thus, Paul is willing to pay 0.5 for his first flip. Similar calculations show that after a tails result, Paul's willingness to pay for a coin flip goes up to 0.7, while after a heads result it declines to 0.4. This is as expected, since the value of information peaks at $P = \frac{2}{3}$, where the expected values of predicting tails and heads are equal.

Expanded Action Sets and Complex Knowledge States

Modifying the refreshment stand example to include a broader action set is simple.

Adding new action possibilities (in Figure 3.2.1 the new option is a hot dog stand with constant payout, $E[V] = V_5 = V_6$), does not change the fundamental statement of the problem, which is that for any state of knowledge there is an expected value based upon the optimal action given that state of knowledge, and that changes in a state of knowledge are valued only in as much as they lead to states of knowledge that have a higher expected value. With the inclusion of a hot dog stand in the entrepreneur's mutually exclusive action set, the value of knowledge states in the middle of the probability range has increased, but no paradigmatic change has occurred.

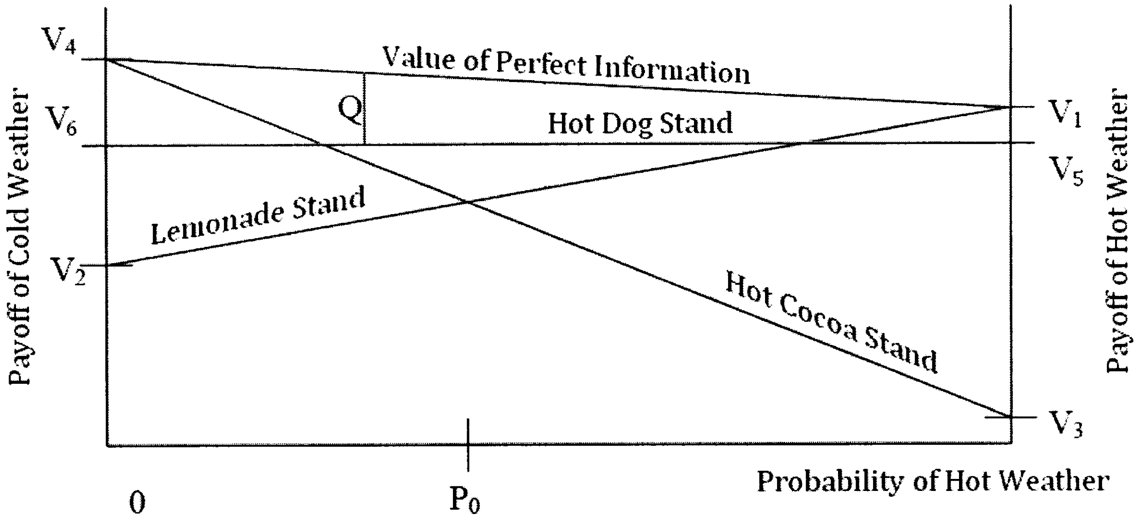


Figure 3-2: The Value of Information, Expanded Action Set

Similarly, increasing the complexity of the state of knowledge does not fundamentally change the problem. The broader the set of possible states of knowledge, the more states there are to assign valuations to, but the calculation of expected value remains as it was before. As long as it is remembered that the benefit of research is dependent upon *actions*, providing valuations to states of knowledge should be relatively simple. Depicting the action set should be the starting point of any attempt at assigning values to states of knowledge— at its heart, determining the value of a knowledge state is simply answering two questions: given this knowledge, what is the optimal action to take, and what would be the expected result of that action?

Expert Review Panels

Given the difficulties in quantitatively measuring research outcomes, it is important to develop reliable subjective means of assessing the value of different states of knowledge. Expert review is similar to the standard practice of peer review, with the exception that expert review groups include not only peer researchers, but also a broad range of other experts, including members of other scientific fields, economists, lawyers, policy analysts, and other end-users of scientific research. The increased breadth of expert review panels relative to peer review gives them the unique ability to judge not only the scientific merit of a research project, but also its value to society at large.

Many federal agencies maintain standing expert review boards to provide independent analysis of agency programs and decisions. These boards regularly make judgments on how well the agency is setting research priorities and managing its resources. Given their availability and capability, expert review boards make ideal bodies for supplying Bayesian networks with valuations of knowledge states. The main focus of these expert panels would not be to assess the scientific merit of a project (an assessment that can be left to traditional peer review), but instead to estimate what the benefits of knowledge derived from a project would be. For most projects, this may be a straightforward estimate of the economic benefits provided by an improved process or product. For a military research project, this may involve estimating how well a new weapons system would improve military capabilities relative to other potential weapons systems. For a project serving the Environmental Protection Agency or the Food and Drug Administration, this might mean estimating the costs and benefits of regulation formed on the basis of research.

In some cases, expert review panels might even be asked to speculate about the benefits provided by basic research. In strict terms, the valuation of research should be derived from the products, processes, and future research that it enables and the definable benefits that these developments provide— but for basic research (as defined Section 1.2.1) it will not be practical to characterize the new research avenues opened by different research outcomes, in which case it will be easier to ask expert panels to directly assess the value of basic research outcomes.

3.2.2 Obtaining Priors

After giving value to research outcomes, the next step is to solicit a prior distribution on each of the relevant parameters of an experiment. In the simple coin-flip examples discussed previously, naive priors were used— it was assumed at the start of each example that no knowledge existed about the bias of the coins, validity of alloys, or future temperatures. In the real world, prior beliefs exist. They come from objective sources, such as previous experimental evidence, and subjective sources, such as expert opinion.

More generally, expert opinion and theoretical knowledge are needed to establish a model of the world. In the hypothetical situations worked through thus far, the model of the world was given as part of the example, i.e.there is a coin that must either have a heads or tails bias of 75%. By establishing this framework, it was made clear what the conditional independencies were between the observations (heads or tails results) and the hypotheses (heads or tails bias). In real applications, it will often be necessary to solicit statements of conditional probability from experts in order to create the links between nodes. This will not always be the case— in some research projects (human drug trials may be a good example), uncertainty analysis is part and parcel with the experiment, and the likelihood functions that should be used to update the prior distribution are self-evident. In other cases however, the statistical power of the experiment cannot be objectively derived either before or after the experiment is conducted. Researchers must provide some estimate of their confidence that an experimental result is free from error. This estimate will depend upon both aleatory and epistemic sources of uncertainty. Aleatory sources of uncertainty may be thought of as imprecision in measurement instruments, natural variability in tested samples, and so on. In other words, aleatory uncertainty is the variability between results that would occur if the experiment were run multiple times using the exact same experimental setup, and as such, aleatory uncertainty can often be determined beforehand through calibration. Epistemic sources of uncertainty may be thought of as errors in the execution or design of the experiment. Sources of epistemic uncertainty are generally unobserved and if observed, corrected. They may include experimenter error, poor control of experimental conditions, and so on. As it is unknown how often and to what degree experimenter error affects a test result, it is difficult to depict some sources of epistemic uncertainty. More interestingly, some sources of epistemic uncertainty can be predicted beforehand but cannot be removed. For example, suppose an experimenter tests a drug on laboratory mice and obtains a result. As a measurement of the effect of the drug on mice, the experimenter might have a very precise answer. But no matter how rigorously the experiment was run, there will always be uncertainty in extrapolating the evidence from the experiment to the relationship the experimenter is truly interested in: the effect of the drug on humans. Thus, in determining the conditional probability of drug efficacy in humans given a result in mice experiments, it is not enough to simply observe the results in mice, it is necessary to solicit from the researcher some form of statistical inference that can be used to relate the evidence to the parameter of interest.

Peer Review Panels

A natural way of soliciting both prior beliefs and conditional relationships between variables is a peer review panel. Because the questions being asked are technical in nature, it is better to bring in a greater degree of specialization than one would find in an expert review panel.

Peer review assessment in this context should be limited to depicting the current state of knowledge on the variables of interest in an experiment, estimating the uncertainty on the potential results of experiments, and ensuring that all the potential results are characterized by the belief network. It is unnecessary to try and estimate the societal value of given results except in the case of basic research when an expert review panel felt that it lacked the technical expertise necessary to appreciate the value of an experiment. Peer review panels tasked with coming up with conditional probability relationships and prior beliefs should include at least one statistician to ensure that the panel understands the nature of the questions they are being asked to answer.

3.2.3 Principal-Agent Relationships

When soliciting prior beliefs or conditional independencies, a problem may arise if the research in question originated with, or is being performed by, an outside party. In these cases of extramural research, it may often be that the extramural party has the most informed understanding of the research problem, and is therefore the source of the most accurate priors and conditional probabilities.

What creates the problem is that the extramural researcher and the funding body may not share the same interests. The priors and conditional probabilities that the researcher submits (in short, his expectations for success), will factor into the decision of not only whether or not the research project should be begun, but also whether or not it is discontinued as new evidence comes in. While the funding body seeks only to maximize the expected return on its research dollar, a researcher has a self-interest in seeing his own research funded, regardless of the return that it might provide for the funding body.

This problem that forms when there is asymmetric information and dissimilar interests between two contracting parties is called a principal-agent problem. The common manifestation is in labor markets, where an employer hires employees, but has imperfect information on the employee's capabilities and level of effort.

In the context of research and development, the agency problem at hand is that the principal (the funding body), has a willingness to pay for research that depends upon the expected value of the research and the effort of the agent (the researcher), while the agent, who has insight into the expected value and control over his own level of effort, has a different calculation of self-interest than that of the principal.

In most principal-agent problems, the primary concern is not the agent's personal estimation of how much benefit he can potentially provide to the principal, but instead the main focus is on the second part of the problem, incentivizing the agent to provide effort when the principal lacks the means to directly observe effort. In this sense, the research agency dilemma problem is unique: it is not generally observed that researchers fail to provide effort— a more common complaint is that funding requests exaggerate the potential benefits that the funds can provide.

Still, the form of the solution is the same. In problems where effort must be induced, a typical solution is to replace lump-sum payments with piece-rate payment. In the optimal solution, an agent's total compensation package is the combination of a lump-sum payment (which can take both positive and negative values) and a reward that is proportional to the output provided: $Payout = C_1 + C_2 x$. C_2, the fraction of an agent's expected compensation that comes from the output/effort-proportional term is increasing with four factors: the marginal benefit created by marginal effort, the accuracy with which effort can be assessed, the agent's responsiveness to incentives, and the agent's tolerance to risk [16].

In cases where honesty must be incentivized, the solution is to replace lump-sum payments with a payment that is proportional to the absolute difference between the output prediction given by the agent and the actual output. A payout function of the form $Payout = C_3 - C_4 \mid x - e \mid$ incentivizes the agent to report the median expectation of output, while a payout function of the form $Payout = C_3 - C_4(x-e)^2$ incentivizes the agent to report the mean expectation of output.

Once again however, in the optimal solution, C_4 is increasing with four factors: the marginal benefit created by marginal increase in accuracy of the agent's expectation, the accuracy with which the difference between the expectation and the result can be assessed, the agent's responsiveness to incentives, and the agent's tolerance to risk.

It is the last of these four factors that creates biggest difficulty in solving the principal-agent problem. Research grants are on the scale of hundreds of thousands of dollars. It would be infeasible to loan researchers hundreds of thousands of dollars to conduct research, and then only compensate them if the results were useful to the funding body or if the research yielded the results the researcher expected. Researchers do not have such high tolerance to risk. We should expect an optimized C_4 to be very low and dominated almost entirely by a researcher's willingness to risk his own money in return for obtaining a research grant. Also, a significantly high C_4 would incentivize researchers to falsify results (decrease the accuracy with which the difference between the expectation and the result can be assessed).

Potential methods of resolving the agency problem discussed here will be revisited, along with methods of expert solicitation, in Chapter 7.

3.2.4 Empirical Updating

After every new experimental result, empirical updating should be performed using Bayes' theorem in conjunction with the prior distributions and likelihood functions solicited from research staff and experts. Because this updating may require difficult integrations to find the normalizing constant, Pr(B), two implementation methods have been developed.

Conjugate Priors

If the prior distribution of the variable being updated is conjugate with the likelihood distribution being used to update it, the updating is a straightforward process where the hyperparameters of the prior and likelihood function are used to determine the hyperparameters of the posterior distribution. Conjugate priors greatly simplify the updating process, and fortunately, most natural sources of uncertainty yield distributions for which there are several conjugate counterparts.

A common distribution used in statistics is the normal distribution, $\mathcal{N}(\mu, \sigma^2)$, which has the probability density function

$$f(x \mid \mu, \rho) = \sqrt{\frac{\rho}{2\pi}} \exp\left(-\frac{\rho(x-\mu)^2}{2}\right) \tag{3.3}$$

If a conjugate prior, π is a normal distribution with hyperparameters m and p tracking a parameter, μ

$$\pi(\mu \mid m, p) = \sqrt{\frac{p}{2\pi}} \exp\left(-\frac{p(\mu-m)^2}{2}\right) \tag{3.4}$$

and a set of measurements, $x_1, \ldots, x_n$, are independently and identically obtained from a normal distribution with unknown mean, μ, and known precision, ρ, then the likelihood function will be

$$\mathcal{L}(\mu \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(\mu \mid x_i) = \left(\frac{\rho}{2\pi}\right)^{n/2} \exp\left(-\frac{n\rho(\bar{x}-\mu)^2}{2}\right) \tag{3.5}$$

and the posterior, $\pi(\mu \mid m\prime, p\prime)$ will be a normal distribution with hyperparameters

$$m\prime = \frac{mp + n\rho\bar{x}}{p + n\rho} \quad \textit{and,} \quad p\prime = p + n\rho \tag{3.6}$$

Similar relationships exist for likelihood functions with known μ and unknown ρ, unknown μ and unknown ρ, and many others, such as lognormal, gamma-normal, and multi-dimensional normal distributions.

One advantage of using conjugate priors is that the updating process is transparent and easily understood by those with limited statistical backgrounds. It is easy to track changes in hyperparameters and understand the significance of the updating process. This makes Bayesian networks more accessible to researchers, more easily implementable, and better helps researchers to design or modify experiments to maximize the expected value of an experiment.

A disadvantage of using conjugate priors is that they may not be the best reflection of prior beliefs or evidential forms. Even though many conjugate distributions depict common, natural processes, they are still a small subset of all possible distributions. Therefore, in many circumstances, it will be necessary to use numerical methods.

Monte Carlo Assessment

Monte Carlo methods are a class of algorithms that generate large numbers of random samples from distributions as a means of approximating the features of the distribution itself. They are particularly useful in the evaluation of integrals— random sampling of points over the domain that is being integrated will yield an estimate of the integral's value. This feature of Monte Carlo numerical methods is particularly useful for evaluating the normalizing constant used in the updating process, $f_Y(y)$ (or Pr(B) if the discrete form is being used), which is the integral of the product of the prior and likelihood functions over the relevant domain of the parameter being tracked.

Monte Carlo methods are far more computationally cumbersome than the analytical methods enabled by conjugate priors, but are necessary in most realistic applications of Bayesian decision making.

3.3 Chapter Summary

Belief networks are a type of linked-node system, with the nodes representing parameters, variables, hypotheses, or forms of evidence, and the linkages between nodes representing statements of conditional probability. Using Bayes' theorem, changes in the state of one node can be used to inform or "update" the states of other nodes. Depending on the types of probabilistic distributions used to define the nodes and their relationships, the process of updating can be simple and analytical, or require more computationally intensive methods. In the context of research decision making, the driving concept behind the use of belief networks is to assign societal value to nodal states (or, equally, define a new node that reflects societal value), and thereby connect research activities (which change nodal states) to ultimate outcomes and benefits. To establish a belief network, two pieces of information are needed: a set of prior beliefs (initial values for the nodes) and conditional probability relationships (the connections between the nodes). These can be solicited from expert and peer-review boards, among other methods, though it should be noted that soliciting from researchers who have a personal interest in project funding may create adverse incentives. To use belief networks to their full potential, one more element is needed: a method of assigning value to nodal states. The value of a state is defined by the expected value of the actions that can be taken in that state. Similarly, the expected value of changing states (and thus the expected value of a research activity) can be defined as the difference, between states, of the expected value of the highest valued action. With this final element, belief networks enable cost-benefit analysis and optimization.

Chapter 4

Description of the Case Study

In order to demonstrate the potential of Bayesian networks as a tool for improving research and development outcomes, we have applied the method to a prototypical case study: the Department of Energy's Nuclear Energy Research Initiative, Project Number 06-038, *The Development and Production of a Functionally Graded Composite for Lead-Bismuth Service.* Funded by public grant (Award Number DE-FC07-06ID14742) the project is led by Dr. Ronald Ballinger of the Massachusetts Institute of Technology in collaboration with the Los Alamos National Laboratory (LANL) and the Idaho National Laboratory (INL). Begun in April of 2006, the project investigates whether a functionally graded composite of T91 steel and Fe-12Cr-2Si alloy, manufactured using standard commercial practices, can meet the operating requirements of a liquid lead cooled nuclear reactor at temperatures greater than existing operational limit of 547°C[14]. If successful, this bilayered alloy could be used to manufacture fuel rod cladding and coolant piping in a lead-bismuth-eutectic-cooled fast-reactor.

