ABSTRACT

VAIL, MATTHEW WILLIAM. Towards the Preservation of Privacy and Legal Compliance in Healthcare Systems. (Under the direction of Dr. Ana I. Antón).

Given the introduction of United States legislation that governs the collection, use, and disclosure of sensitive patient information, there is a need for mechanisms to preserve the privacy of sensitive information in software systems and to ensure these systems comply with law. One such piece of legislation is the Health and Human Services' (HHS) Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. The introduction of such legislation poses many challenges to organizations seeking to comply with the law, and thereby avoid severe penalties.

A study was conduct by Antón et. al, prior to the enactment of HIPAA (pre-HIPAA), to examine the content of online privacy policies. This thesis expounds upon this work by replicating the analysis, after the enactment of HIPAA (post-HIPAA), in order to evaluate the evolution of privacy policies in the presence of legislation. We discovered that since the introduction of HIPAA, the privacy policies of healthcare organizations have evolved significantly. One of the most noteworthy discoveries made during this post-HIPAA study was the lack of clarity and readability of healthcare enterprise privacy policies.

To address the need for more clear and concise privacy policies, we conducted an experiment using an empirical survey instrument that we developed to investigate user perception and comprehension of alternatives to natural language privacy policies. Some of the more compelling observations we made include:

- Users felt more secure and protected by natural language privacy policies.
- Users comprehend alternatives to natural language policies better than the original natural language privacy policies.
- User perception and comprehension of privacy policies are not in alignment with one another.
- Human Computer Interaction (HCI) factors play a significant role in the perception and comprehension of privacy policies.

In addition to evaluating how privacy policies evolve with the introduction of legislation, we attempted to explore whether organizations were actually in compliance with legislation. We developed a methodology for extracting rights and obligations from regulatory texts in order to determine stakeholder obligations. This information can be used to perform a comparative analysis by the organization to ensure compliance, or by external parties to detect potential non-compliance.

Towards the Preservation of Privacy and Legal Compliance in Healthcare Systems

by

MATTHEW WILLIAM VAIL

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

COMPUTER SCIENCE

Raleigh

2006

APPROVED BY:

_____
Dr. Ana I. Antón
Chair of Advisory Committee

_____ _____
Dr. Julia Earp Dr. Ting Yu

DEDICATION

I dedicate this work to my father, Franklin, and my mother, Dee. It was their love and support that made this possible.

BIOGRAPHY

Matthew Vail was born in Garden City, Michigan. He is a son to Franklin and Darrella Vail, a younger brother to Michelle and Lana, and a big brother to Lori. Matthew lived in Michigan, where he attended the University of Michigan at Dearborn, until he was 19 years old. He then moved to North Carolina where he began work as a Distributed Computing Services Engineer for the State of North Carolina's Administrative Office of the Courts Technical Services Division. In 2001, while continuing to work fulltime, he went back to school at the University of North Carolina at Chapel Hill. A year later, he transferred to North Carolina State University where he finished his Bachelor's degree in Computer Science and will receive his Master's degree in Computer Science. While at North Carolina State University, he was awarded the CISCO Information Assurance Scholarship in recognition of his research contributions in the field of privacy.

ACKNOWLEDGEMENTS

I would like to thank my supervisor at AOC, Linda Ward, for her continued encouragement and support in the pursuit of my education. Many thanks to the members of my thesis committee for their valuable feedback and support, to the members of ThePrivacyPlace.Org research group for their assistance, and to the NSF for their financial support in conducting this research (NSF grant #032-5269). Additional thanks to IBM for their support in encouraging survey participants.

Table of Contents

List of Figures

List of Tables

Chapter 1

"Relying on the government to protect your privacy is like asking a peeping Tom to install your window blinds."
- *John Perry Barlow*

INTRODUCTION

The objective of this thesis is to advance the privacy preservation and legal compliance in software-based healthcare systems. To this end, we employ analytical and empirical research methods to support our investigations.

Recent United States legislation has been passed that governs the collection, use, and disclosure of sensitive patient information. This thesis explores the evolution of organizational policies that govern such information. We use the results of these findings to substantiate the need to develop and validate an empirical survey instrument that investigates user perception and comprehension of these various policies within the context of web-based healthcare systems. The results of these two efforts are presented in this thesis.

We present heuristics that we have developed for analyzing legislative text artifacts. These heuristics aid in identifying and classifying the requirements that healthcare systems must abide by in order to be compliant with law. We compare these requirements with the findings from our policy analysis to investigate the level of compliance in various healthcare domains. The results of this analysis may be helpful for forecasting how future legislation will affect the state of online privacy in other domains, as well as aiding healthcare organizations to ensure they are in compliance with such legislation.

1.1 Research Context

The Internet has enabled novel innovations such as e-commerce, online bill payment, online healthcare and prescription drug transactions, scientific collaboration, and distance education. The most notable reason behind the Internet's success and rapid evolution is the wealth and availability of information it provides. On February 18, 2004, Google announced that its breadth spanned over 4.3 billion web pages [CNN04]. Additionally, initiatives are in place to digitize entire libraries of information [WHIR05] and make them searchable via the Internet. This presents Internet users with a virtually boundless pool of information at their fingertips.

The Internet's increase in size, however, has also resulted in the proliferation of sensitive and personally identifiable information (PII). Internet users are increasingly concerned with the electronic transmission and dissemination of their PII [Jup02, WW05], which prevents users from fully enjoying the benefits that these services provide [PM04]. The current state of personal information management is poor. In fact, there are only a few safeguards against misuse and disclosure of sensitive information, all of which are inadequate.

As Internet users continue seeking out more and different services online, they are being asked to yield their personal information in exchange for these services. Studies have shown that users are concerned about their privacy on the Internet [Westin, ACR99]. In fact, one study [Romney04, PM04], revealed that 80 percent of Internet shoppers abandon their shopping carts upon encountering the pre-purchase demand of personal information, while another study [Jup02] confirms that nearly 70 percent of Internet shoppers are worried that their privacy is at risk. This fear is not unfounded, since a recent study found that 91 percent of U.S. websites collect personal information and 90 percent collect PII. However, studies have also shown that, despite their concerns for privacy, users are still utilizing online services and e-commerce is thriving [EStats05].

There are two primary ways for the confidentiality of users' PII to be compromised by an online organization: inadvertent and intentional. Inadvertent disclosures can occur due to various reasons such as employee negligence, insufficient safeguards, and poor business procedures. For example, in 2005, CitiGroup lost the data of nearly 3.9 million customers

when they lost computer tapes being sent via United Parcel Service (UPS) [CNN05], while ChoicePoint, a data brokerage company, was duped into disclosing PII of over 140,000 consumers [CP06]. Furthermore, according to a survey by the Federal Trade Commission (FTC), in 2003, almost 10 million Americans had fallen victim to identity theft, resulting in costs of over $5 billion for consumers and $48 billion for businesses [FTC03]

Not all disclosures to a third party are inadvertent. Intentional disclosures are initiated by an organization in possession of PII with the intent to transfer the PII to a third party. In 2003, JetBlue Airways publicly acknowledged that it provided the travel records of five million JetBlue customers to Torch Concepts, a Department of Defense contractor. This disclosure was intentional and in violation of JetBlue's privacy policy [AHB04].

Intentional disclosures are not always in violation of the privacy policies of an organization. In fact, organizations' often benefit monetarily from the disclosure of PII. Sensitive data has value [PM04], as evidenced by companies such as Choicepoint and Axciom who thrive fiscally on the collection and aggregation of consumer information [AHB04]; and as such, companies that collect consumer information often sell it to interested third parties. Often times, consumers are unaware that their information is being collected for transfer. According to a recent survey at the University of Pennsylvania [TFM05], approximately two-thirds of Americans are unaware that supermarkets are allowed to sell purchase decision information to other companies.

In response to consumer concerns, companies began to self-regulate to prevent consumers' privacy from being compromised. However, in June 1998, the FTC concluded that "the Commission has not seen an effective self-regulatory system emerge," and it urged "that Congress develop legislation to require commercial Web sites that collect personal identifying information from children 12 years of age and under to provide actual notice to the parent and obtain parental consent" [Clarke99]. Legislation in the form of the Children's Online Privacy Protection Act (COPPA)[1] was passed in 1998. While certainly a valiant step towards protecting consumer, this legislation was limited to the domain of protecting the privacy of individuals under the age of 13, which excludes a large portion of the Internet

[1] Children's Online Privacy Protection Act of 1998, 15 U.S.C. §§ 6501- 6506.

community. In addition to COPPA, Congress has also passed the Gramm-Leach-Bliley Act[2], which governs the protection of financial data, and the Health Insurance Portability and Accountability Act (HIPPA)[3], which governs the protection of healthcare data.