Originally, it was intended that Dr. Ballinger's research project would last three years and conclude in April 2009. However, as a consequence of intermittent funding authorization, the project's progress has been delayed by at least one year. As a result, this thesis lacks the time horizon necessary to conduct Bayesian monitoring to the full extent that would be expected of a federal agency or oversight body. However, we deem that sufficient progress has been made by the case study to provide a valid proof of concept and justify further attention.

Project 06-038 has many characteristics that make it ideal for a prototypical case study. The project is an appropriate example of "hard" science: it employs the scientific method and provides quantitative evidence through experimentation. It is a good representation of the basic and applied research to which Bayesian methods offer the greatest potential for improvement: the project is neither theoretical (in which so little is known about the possible results that it is difficult to characterize the project's activities into a belief network) nor a mature development project (in which so much is known that the project's results are more deterministic than probabilistic). Finally, the project is concerned with multiple parameters— corrosion, diffusion, yield strength, etc— and thus provides an implementation challenge that is sufficiently complex without obscuring the main purpose of the case study, which is a proof of concept demonstration.

4.1 Motivation for Alloy Development

Lead-bismuth cooled nuclear reactors offer several design advantages over comparable fast-reactor designs. Molten lead has favorable heat transfer and heat capacity characteristics, and its low melting point, high boiling point, and stability in air and water make it relatively safer than molten sodium for use as a reactor coolant[1]. Unfortunately, lead cooled advanced reactor systems have been historically limited by the high rate of corrosion of nickel-based alloys in the lead environment. This limitation precludes the use of many standard reactor materials, including the set of stainless steels typically employed in a reactor's primary loop construction. Consequently, existing materials are only suitable up to $\tilde{5}50$°C, a limit that reduces economic attractiveness and prevents lead-bismuth reactors from realizing the full thermal efficiency potential of their design.

The most demanding material problem occurs in the fuel cladding. Fuel cladding is a thin metal sheath that surrounds the fuel and maintains the fuel rod's structural integrity during reactor operation. Due to thermal and neutronic requirements in the reactor core, cladding must be very thin, on the order of a fraction of a millimeter, but also thick enough to withstand corrosion and meet strength requirements. To make lead-cooled reactors a viable design choice, there is a great need to develop alloys that satisfy a demanding set of material property requirements.

Previous testing suggested that the use iron-chromium-silicon alloys might raise the temperature limit currently faced by lead cooled designs to 650°C or more[15]. Fe-Cr-Si alloys develop stable protective films over a range of oxygen potentials, provide strong resistance to corrosion, and do not exhibit a vulnerability to stress corrosion cracking.

Based upon this preliminary evidence, a solution to the materials limitations of lead-cooled reactors was proposed: a bilayer composite, made of an inner, structural layer of T91 (a nickel-free steel), and an outer, corrosion resistant layer of Fe-12Cr-2Si. This functionally graded composite might meet both the structural and corrosion requirements of the lead-bismuth environment. The dimensions of the cladding product form to be tested were chosen such that they would be compatible with the established fuel cladding dimensions to be used in the Department of Energy's Advanced Burner Reactor program. Furthermore, it was decided that the product would only be produced using conventional commercial practices, so as to demonstrate readiness for immediate use as a prototype.

4.2 Project Tasks

To be successful as a product form, the proposed functionally graded composite needs to guarantee, to an acceptable degree of certainty, that its material properties and specified product dimensions are sufficient to maintain the clad's structural interity over the course of a three-year fuel operational lifetime. Success therefore depends upon a number of factors, including the corrosion rate of Fe-12Cr-2Si in high temperature lead, the diffusion rate of silicon between layers during operation, the dilution of silicon during production, the mechanical strength of the inner structural layer, the commercial fabricability of the product form, and, as an ancillary factor, the commercial cost of the product form itself. Project 06-038 investigates each of these factors.

4.2.1 Product Form Procurement

Two product forms, piping and cladding, will be produced using standard commercial practices. Of the two, the cladding faces the more stringent materials requirements, and so the focus of the case study is primarily on the challenges faced by the clad product form.

The cladding will be produced in a series of stages. In the first stage, ingots of T91 and Fe-12Cr-2.5Si will be obtained. The T91 ingot will be processed into an extrusion billet through direct melting, and the Fe-12Cr-2.5Si will be custom forged into a billet and then fabricated into weld wire. In the second stage, the T91 will have a weld overlay applied to its outer diameter using the weld wire fabricated from the Fe-12Cr-2.5Si billet. Multiple weld passes will be made in order to build up an outer layer of corrosion resistant material. In the third stage, the weld-overlaid extrusion billet will be reduced through a combination of hot extrusion and roto-rolling (pilgering) to achieve pre-drawing dimensions for the cladding form. In the last stage, the product will be tube drawn to its final product dimensions. The final product form will be a hollow cylinder with an inner diameter of .257" and an outer diameter of .28". It consists of a .02" inner structural layer of T91, and a .003" outer, corrosion resistant layer of Fe-12Cr-2Si.

The procurement process is being performed entirely through commercial metalwork providers and other contracted services. Both fabricability and commercial product cost are relevant to the success of the experiment and can be considered variables in the belief network affecting the value of the research project, to be updated as new service providers are identified and production tasks are completed.

4.2.2 Corrosion Testing

The primary material requirement of the alloy is corrosion resistance. Over the lifetime of the cladding, some thickness of the protective layer will become corroded away through contact with the molten lead. In order for the product to be viable, the corrosion rate must be small enough such that at the end of clad lifetime the outer layer still offers protection to the inner structural layer.

Static corrosion tests will be used to determine the corrosion resistance of the alloy. 100mm diameter disk samples will be immersed in a molten lead bath within a high temperature furnace. The furnace will be equipped with an oxygen control and monitoring system to maintain oxygen potentials within specified levels. 1000-hr tests will be conducted over a broad range of oxygen potentials and a temperature range of 600°C to 700°C. After testing, samples will be analyzed using standard metallographic and scanning electron microscopy, as well as X-ray techniques. These analyses will yield an estimate of the long-run corrosion rate of the alloy. More details on the testing facility are included in Appendix E.

4.2.3 Diffusion Testing

Because the silicon concentration of the inner structural layer is 0.4% while the Si concentration in the outer corrosion resistant layer is expected to be 2%, there will be diffusion of silicon between the two layers during reactor operation. In order for the outer layer to maintain its corrosion resistant properites, it must have a silicon concentration of over 1.25%. Thus, over the lifetime of the cladding, some thickness of the protective layer will become degraded due to the loss of silicon.

The change in silicon concentration over time can be described using a diffusion constant, D_{Si}, and Fick's Second Law

$$\frac{\partial C}{\partial t} = D_{Si} \frac{\partial^2 C}{\partial x^2} \tag{4.1}$$

Solving Fick's Second Law in a one-dimensional setting yields a silicon concentration profile

$$C = C_{avg} + C_{\Delta} erf(\frac{x}{\sqrt{4D_{Si}t}}) \tag{4.2}$$

Thus for a given diffusion constant, D_{Si}, and length of time, t, there is a thickness, x, of the protection layer for which the silicon concentration, C, is degraded below 1.25%.

The diffusion of silicon in iron follows an Arrhenius relationship, and the diffusion constant, D_{Si}, can be described by

$$D_{Si} = D_o e^{\frac{-Q}{RT}} \, (4.3)$$

where D_0 is the diffusion coefficient, Q is the activation energy for the diffusion of silicon in iron, R is the gas constant ($1.986*10^{-3}$ kcal/mol-K), and T is the temperature of the material in Kelvin.

Diffusion tests were performed to obtain an estimate on the diffusion coefficient. Test samples were created by Hot Isostatic Pressing couples of T91 and Fe-12Cr-2Si together at 1050°C for five hours. The couples were then sealed and aged in a furnace for 274, 495.2, and 1219.8 hours at 800°C. Samples were then removed and analyzed for silicon content— the results were fitted to an error function profile to obtain an estimate on D_0.

4.2.4 Dilution Testing

Loss of silicon from the protective layer will also occur during processing. During the weld overlay process there will be migration of silicon from the weld wire to the T91 billet. Although there is no explicit measurement of dilution of silicon, analysis of the as-received product forms will provide a useful estimate of the migration of silicon that should be expected from commercial processing.

4.2.5 Mechanical Strength Testing

Because molten lead coolant is not pressurized, the mechanical strength requirements of the cladding product form are not as demanding a condition as they would be in most light water reactors. In the lead-coolant system, the primary structural requirement placed on the clad is that it must support its own weight, for which T91 should be sufficient: at temperatures of 700°C, T91 has an ultimate tensile strength of 200MPa and a yield strength of 160MPa[25, 26].

Although the mechanical properties or T91 are known to a high degree of certainty, there is a risk that the carbon in the T91 steel will defuse to the outer protective layer. It is unclear to what extent carbon depletion in the structural layer will weaken the clad, and so the couples used previously in diffusion testing will undergo NANO-hardness testing on a Hysitron nanoindenter. Hardness testing should not only provide an estimate on the diffusion of carbon between the layers, but also reveal the mechanical properties of the T91 after some amount of carbon migration has occured. The structural properties of the product forms will also be explored at 600°C and 700°C in the form of tensile and 1000hr stress rupture tests.

4.3 Chapter Summary

For our case study, we selected an ongoing research project at MIT being funded through the Department of Energy. The project is testing the corrosion, diffusion, dilution, and mechanical properties of a bilayered alloy product for use in a specialized environment. The project has a clear motivation and easily characterized test results— as such provides an ideal candidate for a proof of concept demonstration.

Chapter 5

Applying Bayes Theorem to NERI Project 06-038

To model a research project using a Bayesian network, it is necessary to obtain three key pieces of information through expert elicitation:

- A definition of success or method of knowledge valuation that relates societal benefit to the parameters being explored by the experimental activities.
- Statements of prior belief on the parameters being tracked by the Bayesian network.
- Statements of conditional probability (likelihood functions) that relate the parameters tracked to experimental observations.

Prior beliefs and likelihood functions were solicited from the project's principal researcher, Professor Ronald Ballinger, and his graduate assistant, Michael Short. A method of knowledge valuation, which would normally be solicited from the funding body or an expert review panel, was created in consultation with Prof. Ballinger and Mr. Short.

5.1 Choice of Parameters

In order to provide the clearest possible application of Bayesian networks and avoid unnecessary complexity, some relevant parameters were not tracked. Corrosion and diffusion behavior were chosen to be tracked, while fabricability, cost, dilution, and mechanical strength were not. Before continuing, it is important to explain why this particular choice was made and provide a brief description of how the ignored parameters could be added as nodes to create a fuller Bayesian network.

5.1.1 Corrosion and Diffusion

It is self-evident why corrosion behavior was chosen as a parameter to track. Corrosion resistance is the primary motivation for alloy development and occupies the attention of the large majority of the project's proposed experimental activities. The corrosion behavior of the Fe-12Cr-2Si overlay is far and away the single largest source of uncertainty in determining alloy viability and because of this, it would have been impossible to exclude corrosion from the Bayesian network and still obtain meaningful results.

Diffusion warranted attention for two reasons. Firstly, although the diffusion of silicon between the product's bilayers is is of secondary importance in determining the alloy's viability, it still represents a significant source of uncertainty and thus merits tracking. Secondly, diffusion behavior is one of the few areas in which, as of writing, the project has returned experimental results. As such, diffusion represents one of the few areas where a clear demonstration of Bayesian updating can be made.

The explicit corrosion parameter tracked was the long run corrosion rate, as extrapolated from the measurement of material lost over the course of a corrosion test. The explicit diffusion parameter tracked was the estimate on D_o determined by error function fitting of the measured silicon concentrations within the tested diffusion couples.

5.1.2 Other Parameters

Next to corrosion, fabricability and cost are the largest determinants of alloy viability. To date, the majority of the project's activities have been devoted to procurement efforts, obtaining alloy billets and contracting for materials processing services with various metal working companies. The question of whether the proposed product form can be produced at reasonable cost with a sufficiently low rate of defects is one of the most important considerations of the project, and the process of soliciting offers and contracting for services is an important means of information gathering. Were fabricability and cost to be treated as parameters in our Bayesian network, they could be described using three quantitative parameters: the probability that the product form can be produced at all (a parameter that takes values between 0 and 1 and can be estimated as the top-level measure of a fault tree that includes each processing stage as a node), the rate of defect per unit length of final product form (or alternately a probability distribution on the largest undetected defect per fuel rod length), and the cost per unit length of final product form.

Fabricability and cost are important determinants of alloy viability: if the functionally graded composite cannot be produced at all, has too great a rate of failure due to defects in the product form, or can only be procured at a cost that inhibits commercial application, then the usefulness of the alloy is diminished.

Despite the importance of these parameters, it was decided not to include them in the case study's analysis. While procurement represents a significant source of uncertainty for this project, it is not a good representation of the traditional process of research that this thesis is interested in. The forms of evidence are often simple, binary results ('Works' vs 'Doesn't work') that can be easily translated into project success or failure without the aid of belief networks, and because the marginal cost of soliciting estimates from service providers is negligible, it is unclear how Bayesian decision making would improve current practices. Fabricating the alloy product forms is more similar to capital procurement (like the construction of a testing facility) than research and development, and capital procurement, even in the context of research and development, is already well served by existing decision-making practices and efficiency metrics.

The remaining parameters, dilution and mechanical strength, were not tracked for a combination of reasons. Firstly the contribution of these parameters to the uncertainty on alloy viability is small. Dilution of silicon occurs primarily during the weld overlay process in the production stage. During the first welding pass, dilution will occur over a thickness of roughly only .032" of the material, but subsequent weld passes will not make direct contact with the T91 and dilution will therefore not be a concern. Also, because the weld-overlay process occurs while the product is still in a billet form and has an outer diameter of roughly 10", the effect of dilution will be even further diminished: after welding, the product will be reduced through extrusion and tube drawing down to an outer diameter of .28", nearly a forty-fold reduction in size, and the thickness affected by dilution during welding will be reduced proportionally, putting it on the order of .001". This places a physical upper-bound on the effects of dilution and limits its potential impact on the product form's viability. Furthermore, because the weld-wire has a slightly higher concentration of silicon than the aimed for final protective layer concentration of 2%, the effects of dilution are well compensated for and are not expected to provide a significant threat to alloy validity.

Mechanical strength is also a minor contributor to the uncertainty of alloy viability. The material properties of T91 are known to a high degree of certainty, and it is expected that at 650°C, T91 will the meet the cladding yield strength requirements[25, 26]. Furthermore, the project does not devote significant attention to either the issue of silicon migration during the processing stage or the mechanical properties of T91. Evidence of dilution is treated as an ancillary task, and the confirmatory mechanical properties testing is effectively a product of convenience: the samples already exist as a consequence of diffusion testing, and little experimental work is needed to provide an estimate. There is no need to recreate the precision of the older, in-depth studies of T91.

If dilution were a tracked parameter, it could be integrated with the measurement of diffusion to obtain an estimate of the thickness of the protective layer that loses its corrosion resistant properties over the course of the clad life. If mechanical strength were a tracked parameter, it would be used to create a probability that the composite alloy has sufficient mechanical strength to maintain its structural integrity over the course of a fuel rod lifetime.

5.2 Definition of Success Space

Using the selected parameters, the condition for experimental success, "Alloy Viability," was defined in the following way:

If utilized in a commercial lead-bismuth reactor, the cladding product form would be expected to remain in the reactor environment for three years (26280 hours) at 650°C. From the corrosion rate and the diffusion constant, we can extrapolate a total thickness, T_T, which is the sum of the thickness of the protective layer that has been corroded away, T_C, and the thickness that has seen its silicon concentration fall below 1.25% due to diffusion into the structural layer, T_D, over the course of the three years of operation. If this extrapolated T_T is less than the thickness of the protective layer in the as-tested dimensions of the product (.003"/.0762mm), then the alloy will be considered viable, if not, then it will be considered inviable.

T_C is easy to extrapolate from the corrosion rate. If the corrosion rate is given in μm per hour, then T_C in mm is simply the corrosion rate multiplied by 26.28.

T_D is a somewhat more difficult extrapolation. Using Equation 4.2 we can determine that if C = 1.25, C_{avg} = 1.2, and C_Δ = 0.8 (the concentration of silicon in the protective layer is 2.0% and the concentration of silicon in the structural layer is 0.4%), then T_C is equal to 0.055446 $\sqrt{4D_{Si}t}$. If D_{Si} is expressed in units of $\frac{cm^2}{s}$, then the total thickness degraded by diffusion, in mm, is equal to 10786.08 $\sqrt{D_{Si}}$.

Because the parameter being updated is not the diffusion constant, D_{Si}, but instead the diffusion *coefficient*, D_o, it is necessary to relate D_o to D_{Si} using Equation 4.3. Here, one complication is introduced: Q, the activation energy for diffusion of silicon in iron, is not known with absolute certainty. Therefore, to track the effect of a change in the estimate of D_o, it is necessary to account for the uncertainty on the estimate of Q as well.

Alternate definitions of alloy viability can be considered as well. For example, it could be assumed that the thickness of the clad does not necessarily need to equal the dimensions of the as-tested product form, but instead could be thicker or thinner, with the achievement of thinner cladding having more value than that of thicker cladding. For a given clad thickness, a certain set of reactor design options may become available, with thinner cladding enabling a broader set of reactor designs. The value of a given cladding would then be dependent on the set of reactor designs it enabled, with the value of a set of reactor designs dependent on the characteristics of the reactors, their economic advantages, waste generation rates, resource efficiencies and so on. Alloy viability would then be described not as a simple boolean variable, but instead as a set of multiple boolean variables representing the alloy's capability of meeting the materials requirements of an array of reactor designs.

5.3 Valuation of Knowledge States

For simplicity and the purposes of demonstration, we have settled on a simple means of valuation:

The direct benefit of an alloy is determined by the action set with the highest expected value. In its simplest form, we can imagine the action set related to an alloy to be "Use in Reactor Construction" or "Don't Use in Reactor Construction." For the action, "Don't Use" we can assume the value of the alloy to be zero— without some physical application of the alloy, there can be no societal benefit. For the "Use" option we can imagine two potential payoffs: a benefit, B, incurred when a viable alloy is used, and a harm, H, incurred when an inviable alloy is used. For example, using a viable alloy in a lead-bisumth reactor would allow the reactor to achieve greater thermal efficiency, producing benefit, while using an inviable alloy would result in fuel failures, causing harm in the form of damage to and decreased utilization of the reactor, as well as radiological dose to workers.

In a method identical to that described by Figure 3.2.1, we can construct expected value curves for both actions:

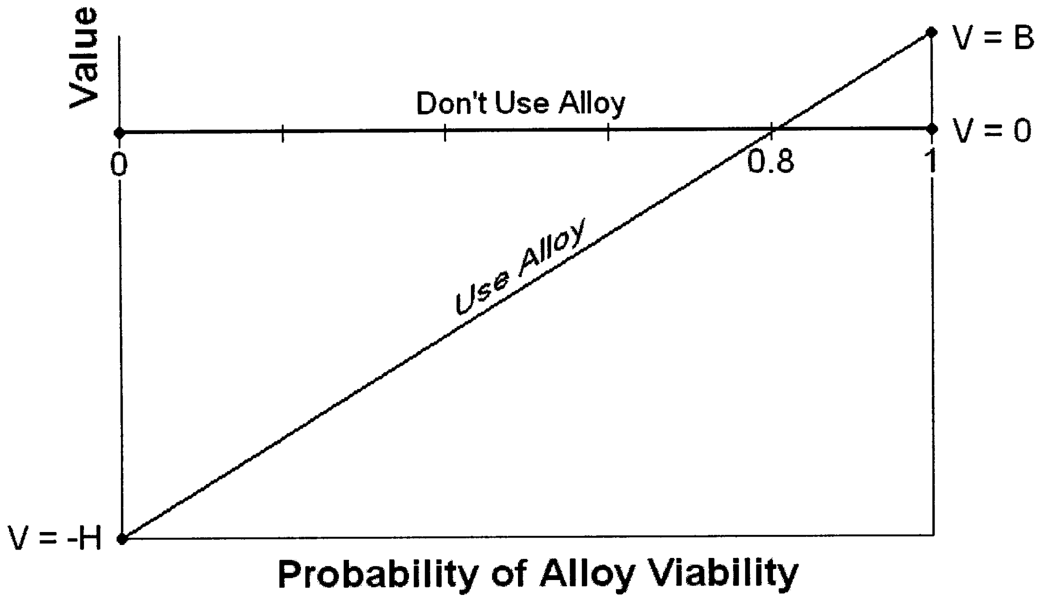


Figure 5-1: Value of Alloy Actions as a Function of Probability of Viability, "80% Standard"

If the H is equal to 4B, then the expected value of using the alloy is greater than the value of not using the alloy and is equal to 5B[Pr(Viable)-0.8]. Under this 80% standard, the alloy would only used if the probability of the alloy being viable is greater than 80%— for all lower probabilities there is a higher expected value to not using the alloy. Similarly, we could assume that H is equal to 999B, in which case the Use Alloy action only has positive value for probabilities greater than 99.9%, with a value function equal to 1000B[Pr(Viable)-0.999].

In our analysis, we use five different knowledge valuations, reflecting 80%, 90%, 95%, 99%, and 99.9% standards. This range of valuations should serve as good proxies for many potential valuation schemes. In general, we may expect valuation functions to be concerned with two quantities: the mean of a tracked parameter, and the variance about that mean.

Though not done so here, in practice, it would also be necessary to add another term to the value of different states of knowledge to obtain a complete picture of the value of a research effort. In many instances, research does not yield direct benefits such as a commercially viable alloy (nor is it necessarily the case that this research project does so— that is merely an assumption made to give a hypothetical means of valuation). More often, research enables other research, and for many projects, this effect may be the dominant contributor to the value of a state of knowledge.

A similar process of valuation can be used when this is the case. Suppose the results of a research project, A, affect the expected value of another research project, B. Just as we developed the action set of "Use Alloy" and "Don't Use Alloy" to assign a value to different states of knowledge concerning the viability of an alloy, we could just as easily develop an action set of "Conduct Research Project B" and "Don't Conduct Research Project B" and assign different values to conducting research project B based upon the range of possible results from research project A. The possible outcomes from research project A could be continuous (like the value of the parameter T_T), or discrete (such as "Viable" and "Inviable"). In either case the benefit of project A can be fully described, with all value ultimately derived from research created by societal benefits that are external to the research itself.

It is not expected that the next step immediately after the completion of NERI Project 06-038 should be a decision of whether or not to use the functionally graded composite in a commercial process— the alloy, not to mention molten lead reactors themselves, are much too immature for such a leap. However, the methods of knowledge valuation used here are expected to serve as a good proxy for the more complete method of knowledge valuation that would take place with a broader and more encompassing Bayesian network.

5.4 Obtained Priors

In order to construct our Bayesian network, prior beliefs were needed for three parameters: the diffusion coefficient, D_o, the activation energy for diffusion of silicon in iron, Q, and the corrosion rate, C_{rate}.

5.4.1 Diffusion

Professor Ballinger reported that his beliefs on the diffusion behavior of silicon in an iron-chromium alloy were primarily informed by earlier experiments performed by R. J. Borg and D. Y. F. Lai. Specifically, Ballinger cited a 1970 paper in the Journal of Applied Physics, *Diffusion in α-Fe-Si Alloys*[2].

Borg and Lai analyzed the diffusion of silicon in high purity iron at different concentrations of silicon at varying temperatures. Through work with several samples, estimates were made of both D_o and Q (kcal). Using seven different concentrations of silicon between 0% and 4.21%, Borg and Lai obtained estimates of Q between 52.2 and 52.7 kcal, and estimated D_o to be 0.91278 for silicon concentrations of 2%. Borg and Lai were fairly confident in the precision of their experiment and given the quality of fit of their data, considered ±1 kcal and ± a factor of 2 to be reasonable uncertainties to apply to their estimates of Q and D_o respectively.

Discussion with Ballinger revealed that he largely agreed with the Borg and Lai estimates of diffusion behavior and agreed with their uncertainty estimation. Ballinger noted that as his product form was an alloy of iron and chromium, and not a pure iron diffusion medium, the presence of chromium could affect the diffusion rate by changing the lattice structure of the alloy. Though the significance of this change could not be determined a priori, Ballinger believed the likely consequence would be that the diffusion constant would be slightly lower than Borg and Lai's estimate. Ballinger therefore depicted his prior beliefs on the diffusion coefficient, D_o, to be a normal distribution, with a mean, μ, equal to the Borg and Lai estimate of D_o (0.91278 at a silicon concentration of 2%) and with uncertainty such that ± 0.45639 (50% of the mean value) would serve as a 90% confidence interval on D_o (implying a σ of 0.234). Prior beliefs therefore reflected the view that the mean value of diffusion would be slightly smaller than the Borg and Lai estimate (the use of a linear confidence interval rather than Borg and Lai's geometric confidence interval results in a slightly smaller mean value on the distribution) and largely agreed with the uncertainty assessment provided by Borg and Lai.

Ballinger's prior belief on Q was also largely informed by the estimates provided by Borg and Lai. Ballinger depicted his prior belief on Q as a normal distribution with a mean, μ equal to 52.45 kcal, and a symmetric 90% confidence interval extending to the most extreme values obtained by Borg and Lai, 52.2 kcal and 52.7 kcal. This confidence interval implied a σ of 0.12755.

5.4.2 Corrosion

Professor Ballinger reported that his beliefs on the corrosion behavior of Fe-12Cr-2Si were primarily informed by a Ph.D thesis[15] completed in February 2006 by Jeongyoun Lim, a former student. Lim tested fifteen different iron-chromium, iron-silicon, and iron-chromium-silicon alloys in 600°C Pb-Bi eutectic for 100, 250, 300, and 500 hours. After testing, the samples were analyzed using Secondary Electron Microscopy, Energy Dispersive X-Ray analysis, Electron Probe Micro-analysis, X-ray Diffraction, and X-ray Photoelectron Spectroscopy.

The primary focus of these tests was to characterize the morphology and qualitative aspects of the interaction layers formed by exposure to the molten lead environment. Some quantitative estimates of the corrosion rate could be made using the interaction layer thicknesses observed on the alloy samples, but nothing approaching a definitive estimate of the corrosion rate was made. Lim's research suggested a synergistic effect between chromium and silicon that would form thin, stable, corrosion resistant layers at sufficient concentrations of the two elements.

Although Fe-12Cr-2Si was not one of the alloy compositions tested, alloys with both higher concentrations (Fe-18Cr-2.55Si) and lower concentrations (Fe-12Cr-1.25Si) of chromium and silicon were tested, providing a reasonable basis upon which to infer the corrosion behavior of the product form of concern in Project 06-038.

Drawing from both the qualitative and quantitative results of Lim's research, Ballinger depicted his beliefs on the corrosion rate, C_{rate} as a lognormal distribution with a geometric mean of 0.0022 microns per hour, and a 90% confidence interval over the geometric mean ± a factor of 2. This implied a lognormal distribution (shown in Equation 5.1) with hyperparameters of $\mu = 0.78855$ and $\sigma = 0.30125$.

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right] \tag{5.1}$$

The relatively high uncertainty given to the prior belief on corrosion reflects a number of factors: Lim's use of alloys with compositions different from that of Fe-12Cr-2Si, the limited time span of the runs (the longest runs were 500h), and lack of rigorous quantitative estimates of the corrosion rate.

5.5 Likelihood Functions and Updating

In order to determine the relationships between nodes within the Bayesian network, likelihood functions were needed for two parameters: the diffusion coefficient, D_o, and the corrosion rate, C_{rate}. Likelihood functions are statements of conditional probability— they represent the likelihood of an underlying parameter (in this case D_o or C_{rate}) being equal to a particular value, given an experimental observation.

5.5.1 Diffusion

Diffusion testing was completed in the second quarter of 2008. The diffusion coefficient, D_o was estimated at 1.97. Ballinger characterized the error on this estimate to be normally distributed, with uncertainty such that the estimate suggested with 90% confidence that the true value of D_o lay between ± 50% of the estimate. This implies a normally distributed likelihood function with hyperparameters $\mu = 1.97$ and $\sigma = 0.50255$.

Uncertainty on the evidence was considered to be greater than that of the prior distribution drawn from the previous work of Borg and Lai. Although the Borg and Lai tests studied diffusion of silicon in pure iron while the diffusion tests run by Project 06-038 examined diffusion couples that had the alloy composition of the final product form, the Borg and Lai tests also ran considerably more samples for longer periods of time. As a consequence, the standard deviation, σ, associated with the diffusion testing was roughly twice that of the standard deviation given to the prior.

5.5.2 Corrosion

As of writing, corrosion testing has not been completed. Ballinger has characterized the uncertainty on the experimental estimate of the corrosion rate to be lognormally distributed: each individual sample tested is drawn from an underlying lognormal distribution with unknown mean and known σ equal to 0.13561. This value of σ is such that for each sample tested, there is a 90% probability that the true value of the corrosion rate will be within ± a factor of 1.25 of the estimated value.

5.6 Assessment Methodology

The updating of the diffusion coefficient was performed analytically. The prior distribution, a normal distribution, was conjugate with the likelihood function that updated it, another normal distribution with unknown mean and known precision. The updating process for a normal conjugate prior was described in Section 3.2.4— in short, the resulting posterior distribution is normally distributed with hyperparameters determined by the mean and variance of the prior distribution and likelihood function.

No updating was performed on the corrosion rate parameter. Instead, because the corrosion experiments are unfinished as of writing, corrosion behavior was selected to demonstrate the Bayesian method of evaluating the value of research.

Monte Carlo methodology was used to generate and evaluate random experimental results as follows:

"True" values of C_{rate} were randomly generated using the prior distribution on C_{rate}. Then, for each randomly obtained underlying value of C_{rate}, a random sample estimate on the corrosion rate, $\bar{x}$, was generated using the likelihood function. This $\bar{x}$ was used to update the prior distribution (the prior distribution, a lognormal, is conjugate with the likelihood function, a lognormal with unknown mean and known precision). Then, Monte Carlo was used once more to integrate the probability distribution function of T_T: 10,000 random values of D_o, Q, and C_{rate} were generated from their related distributions (the C_{rate} values were generated from the posterior distribution created by the $\bar{x}$ estimate) to determine an array of T_T values, and these T_T values were compared with the protection layer thickness of the cladding product form, 0.0762mm, to determine the probability of alloy viability. In this manner, 10,000 new states of knowledge were simulated and a probability of alloy viability was estimated for each potential new state.

Twenty Monte Carlo runs in total were completed, ten using the prior distribution on the diffusion coefficient and ten using the posterior distribution. For each run, 100,000,000 simulations were performed: 10,000 unique corrosion test results were simulated and for each result simulated, 10,000 points were sampled from the resulting distribution on T_T to generate a probability of alloy validity. For each set of ten runs, a different number of corrosion test samples were simulated, ranging from n=1 to n=10. This range of n-values reflects a set of potential choices for experimental design that are available in the upcoming corrosion testing and more generally represents a range of potential tests with varying precision, from a relatively low precision experiment simulated by n=1 and a relatively high-precision experiment simulated by n=10.

For each of the twenty runs, a value was assigned to each of the 10,000 simulated new states of knowledge using the five knowledge valuation methods described in Section 5.2. These values were then averaged to estimate the expected benefit of each permutation of number of samples tested and diffusion state of knowledge.

The Monte Carlo code used to perform these tests is included in Appendix F.

5.7 Chapter Summary

In the interests of clarity, the scope of our case study was limited to two parameters: corrosion and diffusion. These two parameters were used to define "alloy validity," a boolean variable that is true only if the total thickness of the product's outer, protective layer is greater than the thickness of product lost to diffusion and corrosion effects over the expected operational lifetime. The conditional probability relationships between the tracked parameters and experimental results, as well as a set of prior beliefs, were solicited from the principal researcher. Then, value was assigned to different probabilities of alloy validity. This input was used to perform an updating on the diffusion parameter and estimate the expected value of the tests on the corrosion parameter using both analytical and numerical uses of Bayes' Theorem.

Chapter 6

Results of the Case Study

There are two major updating results to report: the change in the distribution of D_o, and the change in the expected value of the corrosion testing following the updating of the D_o distribution.

6.1 Demonstration of Successful Updating - Diffusion

The first important result to note is the change in the estimate of the diffusion coefficient.

As can be seen in Figure 6.1, the updating was fairly minor (the prior and posterior distributions were not significantly different). The evidence from the diffusion testing had a limited effect on the estimate of D_o because the uncertainty associated with the evidence was large relative to the uncertainty on the prior belief, and so the mean of the posterior distribution remained close to that of the prior distribution.

The effect on the total probability of alloy viability was similarly small:

The total probability that the alloy meets the combined diffusion and corrosion requirements declined by only 1.01%, from 73.57% to 72.56%. This result is not particularly surprising— D_o is not the most significant factor affecting T_T, and the change on D_o was not dramatic.