HIPAA was passed in 1996 to improve efficiency in healthcare delivery by standardizing electronic data interchange, and to protect the confidentiality and security of healthcare information by setting forth guidelines and standards to be followed. HIPAA is comprised of various rules that address specific compliance standards for various domains [HIPAA05]. One of these rules, the HIPAA Privacy Rule, was published on December 28, 2000, with a compliance deadline of April 14, 2003. With the compliance date being so recent, little is known about how well companies have adhered to the standards outlined in the Privacy Rule.

1.2 Thesis Outline

This thesis explores the impacts of legislation on sensitive data from three vantage points: organizational policy, consumer comprehension, and legal requirements. We begin by analyzing the evolution of organizational privacy policies within the healthcare domain since the enactment of the HIPAA Privacy Rule. We use these results as the foundation for the rest of the thesis.

Subsequently, we investigate the requirement that healthcare organizations must provide consumers with a plainly written notification of uses and disclosures. To do this, we created an empirical survey instrument that will help aid our investigations of privacy policy expression and user comprehension thereof. The instrument measures how well users perceive policies that are expressed using alternatives to natural language policies. Additionally, the instrument gauges the ease with which users are able to read and comprehend the policies they are presented with and compare these results with the natural language policies.

---

[2] Gramm-Leach-Bliley Act of 1999, 15 U.S.C. §§ 6801- 6809 (2000).

[3] Health Insurance Portability and Accountability Act of 1996, 42 U.S.C.A.1320d to d-8 (West Supp. 1998).

Lastly, we developed heuristics to analyze the HIPAA Privacy Rule to extract and classify the rights and obligations of the various entities governed by the Privacy Rule. We utilize these heuristics to identify the constraints imposed on these rights and obligations by the Privacy Rule, resolve any anomalies that exist, distill the constraints into parameterized statements for classification and conflict resolution, and, finally, extract organizational and system requirements. We then compare these requirements to the previously analyzed privacy policies to investigate the extent to which organizations are compliant with HIPAA.

The remainder of this thesis is organized as follows. Chapter 2 presents our analysis of the evolution of healthcare privacy policies in the presence of legislation. Chapter 3 presents the results of our empirical survey that investigated user comprehension and perception of alternatives to natural language privacy policies. Chapter 4 presents a methodology we developed for extracting rights and obligations from regulatory texts in order to evaluate regulatory compliance. Chapter 5 provides a conclusion to this thesis, as well as summarizes its content.

Chapter 2

"This isn't just a legal compliance issue for us. We consider the privacy issue to be an opportunity to reinforce our brand image."
- *Tom Warga*

ANALYSIS OF PRIVACY POLICY EVOLUTION

This chapter discusses a unique longitudinal study that examines the effects of HIPAA's enactment on a collection of privacy policy documents for a fixed set of organizations over the course of four years. Specifically, we present our analysis of 24 healthcare privacy policy documents from nine healthcare Web sites, analyzed using a content analysis technique called goal-mining. Goal-mining is an analysis method that supports extraction of useful information about institutions' privacy practices from documents. We compare our results (post-HIPAA) to a prior (pre-HIPAA) study [AER02] of these same institutions' online privacy practices and evaluate their evolution in the presence of privacy laws.

For this study, we analyzed 24 online privacy documents from nine healthcare institutions using goal-driven requirements engineering and text readability metrics [AER02]. Our sample consisted of Web sites from three pharmaceutical companies (GlaxoSmithKline, Novartis and Pfizer Inc.), three health insurance companies (Aetna, AFLAC and CIGNA) and three online pharmacies (DestinationRx, Drugstore.com and HealthCentral). Each of these healthcare institution's privacy statements had been previously analyzed using the same goal-driven approach during the summer of 2000 prior to HIPAA's enactment [AER02]. The 24 privacy policy documents examined for this study were in force on September 4, 2003.

2.1 Privacy Policies

The Federal Trade Commission (FTC) states that a privacy policy is a comprehensive description of a domain's information practices, which is located in one place on a website, and may be reached by clicking on an icon or hyperlink [FTC98]. A privacy policy acts as a website's official statement as to how the website will collect, use, and disclose the personal information of consumers that utilize their site. Though not necessarily true in practice, privacy policies are supposed to reflect a website's actual privacy practices and serve as an understanding between the website and the consumer.