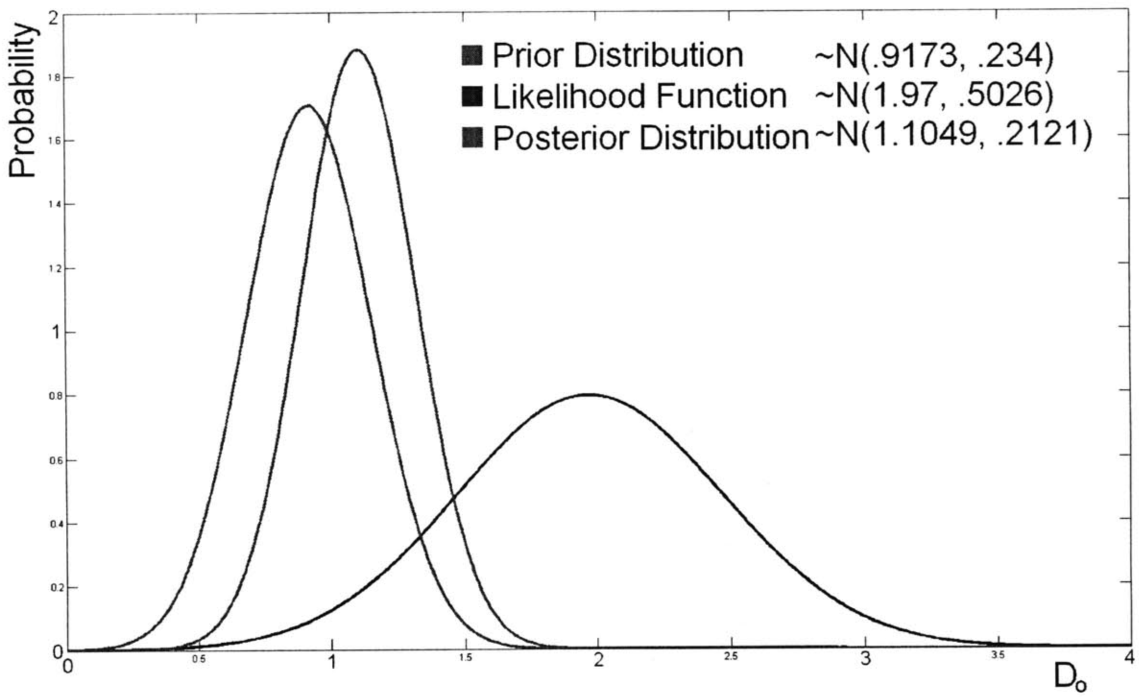


Figure 6-1: Prior and Posterior Distributions on the Diffusion Coefficient, D_o

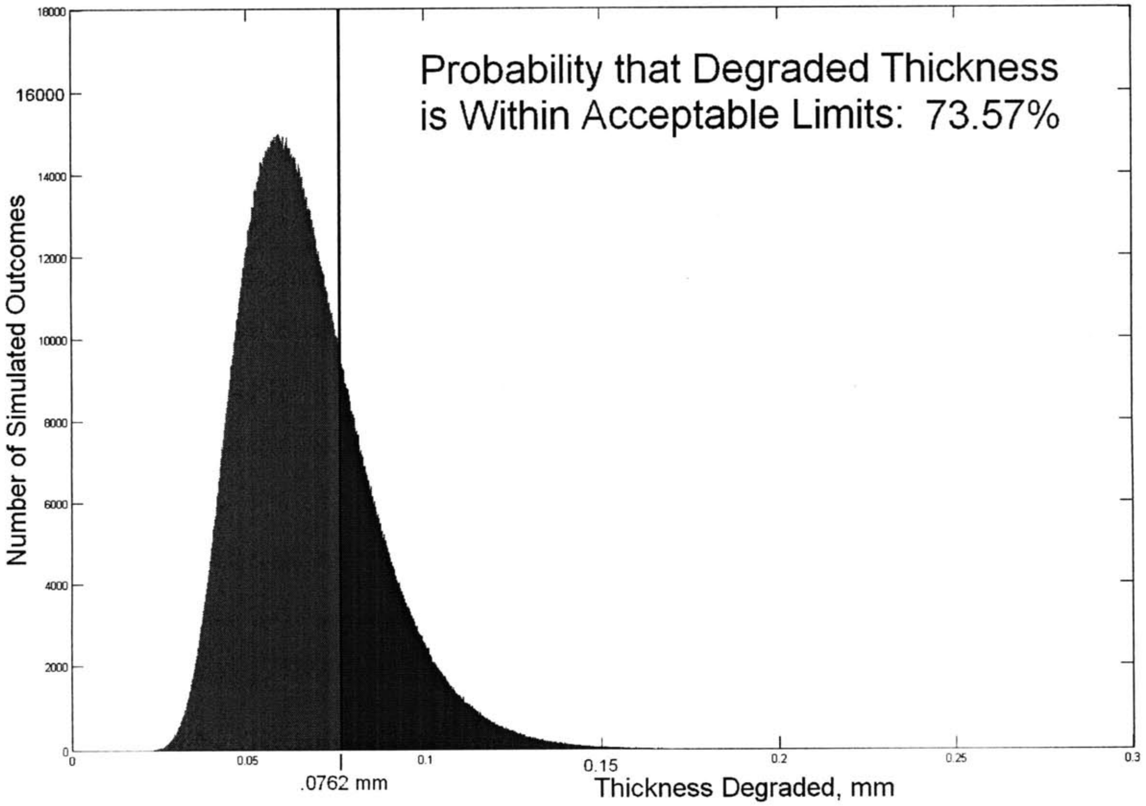


Figure 6-2: Histogram of Protective Layer Thickness Necessary to Ensure Clad Integrity, Prior Distribution, 10,000,000 simulations

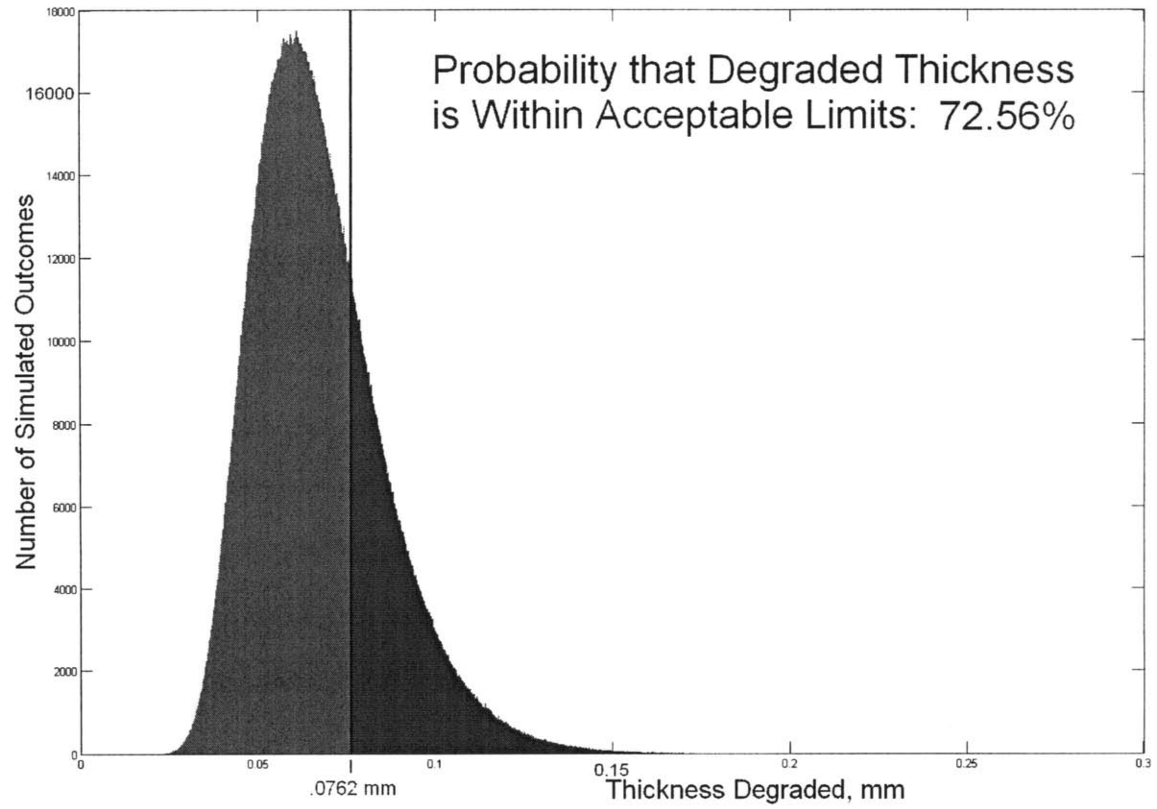


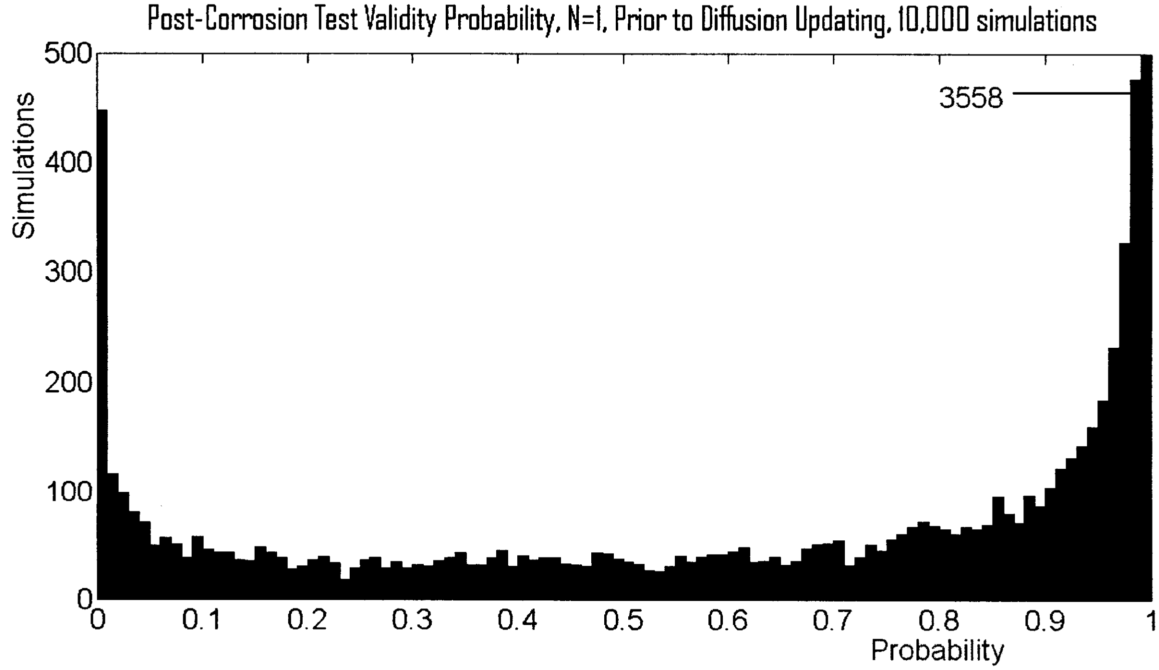
Figure 6-3: Histogram of Protective Layer Thickness Necessary to Ensure Clad Integrity, Posterior Distribution, 10,000,000 simulations

6.2 Demonstration of Successful Adaptation - Corrosion

Even though the evidence gathered from diffusion testing did not lead to a dramatic shift in expected success, it is still interesting to ask how the choice of experimental design or the decision of whether to continue with the experiment at all might change in response to this new evidence.

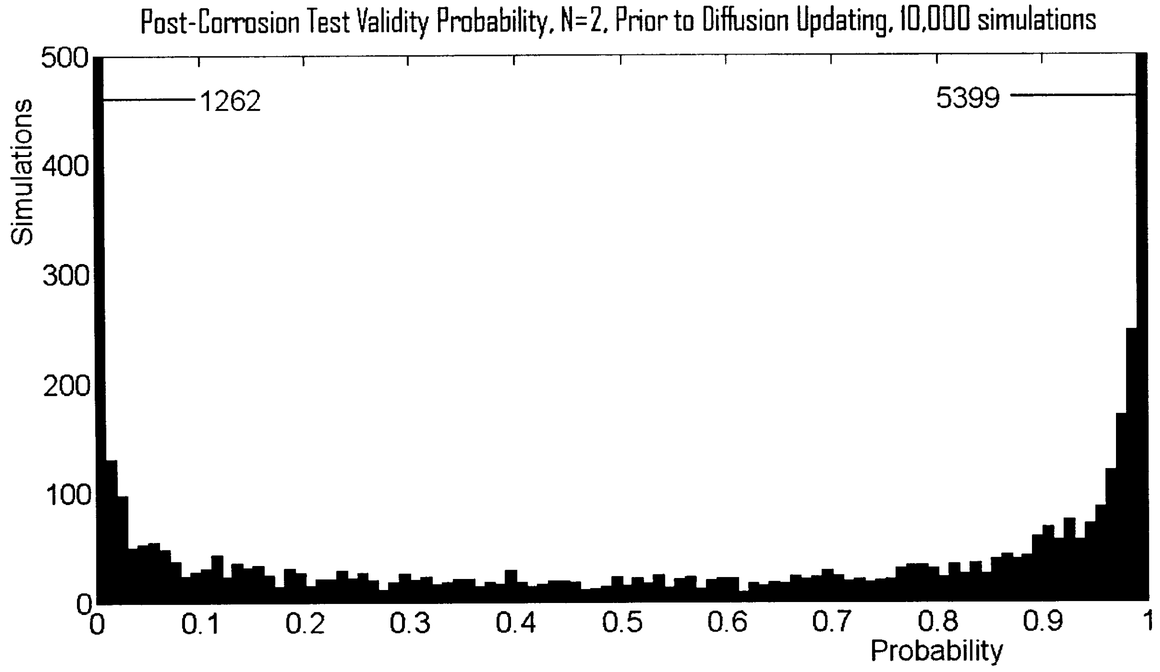
10,000 corrosion test results were simulated for each choice of n ranging from n=1 to n=10. For each of these 10,000 corrosion test results, the probability of alloy viability following an updating on the corrosion rate distribution was estimated using 10,000 random samples drawn from the posterior distribution of T_T.

Due to the precision of the tests used to determine the long-run corrosion rate, even experimental designs with only one sample were fairly conclusive in regards to the viability of the alloy. Prior to any hypothetical corrosion testing or diffusion updating, the probability of alloy viability was equal to 73.57%, a state of knowledge with zero value. As can be seen from the figure below, after testing a single corrosion sample for 1000-hr, the viability of the alloy would most likely be known to a much higher degree of certainty, with nearly 40% of the simulated cases having a probability of alloy viability within one percent of one (certainty that the alloy is viable) or zero (certainty that the alloy is inviable). Because the precision provided by each sample in the experiment is so high (the true corrosion rate lies within ± a factor of 1.25 of the sample value with 90% probability), we find that the consequences of corrosion updating are quite strong.



As n increases, the distribution of posterior validity probabilities concentrates even further in the extremes. This reflects a tightening of the estimate on C_{rate} around the parameter's true value and a reduction of the uncertainty in the alloy's viability.

For each of the Monte Carlo runs, it is possible to calculate the benefit of each of the 10,000 simulated states of knowledge and thus form an expectation on the benefits of the research. Using an 80% standard (which reflects a valuation where the cost of using an invalid alloy is 4x that of the benefit of using a valid alloy), the value

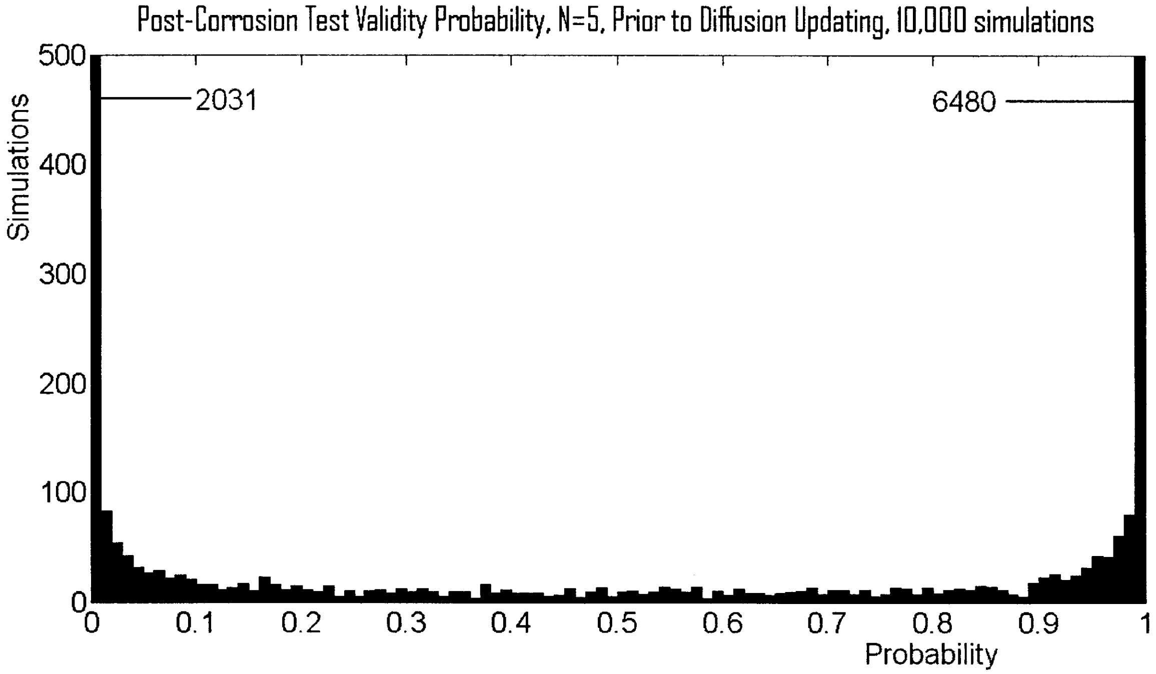


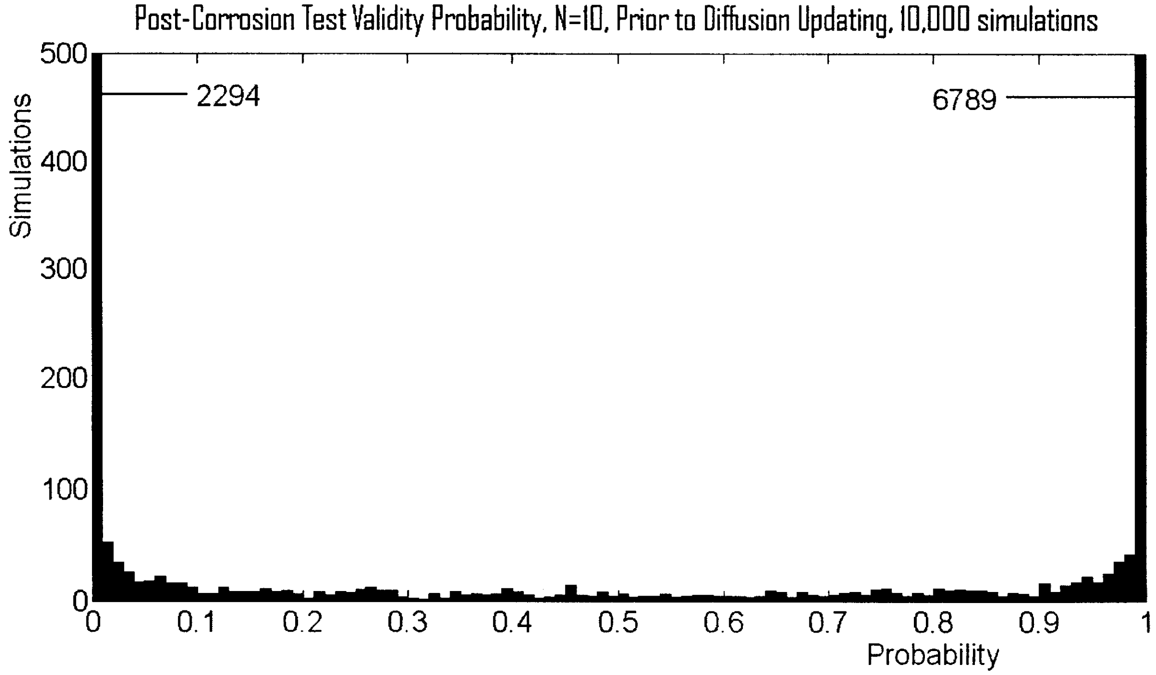
of the initial state of knowledge (Pr(Valid) = 73.57%) is zero— if the probability of alloy validity is less than 80% then it is advantageous not to use the alloy (an action whose value we have normalized to zero). Meanwhile, the expected value of the state of knowledge after a n=1 corrosion test is more than 50% of the benefit of a known valid alloy, and a two-sample experiment has an expected benefit of more than 60%. For each of the five proposed knowledge state valuation methods, the expected experiment benefits (prior to diffusion updating) are summarized in Table 6.1 as a fraction of B, the benefit of having a valid alloy with absolute certainty:

Table 6.1: Expected Value of Experiment as a Function of N and Benefit Structure (Before Diffusion Updating)

	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
80.0%	.5220	.6240	.6506	.6675	.6776	.6840	.6882	.6913	.6936	.6955
90.0%	.4593	.5919	.6328	.6539	.6666	.6748	.6801	.6838	.6867	.6892
95.0%	.4050	.5647	.6165	.6415	.6562	.6662	.6728	.6774	.6808	.6834
99.0%	.3041	.5097	.5831	.6168	.6367	.6493	.6585	.6650	.6694	.6728
99.9%	.2059	.4517	.5409	.5872	.6129	.6298	.6401	.6488	.6552	.6597

By taking the value difference between successive n's, these data can be used to


Post-Corrosion Test Validity Probability, N=5, Prior to Diffusion Updating, 10,000 simulations
Simulations
500
400
300
200
100
0
2031
6480
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Probability


Post-Corrosion Test Validity Probability, N=10, Prior to Diffusion Updating, 10,000 simulations
Simulations
500
400
300
200
100
0
2294
6789
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Probability

derive the marginal benefit of n, shown in Table 6.2

Table 6.2: Marginal Value of N for each Benefit Structure (Before Diffusion Updating)

	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
80.0%	.5220	.1020	.0266	.0169	.0101	.0064	.0042	.0031	.0023	.0019
90.0%	.4593	.1326	.0409	.0211	.0127	.0082	.0053	.0037	.0029	.0025
95.0%	.4050	.1597	.0518	.0250	.0147	.0100	.0066	.0046	.0034	.0026
99.0%	.3041	.2056	.0734	.0337	.0199	.0126	.0092	.0065	.0044	.0034
99.9%	.2059	.2458	.0892	.0463	.0257	.0169	.0103	.0087	.0064	.0045

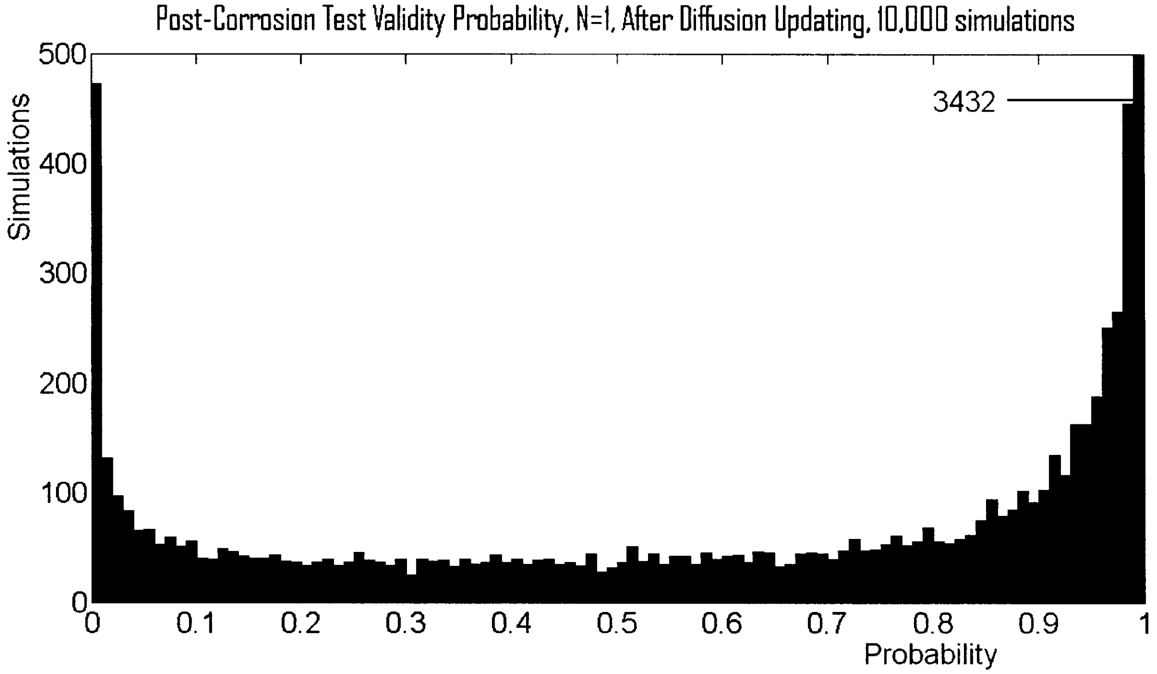
Using these data, one method of experimental design optimization would be to determine the marginal cost of adding samples to the corrosion test runs and set marginal benefit equal to marginal cost. In this manner, the total benefit of the corrosion test would be maximized.

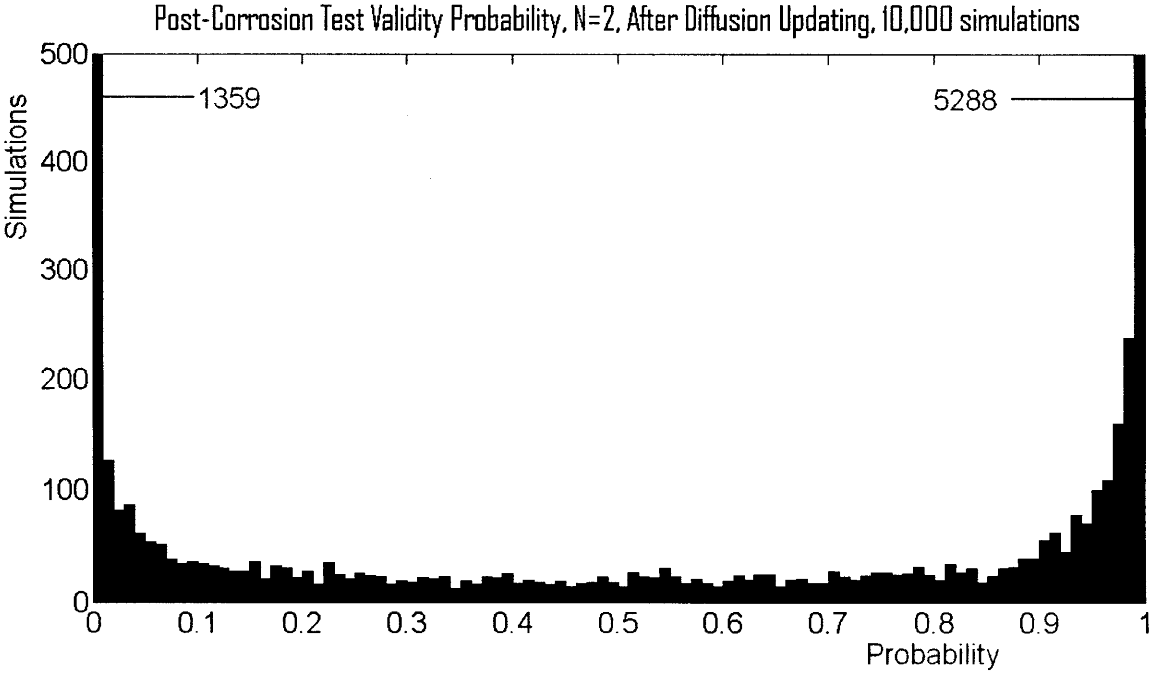
By comparing these results to those obtained using the new distribution on the diffusion coefficient, it is possible to measure how the expected benefit of research changes as a consequence of the new diffusion evidence. The expected and marginal value tables for the updated diffusion probability distribution are provided below along with histograms of the simulated viability probabilities for the n=1 and n=2 experimental designs following diffusion updating.

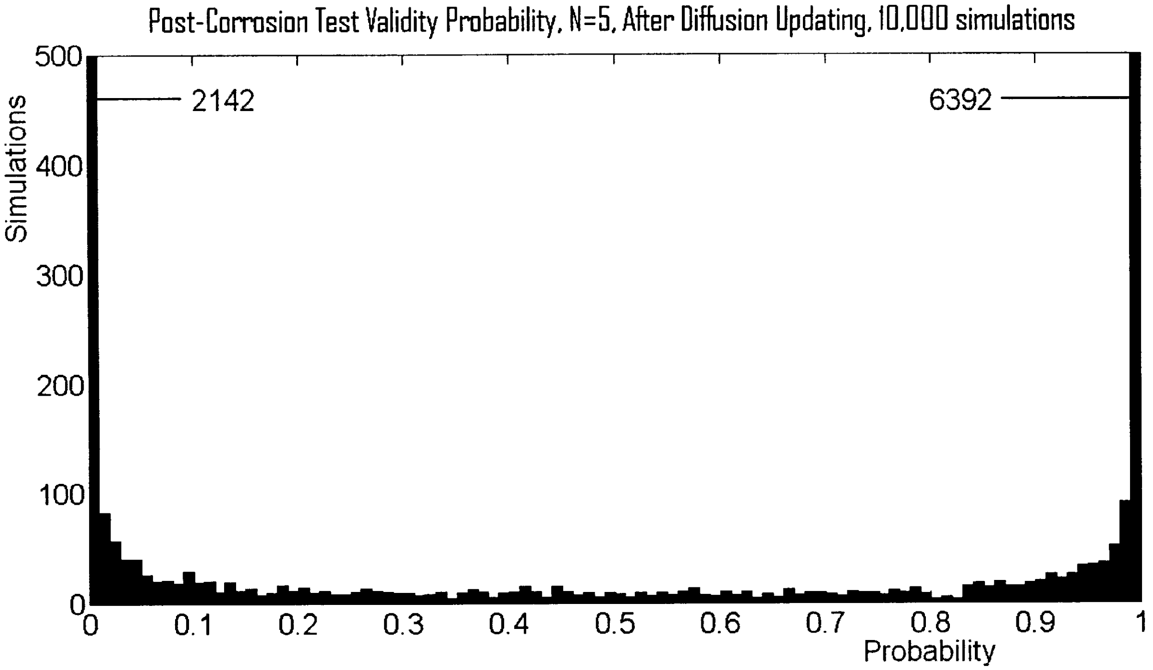
Table 6.3: Expected Value of Experiment as a Function of N and Benefit Structure (After Diffusion Updating)

	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
80.0%	.5058	.6077	.6413	.6585	.6691	.6757	.6801	.6832	.6857	.6878
90.0%	.4424	.5790	.6230	.6447	.6574	.6661	.6720	.6760	.6791	.6814
95.0%	.3886	.5522	.6064	.6324	.6475	.6575	.6648	.6697	.6733	.6762
99.0%	.2925	.4989	.5725	.6073	.6282	.6406	.6498	.6571	.6621	.6659
99.9%	.1950	.4377	.5314	.5776	.6040	.6208	.6323	.6403	.6470	.6526

Because the updated distribution on the diffusion coefficient reduces the probability of alloy viability, we see three effects: a decrease in the expected benefit for all of the experiment options under analysis, a decrease in the marginal benefit of the


Post-Corrosion Test Validity Probability, N=1, After Diffusion Updating, 10,000 simulations
Simulations
Probability
3432
0
100
200
300
400
500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Post-Corrosion Test Validity Probability, N=2, After Diffusion Updating, 10,000 simulations
Simulations
Probability
1359
5288
0
100
200
300
400
500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Post-Corrosion Test Validity Probability, N=5, After Diffusion Updating, 10,000 simulations
Simulations
500
400
300
200
100
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Probability
2142
6392

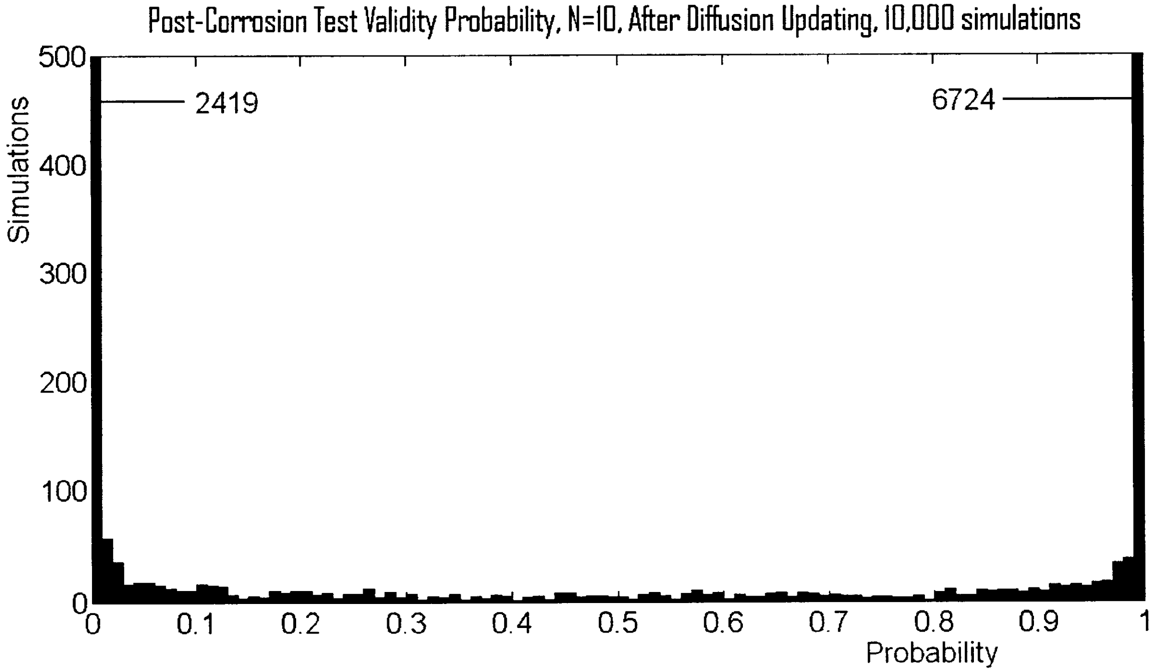

Post-Corrosion Test Validity Probability, N=10, After Diffusion Updating, 10,000 simulations
Simulations
500
400
300
200
100
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Probability
2419
6724

Table 6.4: Marginal Value of N for each Benefit Structure (After Diffusion Updating)

	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
80.0%	.5058	.1019	.0336	.0172	.0106	.0066	.0044	.0031	.0025	.0021
90.0%	.4424	.1366	.0440	.0217	.0127	.0087	.0059	.0040	.0031	.0023
95.0%	.3886	.1636	.0542	.0260	.0151	.0100	.0073	.0049	.0036	.0029
99.0%	.2925	.2064	.0736	.0348	.0209	.0124	.0092	.0073	.0050	.0038
99.9%	.1950	.2427	.0937	.0462	.0264	.0168	.0115	.0080	.0067	.0056

n=1 experimental design, and a slight increase in the marginal benefit of the higher n designs.