In the 1990's, only 14% of highly visited websites posted any notice regarding information privacy practices [FTC98]. The Progress for Freedom Foundation (PFF) reports that this increased to 83% in 2001. This seems to be due to an increase in an organization's perceived trustworthiness when the privacy policies are prominently displayed and clearly stated [EASS05]. However, privacy policies for web-based systems are developed as an afterthought [AE01], which may lead to less than adequate privacy policies that may not be compliant with legislation it is subject to.

Recent studies have shown that an encouraging number of people that first visit a website choose to read the website's privacy policy. However, during an analysis of financial privacy policies, Antón et. al discovered that many of the privacy policies were less than clear and conspicuous [AEB04]. Their results show that it took an average of 14.1 years of education, the equivalent of two years of college, to understand the privacy policies of those institutions. This leaves approximately 28 percent of the Internet population unable to comprehend the policies, and thus unable to understand how their sensitive data may be used, and possibly compromised, within the organization.

Unfortunately, simply posting a privacy policy has not proven sufficient for protecting consumer privacy. With the current lack of legislation, outside of specific domains such as healthcare, to govern the content and enforcement of privacy policies, organizations resolve to self-regulation. Many argue that this is simply not sufficient for protecting consumer PII [PM04, Romney04].

This thesis is concerned with policies, opinions, and legislation related to the healthcare domain. Approximately eighty percent of adult Internet users consult online

healthcare services for health information, making healthcare research the third most popular online activity behind email and researching a product before buying it [Fox03]. The Pew Project surveyed 2,038 adults in 2002 to examine the kinds of information Internet users seek online; this survey revealed that people searching for health information use the Internet to become informed, share information, seek and provide support, as well as schedule appointments [Fox03]. The evolving trend toward Internet supported healthcare services has resulted in increased information sharing among providers, pharmacies and insurers. However, according to recent studies [AER02, GHS00], inconsistencies exist between privacy policies and the actual privacy practices of healthcare-related Web sites.

This chapter of the thesis examines the privacy practices of three categories of healthcare websites – health insurance companies, pharmaceuticals, and online drugstores – by analyzing 24 online privacy documents from nine institutions, three from each category. We present a unique, longitudinal study that examines the effects of HIPAA's enactment on a collection of privacy policy documents for a fixed set of organizations over the course of four years. Specifically, we present our analysis of 24 healthcare privacy policy documents from nine healthcare Web sites, analyzed using a content analysis technique called goal-mining. We then compare our results to a pre-HIPAA study [AER02] of these same institutions' online privacy practices and evaluate their evolution in the presence of privacy laws.

2.2 Goal-mining

Goal mining is an approach that can be applied to identify strategic and tactical goals as well as requirements. This approach has been successfully applied to identify and refine goals for software systems so they may be converted into software requirements. Goal mining extracts goals from data sources by applying goal-based requirements analysis methods [AEB04]. In this study, we employ goal mining to extract pre-requirements from post-requirements text artifacts. We then use these goals to reconstruct the implicit requirements met by the privacy policies.

The goal-mining process begins by analyzing a policy document to gain an in-depth understanding. We first extract goals by analyzing each statement within the policy and

asking, "*What* goal(s) does this statement exemplify?" and "What goal(s) does this statement obstruct or thwart?" Any statement that contains an *action word* (verb) is a goal candidate. Consider the following excerpt from the Drugstore.com privacy policy:

> *"We may **enter** into an agreement with other companies or include individuals to perform functions on our behalf. These functions may include sending promotional e-mails on our behalf to such companies customers'; serving advertisements on our behalf; providing marketing assistance; processing credit card payments; and have access to information necessary to perform their functions."* [Drug03]

When analyzing this statement, we identify action words. The first action word in this example is "enter." We then identify *actors* (person performing the action), *targets* (who the action is being performed on), *instruments* (how the action will be performed), and *purpose* (why the action is being performed). By asking the goal identification questions, we extracted the following goals from the above example.

> G_{867}: USE customer email address for marketing and promotional purposes
>
> G_{642}: SHARE CI with subsidiaries to recommend services to customer
>
> G_{1166}: SHARE CI with 3rd parties to perform marketing services on our behalf
>
> G_{1167}: SHARE CI with 3rd parties to provide valuable financial services we do not offer (e.g. credit card)

After identifying the goal components, we classify each goal statement according to *keywords* and express it in structured natural language [AEB04]. Each action word maps to a keyword in the Antón goal mining taxonomy [AEB04]. A list of common privacy keywords

is presented in Table 2.1. In the above example, the action word “sending” maps to the keyword “SHARE”.

Table 2.1 Common privacy keywords [AEB04]

ACCESS	CONTACT	EXCHANGE	OBLIGATE	RESERVE
AGGREGATE	CONTRACT	HELP	OPT-IN	REVIEW
ALLOW	CUSTOMIZE	HONOR	OPT-OUT	SHARE
APPLY	DENY	IMPLY	INVESTIGATE	SPECIFY
AVOID	DESTROY	INFORM	POST	STORE
BLOCK	DISALLOW	LIMIT	PREVENT	UPDATE
CHANGE	DISCIPLINE	MAINTAIN	PROHIBIT	URGE
CHOOSE	DISCLAIM	MAKE	PROTECT	USE
COLLECT	DISCLOSE	MAXIMIZE	PROVIDE	VERIFY
COMPLY	DISPLAY	MINIMIZE	RECOMMEND	
CONNECT	ENFORCE	MONITOR	REQUEST	
CONSOLIDATE	ENSURE	NOTIFY	REQUIRE	

Identified goals are classified as either protection goals or vulnerabilities. *Privacy protection goals* express ways in which sensitive information is protected. *Privacy vulnerabilities* reflect ways in which sensitive information may be susceptible to privacy invasions. Goals not relevant to privacy or privacy-related functionality are marked as unclassified for purposes of this research because they are outside of the scope of privacy and security requirements.

Once goals have been classified, they are further categorized by type [AE04]. There are five types of privacy protection goals: notice and awareness, choice and consent, access and participation, integrity and security, and enforcement and redress [AER02]. *Notice and awareness* goals reflect ways in which customers are notified and/or made aware of an organization’s information practices before any information is actually collected from them (e.g. G_{1270}: `INFORM customer that cookies are not required to browse our site`). *Choice and consent* goals reflect ways in which a Web site ensures that consumers are given options as to what personal information is collected, how it may be used and by whom (e.g. G_{223}: `CHOOSE not to receive product info by mail/email/phone`). *Access and participation* goals reflect ways in which consumers access, correct and challenge any data about

themselves; for example, by providing a means for consumers to ensure their data is accurate and complete (e.g. G_1: ALLOW customers to modify/remove their PII). *Integrity and security* goals reflect ways in which a Web site ensures that data is both accurate and secure (G_{88}: GUARD data during transmission using SSL encryption technology). Finally, *enforcement and redress* goals reflect ways in which a Web site enforces its policies (e.g. G_{44}: DISCIPLINE employees who violate PP).

There are seven types of vulnerabilities: information collection, monitoring, personalization, storage, transfer, aggregation and contact [AER02]. *Information collection* addresses how and what information is being collected from the consumer by an institution, either by directly requesting information or by collecting information without consent (e.g. G_{633}: COLLECT CI to facilitate transaction). *Information monitoring* reflects ways in which organizations may track when consumers use their site (e.g., via cookies) often times with the expressed intent of providing benefits to the consumer, such as a customized online experience (e.g. G_{503}: USE session cookies to facilitate online transactions). *Information personalization* reflects Web site customization and tailoring of the functionality and content offered to individual users (e.g. G_{110}: CUSTOMIZE website using PII). *Information storage* refers to what and how information is maintained in an institution's database (e.g. G_{764}: STORE PII long enough to meet legal or regulatory requirements). *Information transfer* concerns any transfer of information from one entity to another (e.g. G_{190}: SHARE PII w/ affiliates to provide marketing support). *Information aggregation* concerns the combination of previously gathered PII with data acquired from other sources (e.g. G_{909}: AGGREGATE statistical info about frequency of visits). *Contact* concerns how, and for what purposes, an organization contacts a consumer (e.g. G_{856}: CONTACT cust for marketing purpose).

Since many policies convey similar privacy practices, many goals are repetitive across different organizations as well as within a given organization's policy document. To abate the occurrence of goal repetition, we employed the Privacy Goal Management Tool (PGMT) to support our goal mining efforts. The PGMT maintains a goal repository containing goals from over 100 Internet privacy policy documents. Each goal in the repository is associated with a goal ID, a description, an actor, its sources, and a taxonomy

classification [AEB04]. When a statement has been analyzed and a goal has been identified, the PGMT can be searched to determine whether the goal already exists. This saves the researcher valuable time, as well as ensure consistency within the goal set.