The first two effects are intuitive: because the value of research outcomes are tied to alloy validity and the probability of alloy validity has decreased, the benefit of further testing is expected to decrease. The third result, that the marginal benefit of higher n experimental designs would increase, is slightly non-intuitive. Although the probability of alloy validity has decreased due to the expected diffusion characteristics becoming less favorable, in the cases where the alloy *is* valid, a higher degree of certainty is needed on the estimate of the corrosion rate in order to achieve the confidence necessary to justify commercial use. In other words, the updating of diffusion had an unambiguously negative effect on the marginal benefit of running an experiment (the comparison between expected value at n=0 and n = some positive value), but it had an ambiguous effect on the marginal benefit of precision— because of the decrease in probability that the alloy is valid, higher precision tests are needed to guarantee an alloy's validity and thus realize benefits.

One interpretation is as follows: if we do not assume anything about the nature of the marginal cost of n but know that there must be an optimal n, we could say that the probability of optimality has increased for n=0 (cessation of further tests), decreased for all n>0, but decreased more strongly for n=1 such that relative to n=1, the probability of optimality of n>1 tests have increased.

Professor Ballinger, before the diffusion updating, intended to run a five-sample experiment. If his intention can be taken as revealed preference vis-a-vis the benefits and costs of the experiment, we could very well conclude that in the face of discouraging evidence from the diffusion tests, the optimal number of samples for Ballinger to test might actually *increase* slightly to n=6.

This finding should come with one caveat on generalization: it assumes that there is not an option to run samples sequentially at equal cost to running them concurrently. If there is the option of running one sample, observing the results, and then making a decision concerning whether to run another, without incurring a penalty in terms of cost, then this sequential process would be guaranteed to be preferrable to the alternatives, and thus the key finding from the diffusion updating would be the decreased marginal value of the n=1 option. The difference in value between the n=1 and n>1 experimental designs comes about because sometimes the n=1 results on a valid alloy are encouraging, but not enough to offer sufficient precision to yield a higher value state of knowledge, whereas a higher precision test would capture this benefit. If tests could be run sequentially, the results from a first sample could be used to better inform the decision to run a second sample, making n=1 the optimal choice, and making the decreased marginal value of n=1 the more relevant indicator of the experiment's change in favorability. Given the option of sequential test runs, the optimality of lower precision tests would actually improve relative to higher precision tests.

Lastly, it should be noted that the method used above to evaluate the benefit of varying sample sizes in a corrosion experiment could more generally be applied to a set of distinct experiments. In this manner, Bayesian updating could be used to reallocate funding across a range of unique research projects in response to new scientific information.

6.3 Chapter Summary

The diffusion updating was performed analytically and resulted in a slightly less favorable distribution on the tracked diffusion parameter. Because diffusion is a relatively minor contributor to alloy validity, and because the change in the distribution on the diffusion paramter was slight, the total probability of alloy validity fell only 1%, from

7̃3.6 to 72.6 percent. From the estimation of the value of corrosion testing, we found that due to the lower probability of alloy validity following the diffusion testing, the value of corrosion testing *fell* for both low and high precision corrosion testing. Furthermore, we found that the marginal value of precision actually *increased* slightly for higher precision tests even as it fell for lower ones. Overall, the Bayesian method worked quite well, and the case study showed a strong, generalizable way of using Bayesian decision analysis.

Chapter 7

Discussion and Conclusions

7.1 Case Study Results

The process of Bayesian updating proceeded smoothly, and the results from Monte Carlo assessment of expected research value suggest that belief networks have the potential to guide the allocation of research and development funds and improve the efficiency of R&D. We were able to track the most important parameters of the project, measure the change in expected benefit of the research project as new evidence came in, and make statistically informed suggestions on experimental design. The results were straight-forward and the methods used were generalizable to other research projects. All of these results lead to the conclusion that it is possible to use belief networks to develop measurements of research outcomes that would satisfy the OMB's wish for efficiency metrics in federal agency spending.

7.2 Problems Facing Implementation

During the thesis work, a few minor hurdles were encountered that deserve mention.

7.2.1 Researcher Overconfidence

One problem that became apparent during the work was researcher overconfidence. If the prior belief on the diffusion coefficient and the related likelihood function were correct, the probability of observing what we did— an estimate on D_o that was more than twice the expected value provided by the prior distribution— would have been vanishingly small. This suggests that the uncertainty on the diffusion coefficient could have been underestimated when prior beliefs were solicited.

Since there is only one instance to suggest overconfidence, it cannot be objectively claimed that overconfidence constitutes a proven trend among research project managers— however there is also *a priori* reason to believe that overconfidence may be a problem in the solicitation of priors. Psychology experiments have revealed that in many professional areas, experts display an "overconfidence effect" when estimating their uncertainty on a professional judgement[5].

The existence of overconfidence in solicited expert opinions is not necessarily a disadvantage for the use of belief networks— after all, if research project decision making suffers from the overconfidence effect when NOT using belief networks, it could easily be possible that belief networks would not operate under a disadvantage relative to their alternatives. Still, it is not encouraging to think that belief networks will, as part of their regular operation, import the cognitive biases that they designed to avoid.

There are a variety of options for expert solicitation, ranging from common practice to exotic, these may include peer review panels, Delphi methods, and prediction markets. The correct option may vary from agency to agency and depend upon the circumstances of the research (whether it involves classified or sensitive material, whether there is an appropriate pool of in-house expertise, etc). If belief networks are adopted as a means of measuring research efficiency and making funding allocation decisions, special care should be made to ensure that overconfidence is reduced where possible. There is future work to be done in developing methods of expert solicitation that correct for common biases such as overconfidence.

7.2.2 Principal-Agent Relationships

One conclusion from the case study is that prior beliefs matter. The relative certainty that was assigned to the prior belief on the diffusion coefficient had a significant impact on how the evidence affected the state of beliefs, and affected the extent to which the total probability of success was revised downward. In the case of extramural research, if researcher input is solicited to form high stakes belief networks, there is a risk of dishonest reporting and "gaming" of the system. As mentioned previously, it is difficult to correct for this principal-agent game due to the low risk tolerances of researchers relative to the financial stakes involved in R&D projects.

There are two potential solutions to this problem: the first is simply not to solicit the input of agents (the researchers) in forming the belief network. If peer review and other methods are sufficient, it may be easier to bypass the principal-agent problem rather than set up an incentive structure to correct for it.

The other potential solution is to apply an incentive structure that utilizes not just the results from a single project with a researcher, but instead makes use of the entire lifetime of experience with a researcher. Two methods are explored below through the use of a hypothetical example.

Let's use an example similar to that of Section 3.1.1. An agent, Angie, is applying for grant funding from a principal, Paul. Angie has a coin that is biased either towards head or tails such that the biased side comes up 75% of the time, but neither she nor Paul knows which side it is biased towards. Paul values coins that are biased towards heads and has no use for coins that are biased towards tails.

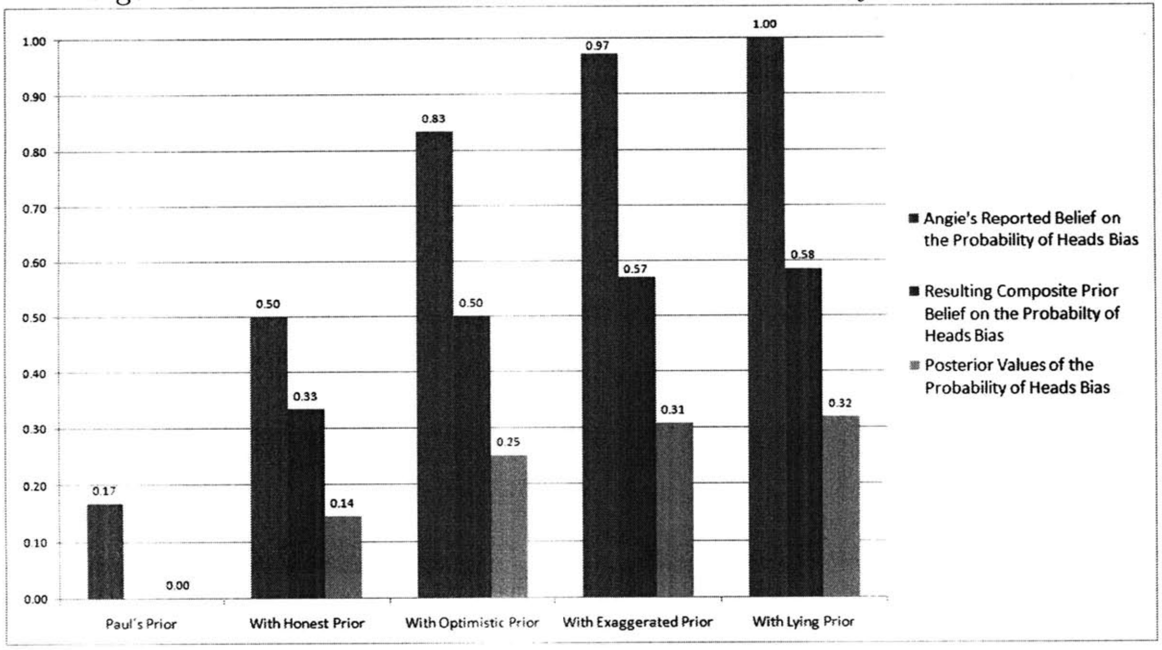
Both Paul and Angie study the coin and develop prior beliefs about the bias. Paul, after consulting with a trusted internal review board, believes that with $\frac{1}{6}$ probability the coin is biased towards heads, while Angie believes with $\frac{1}{2}$ probability the coin is biased towards heads (with complementary probabilities for bias towards tails). Paul, deciding that Angie's input is relevant and equally likely to be correct as his internal review board's, is willing to form the prior of the belief network by simply averaging the probability of heads bias that Paul believes with the probability given by Angie. Angie must now decide whether to submit her honest expectation (there is a $\frac{1}{2}$ probability the coin is biased towards heads), or one of three variations: an optimistic expectation ($\frac{5}{6}$ probability of heads), an exaggerated expectation ($\frac{35}{36}$ probability of heads) or outright lie (Angie proclaims with 100% certainty that the coin is heads-biased).

Whatever the expectation that the Angie submits, let's assume that Paul agrees to fund a study of the coin. The study consists of one coin flip, and it comes up tails. Paul must now decide whether to continue funding the study of this coin.

We can adopt a variety of interesting assumptions on benefits or cost: constant cost per coin flip, cost gets cut in half for each successive coin flip (i.e, for your first dollar you get one flip, for your second dollar you can determine with absolute certainty the bias of the coin), but the more relevant consideration this: on what grounds should Paul base his decision to continue or not continue the funding, and how should his experience change his future interactions with Angie?

First, let's look at the different priors and posteriors that form from the different beliefs that Angie can report.

Figure 7-1: Prior and Posterior Beliefs on the Probability of Heads Bias



In Figure 7-1 we see an expected result: as Angie's reported beliefs become more favorable towards a heads bias, both the prior and posterior values on the probability of the coin having a heads bias rise. As a consequence, Paul is more likely not only to begin funding Angie, but also to continue funding after negative results have been received. This is the heart of the principal-agent problem: Angie lacks an incentive to report accurate prior beliefs.

Paul lacks the ability to penalize Angie for misreporting her prior beliefs. There is no way to recover the grant money he has already given and no reasonable way to contract beforehand for making grant money contingent on favorable results. What Paul *can* do is deny future grant funding and discount future beliefs reported by Angie.

In Figure 7-2, the relative likelihood of Angie and Paul's reported beliefs is shown. What the figure depicts is the probability, given that Angie's reported beliefs are true, of witnessing the experimental results (a tails flip outcome), relative to the probability of witnessing a tails outcome given that Paul's beliefs are true. It is the ratio of the likelihood that Paul assigned to a tails outcome to the likelihood that Angie assigned to a tails outcome, normalized by the probability that either Paul or Angie's beliefs were true. In other words, the weightings shown in Figure 7-2 reflect a miniature Bayesian updating: Paul began by assigning equal probability to both his and Angie's reported beliefs. Having seen the outcome of the experiment, were Paul to re-solicit Angie's beliefs, he would not assign her beliefs a probability equal to his own, but instead would discount them as shown.

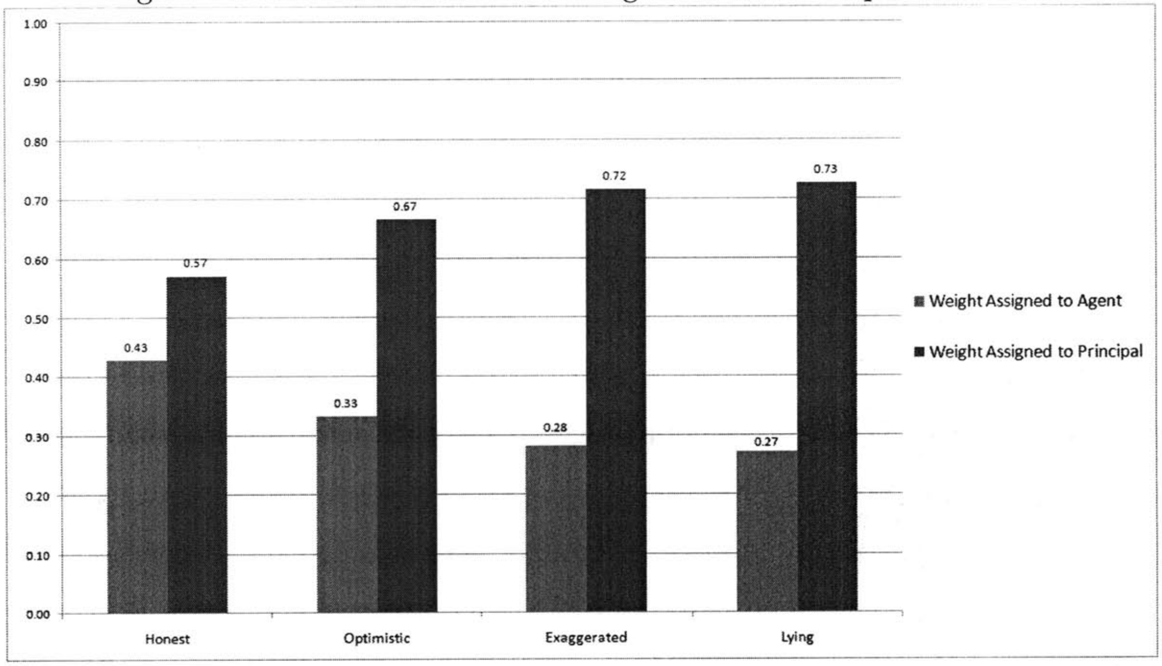
The obvious result is that the greater the extent to which Angie misreported her beliefs, the greater the extent to which her beliefs are discounted in future rounds.

Paul can use this Bayesian updating to penalize Angie in two ways: a punitive system and non-punitive system.

In a punitive system, Paul would penalize Angie by denying future grant funding. The greater the extent to which Angie's beliefs deviate from the evidence, the greater the extent to which funding would be denied in future rounds.

In a non-punitive system, Paul would not deny Angie future funding, but would discount her beliefs in future rounds by tracking the accuracy of Angie's reported

Figure 7-2: Relative Likelihood of Angie and Paul's Reported Beliefs



beliefs relative to those of his own and applying the appropriate weighting to future priors.

The disadvantage of the punitive system is that while it can be geared to make honest submission of priors the optimal choice for Angie, the penalty that is applied on Angie harms Paul as well. If Angie suggests a research project that Paul's internal review board concludes has a high benefit to cost ratio, Paul may be bound to not grant Angie money in order to apply a penalty to past performance.

The disadvantage of the non-punitive system is that it is impossible to reward honesty— the system merely curbs the effects of dishonesty (and inaccuracy in general). As a consequence, if a researcher does not care if they have no control over how the priors of their future research projects will formed, there is nothing to discourage them from reporting dishonest prior beliefs.

In both systems, the fundamental idea is to construct some sort of agent "reputation" that tracks past performance and modifies future grant allocation. The basic principle involved is that while it is difficult to properly solve the principal-agent problem within one research project, it may be possible to solve it over a broader time horizon where, by spreading the penalty assigned to inaccurate priors, agent risk aversion will be diminished as a factor.

7.2.3 Computational Limitations

Even for relatively limited experimental options, the computational requirements for analysis expand quickly. In this experiment we analyzed how one simple choice of experimental design, the number of samples to be tested in the corrosion experiment, was updated by results from other experiments. Had we looked at several different choices of experiments, each with their own experimental design choices, and analyzed the full range of sequential choices available (such as the choice to run Experiment 1 before running Experiment 2 and so on), the computational requirements would have been much greater. With only ten distinct experiments and the choice to run them in any given order, there would be nearly nine million different possible permutations, requiring several computer-centuries to analyze.

Belief networks are most effective when they are broad and encompassing— incomplete networks run the risk of overlooking important contributions to research value or missing connections between critical parameters. However, as belief networks grow, the computational needs to support them grow as well.

If belief networks are adopted on a large scale, it will be important to develop reliable heuristics in order to reduce their computational requirements.

To illustrate, suppose a simple example involving an alloy, Unobtainium, whose utility depends on three parameters of the alloy: X, Y, and Z. The three parameters can take values of either 0 or 1, and the alloy only has utility if all three parameters are equal to 1. Prior beliefs are solicited— the probability that X=1 is 10%, Y=1 is 50%, and Z=1 is 90%. The expected value of using the alloy is only positive if the probability of X=Y=Z=1 is greater than 90%. For each parameter, a researcher has a single experiment that he can perform that will determine the value of the parameter with absolute certainty.

We would like to know whether the researcher should perform an experiment (if the expected benefits of the experiment are greater than the expected costs) and of the experiments he can run, which one he should prefer to run first (the experiment with the highest ratio of expected benefits to expected costs).

If we construct a valuation function that looks at the value of only one action set, Use Alloy vs. Don't Use Alloy, we would find that the expected value of any single experiment is zero— none of the three experiments are sufficient, by themselves, to establish the degree of certainty necessary to give the Use Alloy action a positive expected value.

Instead, it is necessary to construct a valuation function that looks at not only the change in expected value of the Use/Dont Use Alloy action set, but also the Test/Dont Test action sets on X, Y, and Z. For example, the expected value of testing X would be equal to the highest valued action of the set of actions Test Y/Test Z/End Testing. In turn, the expected value of the actions Test Y and Test Z would be equal to the highest valued action of the set of actions Test Z/End Testing and Test Y/End Testing, respectively.

In other words, in order to determine the value of the research options, it would be necessary to evaluate the expected value of every single permutation of testing options.

In this example, it is apparent that if the cost of testing is identical for each parameter, then the researcher should test X, then Y, then Z. For a test cost, C, the expected cost of this testing order is 1.15C (in 90% of scenarios, the test for X will come back positive and only 1 C will be paid, in 5% of scenarios the test for X will come back positive but the test for Y will come back negative, and so only 2 C will be paid, and in the remaining 5% of scenarios, both X and Y will test positive and 3 C will be paid) and it is expected that Unobtanium will be proven useful in 4.5% of scenarios.

In more complicated examples, where evaluating the full set of possible permutations of actions would be unrealistic, it will be necessary to limit the range of permutations that are considered. A similar problem is faced in other machine-aided decision-making applications: for example, in computer chess playing software, where there may be thirty or more possible moves per position and a program might easily be expected to analyze forty to sixty moves deep, even a powerful computer with the capability to analyze one million positions per second would spend more than a year looking less than ten moves deep. Instead, to achieve greater depth of analysis, computer programs regularly prune options from the decision tree they are analyzing.

At the level of an individual research project, (such as the case study this thesis analyzed), little pruning is necessary to satisfactorily answer questions of interest. If the scope were expanded, say, from studying a single alloy to studying a range of alloys, or a range of lead reactor designs, or a range of energy technologies, considerably more thought would need to be given to the selection of outcomes to simulate and decisions to analyze.

7.3 Conclusions

Overall, the proof of concept demonstration was encouraging. A belief network was used to estimate and update the probability of experimental success in a way that enables valuation of research outcomes. In the process, the network was used to assign expectations of benefit to different experiments. Some work remains in formalizing how Bayesian decision making can be integrated into existing regulations and practices, and more effort is needed to address the computational challenge posed by large scale belief networks, however the success in applying belief networks to this case study demonstrates a potential for belief networks as not just a passive metric of research efficiency, but also as an active tool to be used in funding allocation and experiment design. Their use in research and development oversight seems warranted.

Appendix A

Program Assessment Rating Tool (PART) Questions

The PART Questionaire consists of 25 questions divided into four sections. Three additional questions are asked of research and development projects. The full text of each question, including the R&D specific questions, is provided below.

A.1 Section I, Program Purpose and Design

1.1: Is the program purpose clear?

1.2: Does the program address a specific and existing problem, interest, or need?

1.3: Is the program designed so that it is not redundant or duplicative of any other Federal, State, local or private effort?

1.4: Is the program design free of major flaws that would limit the programs effectiveness or efficiency?

1.5: Is the program design effectively targeted so that resources will address the programs purpose directly and will reach intended beneficiaries?

A.2 Section II, Strategic Planning

2.1: Does the program have a limited number of specific long-term performance measures that focus on outcomes and meaningfully reflect the purpose of the program?

2.2: Does the program have ambitious targets and timeframes for its long-term measures?

2.3: Does the program have a limited number of specific annual performance measures that can demonstrate progress toward achieving the programs long-term goals?

2.4: Does the program have baselines and ambitious targets for its annual measures?

2.5: Do all partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) commit to and work toward the annual and/or long-term goals of the program?

2.6: Are independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need?

2.7: Are budget requests explicitly tied to accomplishment of the annual and long-term performance goals, and are the resource needs presented in a complete and transparent manner in the programs budget?

2.8: Has the program taken meaningful steps to correct its strategic planning deficiencies?

2.RD1: If applicable, does the program assess and compare the potential benefits of efforts within the program and (if relevant) to other efforts in other programs that have similar goals?

2.RD2: Does the program use a prioritization process to guide budget requests and funding decisions?

A.3 Section III, Program Management

3.1: Does the agency regularly collect timely and credible performance information, including information from key program partners, and use it to manage the program and improve performance?

3.2: Are Federal managers and program partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) held accountable for cost, schedule and performance results?

3.3: Are funds (Federal and partners) obligated in a timely manner, spent for the intended purpose, and accurately reported?

3.4: Does the program have procedures (e.g., competitive sourcing/cost comparisons, IT improvements, appropriate incentives) to measure and achieve efficiencies and cost effectiveness in program execution?

3.5: Does the program collaborate and coordinate effectively with related programs?

3.6: Does the program use strong financial management practices?

3.7: Has the program taken meaningful steps to address its management deficiencies?

3.RD1: For R&D programs other than competitive grants programs, does the program allocate funds and use management processes that maintain program quality?

A.4 Section IV, Program Results/Accountability

4.1: Has the program demonstrated adequate progress in achieving its long-term performance goals?

4.2: Does the program (including program partners) achieve its annual performance goals?

4.3: Does the program demonstrate improved efficiencies or cost effectiveness in achieving program goals each year?

4.4: Does the performance of this program compare favorably to other programs, including government, private, etc., with similar purpose and goals?

4.5: Do independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results?

Appendix B

Derivation of Bayes' Theorem from the Definition of Conditional Probability

Bayes Theorem can be developed from the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{B.1}$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{B.2}$$

Combining these two equations yields

$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A) \tag{B.3}$$

Dividing each side by Pr(B) then yields Bayes' Theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)} \tag{B.4}$$

Appendix C

Research and Development Program Investment Criteria

[1] In 2007, the Office of Management and Budget issued a memorandum to federal agencies concerning compliance with PART requirements. Included in the memorandum was guidance to federal agencies on the factors to be considered when assessing the investment quality of research and development projects. The text from the memorandum devoted to R&D is given below.

As another initiative of the Presidents Management Agenda, the development of explicit R&D investment criteria builds on the best of the planning and assessment practices that R&D program managers use to plan and assess their programs. The Administration has worked with experts and stakeholders to build upon lessons learned from previous approaches.

Agencies should use the criteria as broad guidelines that apply at all levels of Federally funded R&D efforts, and they should use the PART as the instrument to periodically evaluate compliance with the criteria at the program level. To make this possible, the R&D PART aligns with the R&D criteria. The R&D criteria are reprinted here as a guiding framework for addressing the R&D PART.

The R&D criteria address not only planning, management, and prospective assess-

[1] The Research and Development Program Investment Criteria is drawn from the website of the USDA [22]

ment but also retrospective assessment. Retrospective review of whether investments were well-directed, efficient, and productive is essential for validating program design and instilling confidence that future investments will be wisely invested. Retrospective reviews should address continuing program relevance, quality, and successful performance to date.

While the criteria are intended to apply to all types of R&D, the Administration is aware that predicting and assessing the outcomes of *basic* research in particular is never easy. Serendipitous results are often the most interesting and ultimately may have the most value. Taking risks and working toward difficult-to-attain goals are important aspects of good research management, and innovation and breakthroughs are among the results. However, there is no inherent conflict between these facts and a call for clearer information about program goals and performance toward achieving those goals. The Administration expects agencies to focus on improving the management of their research programs and adopting effective practices, and not on predicting the unpredictable.

The R&D investment criteria have several potential benefits:

- Use of the criteria allows policy makers to make decisions about programs based on information beyond anecdotes, prior-year funding levels, and lobbying of special interests.

- A dedicated effort to improve the process for budgeting, selecting, and managing R&D programs is helping to increase the return on taxpayer investment and the productivity of the Federal R&D portfolio.

- The R&D investment criteria will help communicate the Administrations expectations for proper program management.

- The criteria and subsequent implementation guidance will also set standards for information to be provided in program plans and budget justifications.

- The processes and collected information promoted under the criteria will improve public understanding of the possible benefits and effectiveness of the Federal investment in R&D.

Details on the Criteria

The Relevance, Quality, and Performance criteria apply to all R&D programs. Industry— or market-relevant applied R&D must meet additional criteria. Together, these criteria can be used to assess the need, relevance, appropriateness, quality, and performance of Federal R&D programs.

I. Relevance

R&D investments must have clear plans, must be relevant to national priorities, agency missions, relevant fields, and "customer" needs, and must justify their claim on taxpayer resources. Programs that directly support Presidential priorities may receive special consideration with adequate documentation of their relevance. Review committees should assess program objectives and goals on their relevance to national needs, customer needs, agency missions, and the field(s) of study the program strives to address. For example, the Joint DOE/NSF Nuclear Sciences Advisory Committees Long Range Plan and the Astronomy Decadal Surveys are the products of good planning processes because they articulate goals and priorities for research opportunities within and across their respective fields.

OMB will work with some programs to identify quantitative metrics to estimate and compare potential benefits across programs with similar goals. Such comparisons may be within an agency or among agencies.

A. Programs must have complete plans, with clear goals and priorities.

Programs must provide complete plans, which include explicit statements of:

- specific issues motivating the program;
- broad goals and more specific tasks meant to address the issues;
- priorities among goals and activities within the program;
- human and capital resources anticipated; and
- intended program outcomes, against which success may later be assessed.

B. Programs must articulate the potential public benefits of the program.

Programs must identify potential benefits, including added benefits beyond those of any similar efforts that have been or are being funded by the government or others. R&D benefits may include technologies and methods that could provide new options in the future, if the landscape of todays needs and capabilities changes dramatically. Some programs and sub-program units may be required to quantitatively estimate expected benefits, which would include metrics to permit meaningful comparisons among programs that promise similar benefits. While all programs should try to articulate potential benefits, OMB and OSTP recognize the difficulty in predicting the outcomes of basic research. Consequently, agencies may be allowed to relax this as a requirement for basic research programs.

C. Programs must document their relevance to specific Presidential priorities to receive special consideration.

Many areas of research warrant some level of Federal funding. Nonetheless, the President has identified a few specific areas of research that are particularly important. To the extent a proposed project can document how it directly addresses one of these areas, it may be given preferential treatment.

D. Program relevance to the needs of the Nation, of fields of Science & Technology, and of program "customers" must be assessed through prospective external review.

Programs must be assessed on their relevance to agency missions, fields of science or technology, or other customer needs. A customer may be another program at the same or another agency, an interagency initiative or partnership, or a firm or other organization from another sector or country. As appropriate, programs must define a plan for regular reviews by primary customers of the programs relevance to their needs. These programs must provide a plan for addressing the conclusions of external reviews.

E. Program relevance to the needs of the Nation, of fields of S&T, and of program "customers" must be assessed periodically through retrospective

external review.

Programs must periodically assess the need for the program and its relevance to customers against the original justifications. Programs must provide a plan for addressing the conclusions of external reviews.

II. Quality

Programs should maximize the quality of the R&D they fund through the use of a clearly stated, defensible method for awarding a significant majority of their funding. A customary method for promoting R&D quality is the use of a competitive, merit-based process. NSFs process for the peer-reviewed, competitive award of its R&D grants is a good example. Justifications for processes other than competitive merit review may include "outside-the-box" thinking, a need for timeliness (e.g., R&D grants for rapid response studies of *Pfisteria*), unique skills or facilities, or a proven record of outstanding performance (e.g., performance-based renewals).

Programs must assess and report on the quality of current and past R&D. For example, NSFs use of Committees of Visitors, which review NSF directorates, is an example of a good quality assessment tool. OMB and OSTP encourage agencies to provide the means by which their programs may be benchmarked internationally or across agencies, which provides one indicator of program quality.

A. Programs allocating funds through means other than a competitive, merit-based process must justify funding methods and document how quality is maintained.

Programs must clearly describe how much of the requested funding will be broadly competitive based on merit, providing compelling justifications for R&D funding allocated through other means. (See OMB Circular A-11 for definitions of competitive merit review and other means of allocating Federal research funding.) All program funds allocated through means other than unlimited competition must document the processes they will use to distribute funds to each type of R&D performer (e.g., Federal laboratories, Federally-funded R&D centers, universities, etc.). Programs are encouraged to use external assessment of the methods they use to allocate R&D and maintain program quality.

B. Program quality must be assessed periodically through retrospective expert review.

Programs must institute a plan for regular, external reviews of the quality of the program's research and research performers, including a plan to use the results from these reviews to guide future program decisions. Rolling reviews performed every 3-5 years by advisory committees can satisfy this requirement. Benchmarking of scientific leadership and other factors provides an effective means of assessing program quality relative to other programs, other agencies, and other countries.

III. Performance

R&D programs should maintain a set of high priority, multi-year R&D objectives with annual performance outputs and milestones that show how one or more outcomes will be reached. Metrics should be defined not only to encourage individual program performance but also to promote, as appropriate, broader goals, such as innovation, cooperation, education, and dissemination of knowledge, applications, or tools.

OMB encourages agencies to make the processes they use to satisfy the Government Performance and Results Act (GRPA) consistent with the goals and metrics they use to satisfy these R&D criteria. Satisfying the R&D performance criteria for a given program should serve to set and evaluate R&D performance goals for the purposes of GPRA. OMB expects goals and performance measures that satisfy the R&D criteria to be reflected in agency performance plans. Programs must demonstrate an ability to manage in a manner that produces identifiable results. At the same time, taking risks and working toward difficult-to-attain goals are important aspects of good research management, especially for basic research. The intent of the investment criteria is not to drive basic research programs to pursue less risky research that has a greater chance of success. Instead, the Administration will focus on improving the management of basic research programs.

OMB will work with some programs to identify quantitative metrics to compare performance across programs with similar goals. Such comparisons may be within an agency or among agencies.

Construction projects and facility operations will require additional performance metrics. Cost and schedule earned-value metrics for the construction of R&D facilities must be tracked and reported. Within DOE, the Office of Sciences formalized independent reviews of technical cost, scope, and schedule baselines and project management of construction projects ("Lehman Reviews") are widely recognized as an effective practice for discovering and correcting problems involved with complex, one-of-a-kind construction projects.

A. Programs may be required to track and report relevant program inputs annually.

Programs may be expected to report relevant program inputs, which could include statistics on overhead, intramural/extramural spending, infrastructure, and human capital. These inputs should be discussed with OMB.

B. Programs must define appropriate output and outcome measures, schedules, and decision points.

Programs must provide single- and multi-year R&D objectives, with annual performance outputs, to track how the program will improve scientific understanding and its application. Programs must provide schedules with annual milestones for future competitions, decisions, and termination points, highlighting changes from previous schedules. Program proposals must define what would be a minimally effective program and a successful program. Agencies should define appropriate output and outcome measures for all R&D programs, but agencies should not expect fundamental basic research to be able to identify outcomes and measure performance in the same way that applied research or development are able to. Highlighting the results of basic research is important, but it should not come at the expense of risk-taking and innovation. For some basic research programs, OMB may accept the use of qualitative outcome measures and quantitative process metrics. Facilities programs must define metrics and methods (e.g., earned-value reporting) to track development costs and to assess the use and needs of operational facilities over time. If leadership in a particular field is a goal for a program or agency, OMB and OSTP encourage the use of benchmarks to assess the processes and outcomes of the program with respect to leadership. OMB encourages agencies to make the processes they use to satisfy

GPRA consistent with the goals and metrics they use to satisfy these R&D criteria.

C. Program performance must be retrospectively documented annually.

Programs must document performance against previously defined output and outcome metrics, including progress toward objectives, decisions, and termination points or other transitions. Programs with similar goals may be compared on the basis of their performance. OMB will work with agencies to identify such programs and appropriate metrics to enable such comparisons.

IV. Criteria for R&D Programs Developing Technologies That Address Industry

The purpose of some R&D and technology demonstration programs and projects is to introduce some product or concept into the marketplace. However, some of these efforts engage in activities that industry is capable of doing and may discourage or even displace industry investment that would occur otherwise. For the purposes of assessing Federal R&D investments, the following criteria should be used to assess industry-relevant R&D and demonstration projects, including, at OMB discretion, associated construction activities.

OMB will work with programs to identify quantitative metrics to measure and compare potential benefits and performance across programs with similar goals, as well as ways to assess market relevance.

A. Programs and projects must articulate public benefits of the program using uniform benefit indicators across programs and projects with similar goals.

In addition to the public benefits required in the general criteria, *all* industry-relevant programs and projects must identify and use uniform benefit indicators (including benefit-cost ratios) to enable comparisons of expected benefits across programs and projects. OMB will work with agencies to identify these indicators.

B. Programs and projects must justify the appropriateness of Federal investment, including the manner in which the market fails to motivate private sector investment.

A lack of market incentives discourages private firms from investing in research where the benefits may occur far in the future, the risks may be too great for non-

Federal participants, or the benefits accrue to the public rather than private investors. Programs and projects must demonstrate that industry investment is sub-optimal and explain in what way the market fails that prevents the private sector from capturing the benefits of developing the good or service.

C. Programs and projects must demonstrate that investment in R&D and demonstration activities is the best means to support the Federal policy goals, compared to other policy alternatives.

When the Federal government chooses to intervene to address market failures, there may be many policy alternatives to address those failures. Among the other tools available to the government are legislation, tax policy, regulatory and enforcement efforts, and an integrated combination of these approaches. In this context, projects to address issues of genuine Federal concern should be able to illustrate how R&D and demonstration activities are superior to other policy tools in addressing Federal goals, either by themselves or as part of an integrated package.

D. Programs and projects must document industry or market relevance, including readiness of the market to adopt technologies or other outputs.

Programs must assess the likelihood that the target industry will be able to adopt the technology or other program outputs. The level of industry cost sharing is one indicator of industry relevance. Before projects move into demonstration or deployment stages, an economic analysis of the public and private returns on the public investment must be provided.

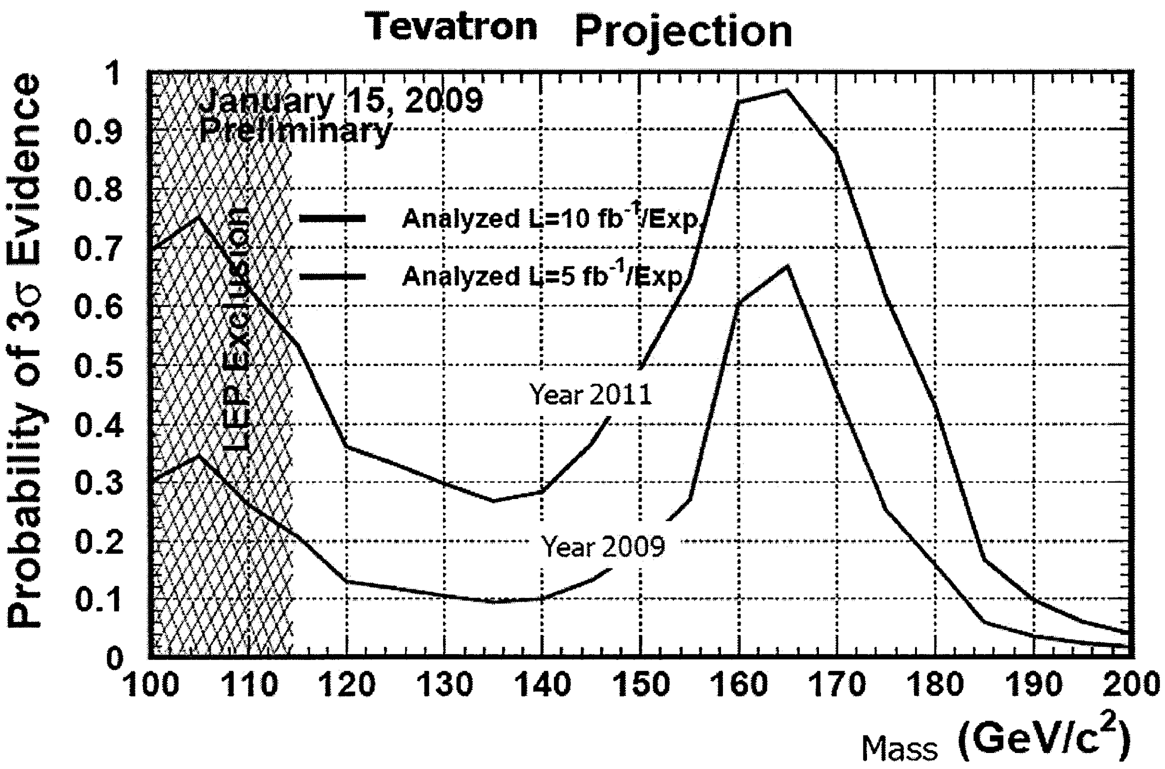
E. Program performance plans and reports must include off ramps and transition points.

In addition to the schedules and decision points defined in the general criteria, program plans should also identify whether, when, and how aspects of the program may be shifted to the private sector.

Appendix D

Estimated Probability of Discovering the Higgs Boson

The following figure is drawn from a presentation given by Dmitri Denisov, a lead researcher at the Fermi National Laboratory:



[4]

Figure D-1: Estimated Probability of Discovering the Higgs Boson, Given Boson Mass

Appendix E

Description of the Static Corrosion Test Facility at MIT

The following is drawn from the quarterly reports provided to the Nuclear Energy Research Initiative by the principal researcher of Project 06-038, Professor Ronald Ballinger.

The static corrosion test system consists of three subsystems that can be described as follows:

Corrosion cell system (2): Each cell consists of a high temperature furnace equipped with a controller and a rotating feed-through that allows for the rotation of the sample if required. Each cell can be operated in the non-rotating ("static") mode simple immersion tests. However, the test cells are not designed to simulate a loop type system. The maximum temperature for test is 1000°C and the maximum rotational velocity is to 2 m/s for a 100mm diameter disk sample. A vacuum-tight outer boundary is achieved by a stainless steel vessel whose joints are welded and leak-tested using helium gas leak detector at a maximum rate of 10-7 cc/sec. A ceramic, 4 liter, crucible, ZrO_2, is used as the container for the molten alloy. Thin, disc type specimens can be attached to a Molybdenum rod to be immersed in the melt. To prevent a galvanic couple between the shaft rod and specimen material, ceramic washers are inserted between those joints. Immersion tests can also be accomplished via the use of individual crucibles for each sample set.

Oxygen probe: Oxygen potential measurement is achieved using a solid electrolyte and reference electrode. A Bi/Bi_2O_3 reference electrode is used to measure the oxygen concentration in the molten lead. The EMF signal of probe is converted into oxygen concentration using the Nernst equation. The current oxygen probe is the product of a long collaborative development effort between MIT and LANL.

Oxygen control/monitoring system: Oxygen control is achieved by means of the thermodynamic equilibrium of H_2O-H_2-O_2. With this technique a very low oxygen partial pressure can be obtained. With the MIT system the main variable is the pressure ratio of H_2/H_2O controlled by hydrogen gas flow and moisture intake by bubbling. Pre-purified Argon is used as a carrier gas and mixed with hydrogen (also pre-purified) below the flammable point by mass flow controller. An oxygen gettering furnace is used to eliminate impurity oxygen still remaining at ppm levels in the argon. Hydrogen also passes through a purifier and drier before being mixed with argon. Gas flow passing through the constant water bath contains moisture up to several thousand ppm to the corrosion cell system. Several measurement units for moisture and gas phase oxygen are installed in order to track the changes in the mixture concentration.

In addition to the rigorous oxygen control system, each test contains several samples of pure metals (Fe or Cr) that are exposed $\frac{1}{2}$ in the gas phase above the molten metal and $\frac{1}{2}$ in the molten metal. After each test these samples are examined for the presence of an oxide layer which would indicate the presence of oxygen above a certain threshold. The absence of an oxide layer, conversely, is an indicator that the oxygen potential is below a certain potential.

Appendix F

Monte Carlo Source Code

The Monte Carlo code was executed using the the MATLAB numerical computing environment. The value of n and hyperparameters used to generate D_o were changed between runs as necessary.

```
   m = 0.788546;
   sigmaone = .30125;
   sigmatwo = .13561;
   lengthone = 10000;
   lengthtwo = 10000;
   counter=0;
   n = 1;
   record = zeros(lengthone, 1);
   TotalT = 0;
   musample = 0;
   C_Rate= 0;
   Q = 0;
   for i=1:lengthone
   musample = lognrnd(m, sigmaone);
   xbarsample = log(lognrnd(log(musample), sigmatwo/sqrt(n)));
   mprime = ((m /(sigmaone^2) + n^2 * xbarsample / (sigmatwo^2))/(1/sigmaone^2
+ n^2 / sigmatwo^2));
   sigmaprime = (1/(sigmaone^2) + n^2 / (sigmatwo^2))^(-.5);
   for j=1:lengthtwo
      D_o = normrnd(.91728, .234);
      Q = normrnd(52.45, .12755);
      D_Si = D_o * exp(-1000*Gibbs/(1.9858775*923.15));
      if D_Si0
         D_Si=0;
```

```
    end
    C_Rate = lognrnd(mprime, sigmaprime);
    TotalT = C_Rate * .02628 + 10786.0827 * sqrt(D_Si);
    if TotalT  .0762
        counter = counter+1;
    end
end
record(i) = counter / lengthtwo;
counter = 0;
end
save data01prior record
dataarray = zeros(20,5);
load data01prior.mat
valueeighty = 0;
valueninety = 0;
valueninetyfive = 0;
valueninetynine = 0;
valueninenineninenine = 0;
for i=1:10000
if record(i)  0.999
    valueninenineninenine = (valueninenineninenine + .1*(record(i)-0.999));
end
end
for i=1:10000
if record(i)  0.99
    valueninetynine = (valueninetynine + .01*(record(i)-0.99));
end
end
for i=1:10000
if record(i)  0.95
    valueninetyfive = (valueninetyfive + .002*(record(i)-0.95));
end
end
for i=1:10000
if record(i)  0.9
    valueninety = (valueninety + .001*(record(i)-0.9));
end
end
for i=1:10000
if record(i)  0.8
    valueeighty = (valueeighty + .0005*(record(i)-0.8));
end
end
dataarray(1,1)=valueeighty;
dataarray(1,2)=valueninety;
dataarray(1,3)=valueninetyfive;
dataarray(1,4)=valueninetynine;
dataarray(1,5)=valuenineninenine;
```

Bibliography

[1] Michael P. Short Sean Morton R. G. Ballinger. Diffusional stability of a ferritic-martensitic steel composite for service in advanced lead-bismuth cooled nuclear reactors, January 2009.

[2] R. J. Borg and D. Y. F. Lai. Diffusion in α-fe-si alloys. *Journal of Applied Physics*, 41(13), December 1970.

[3] National Research Council. Evaluating federal research programs: Research and the government performance and results act. Print, Washington, DC: The National Academies Press, January 1999.

[4] Dmitri Denisov. Closing in on the higgs with tevatron data. In *High-Energy Physics Discoveries: From the Tevatron to the Large Hadron Collider*, AAAS Annual Meeting, Chicago, February 2009. AAAS.

[5] Sarah Lichtenstein Baruch Fischoff and Lawrence D. Phillips. Calibration of probabilities: The state of the art to 1980, January 1982. Cambridge University Press.

[6] Organization for Economic Cooperation and Development. Main science and technology indicators. online, http://www.oecd.org/dataoecd/9/44/41850733.pdf, August 2008.

[7] American Association for the Advancement of Science. Nih by funding mechanism. online, http://www.aaas.org/spp/rd/09ptbii10.pdf, February 2008.

[8] American Association for the Advancement of Science. R&d as percent of the federal budget, fy 1962-2009. online, http://www.aaas.org/spp/rd/rdbdg09p.pdf, February 2008.

[9] American Association for the Advancement of Science. R&d in the department of defense. online, http://www.aaas.org/spp/rd/09ptbii2.pdf, February 2008.

[10] American Association for the Advancement of Science. Research funding by character of work. online, http://www.aaas.org/spp/rd/trcha09p.pdf, March 2008.

[11] American Association for the Advancement of Science. Total r&d by agency, fy 1976-2009. online, http://www.aaas.org/spp/rd/hist09p2.pdf, March 2008.

[12] American Association for the Advancement of Science. Trends in federal r&d as % of gdp, fy 1976-2009. online, http://www.aaas.org/spp/rd/tbrdgdp09.pdf, March 2008.

[13] Eliezer Geisler. *The Metrics of Science and Technology.* Greenwood Publishing Group, 2000.

[14] F. J. Martin L Soler F Hernandez D. Gomez-Briceno. Oxide layer stability in lead-bismuth eutectic up to 600°c. *Journal of Nuclear Materials*, 335(2), November 2004.

[15] Jeongyoun Lim. *Effects of Chromium and Silicon on Corrosion of Iron Alloys in Lead Bismuth Eutectic.* PhD dissertation, Massachusetts Institute of Technology, Department of Nuclear Science and Engineering, February 2006.

[16] Paul Milgrom and John Roberts. *Economics, Organization and Management*, chapter 7. Prentice-Hall, February 1992.

[17] Division of Science Resources Statistics National Science Foundation. National patterns of r&d resources: 2007 data update. online, http://www.nsf.gov/statistics/nsf08318/, November 2008.

[18] Office of Management and Budget. Part frequently asked questions. http://www.whitehouse.gov/omb/part/2004_faq.html, January 2004.

[19] Office of Management and Budget. Rating the performance of federal programs; the budget for fiscal year 2004. online, http://www.gpoaccess.gov/usbudget/fy04/pdf/budget/performance.pdf, March 2004.

[20] Office of Management and Budget. Program assessment rating tool guidance ' no. 2007-02. http://www.whitehouse.gov/omb/assets/omb/part/fy2007/2007_guidance_final.pdf, January 2007.

[21] Office of Management and Budget. Rating the performance of federal programs; the budget for fiscal year 2004. online, http://www.whitehouse.gov/omb/expectmore/, March 2007.

[22] Office of Management and Budget. Research and development program investment criteria, pp 73-78 in guide to the program assessment rating tool (part). online, http://www.csrees.usda.gov/business/reporting/part/r_d_invest_criteria.pdf, January 2007.

[23] Office of Management and Budget. Omb circular no, a-11, section 84. online, http://www.whitehouse.gov/omb/circulars/a11/current_year/s84.pdf, June 2008.

[24] National Research Council of the National Academies. Evaluating research efficiency in the u.s. environmental protection agency. The National Academies Press, Washington D.C., February 2008.

[25] G. Guntz M. Julien G. Kottmann F. Pellicani A. Pouilly and J.C. Vaillant. The t91 book; ferritic tubes and pipe for high temperature use in boilers, January 1990. Vallourec Industries, France.

[26] R. W. Swindeman V. K. Sikka and P. J. Maziasz. Evaluation of t91 after 130,000 hours in service, July 1998. Paper for ASME Pressure Vessels and Piping Conference.