Once goals are classified in the PGMT, they must be further refined. Goal refinement entails removing synonymous and redundant goals and resolving any inconsistencies that exist within the goal set. Consider the following two goals:

> G_{1187}: AVOID disclosing names, address, email to 3rd party without consent
>
> G_{1183}: AVOID sharing sensitive PII with 3rd parties without customer consent

These goals are synonymous and thus redundant because names, addresses and email addresses are all considered forms of PII. Thus, these goals were merged by eliminating goal G_{1187} and replacing it with goal G_{1183}. The objective here is to enable goal reuse by capturing the high level intent of the goal. The goal refinement process helps create a standardized, non-redundant goal set in the PGMT repository and aids in conducting a comparative analysis of different organizations' policy documents.

As part of the goal refinement process, we distinguish between various types of information. By doing so, we can improve the granularity of our goal classification. For example, PII is distinguished from customer information (CI) because PII contains information that can lead to the identification of an individual, while CI may or may not contain PII. Since PII is a subset of CI, PII is finer in granularity and can be distinguished from CI in a goal. Such classifications aid the analysis of policies when performing goal mining, ensuring that the goals representing a policy accurately preserve the original policy content.

2.3 Evaluating Privacy Documents for Readability

As previous research has illustrated [AEB04], natural language privacy policies are difficult for the average Internet user to comprehend. These findings were validated during

our analysis of the online healthcare privacy documents. Several of the documents proved to be difficult to comprehend.

To quantify the readability of these documents, we employed the Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FGL) score methods [Fle49]. The FRES and FGL methods provide a standardized and statistical metric to objectively analyze the text contained in documents. The Flesch metrics give an approximate measure of a text's difficulty. The FRES is often used to evaluate legal documents and is used to regulate the complexity of insurance policies in more than 16 states [AEB04]. The FGL is a number that estimates the number of years of schooling required for an individual to be able to read and understand a document; for example, a score of 9.0 means that a ninth grader would be able to understand a document. These two readability metrics (see Table 2.2) are based on a formula that considers the sentence length and word choice (based upon number of syllables) to determine the overall readability of a document. The FRES is a scale from 0 to 100, with 0 representing a difficult document to read, and 100 representing an easy document to read. Average sentence length for a score of 0 is 37 words and for a score of 100 is 12 words or less [Gno04].

The FRES and FGL scores for each analyzed privacy document are shown in Table 2.3. The FGL reading difficulty scores were calculated using Microsoft Word[4], which returns correct grade levels for values less than 12. For values greater than 12, the average syllables per word were extracted from the Flesh Reading Ease score returned by Word, and then used in the Flesch-Kincaid grade level formula; 206.835 – (1.015 x average sentence length) – (84.6 x average syllables per word).

Table 2.2 Flesch Metrics Formulas [Flesch49]

Flesch Reading Ease Score: 206.835 – 84.6 * (total syllables / total words) – 1.015 * (total words / total sentences)
Flesch-Kincaid Grade Level: (0.39 * average sentence length (in words)) + (11.8 * average number of syllables per word) – 15.59

[4] http://www.microsoft.com/office/word/

The pre-HIPAA scores were calculated for privacy policy documents that were in effect during the summer of 2000 before HIPAA went into effect. Table 2.3 also provides an average score for all the post-HIPAA privacy documents for a given site, including Legal Disclaimers and Terms of Use.

2.4 Findings

In this section, we present and discuss the results of our analysis of the HIPAA Privacy Rule.

2.4.1 Privacy Practice Transparency

As part of our study, we examined the various categories of vulnerabilities and protection goals and compared them with the pre-HIPAA analysis findings. We found that both the number of protection goals and vulnerabilities increased within the nine companies' policy documents. The number of vulnerabilities increased by 6 times in contrast to an increase in the number of protection goals of 3.5 times (see Figure 2.2). This differs from the pre-HIPAA study in which we identified more protection goals than vulnerabilities.

This might suggest that consumers' sensitive information is more susceptible to privacy breaches now than prior to HIPAA's enactment. However, the more likely explanation for such a substantial increase in vulnerabilities is a greater level of transparency. Pre-HIPAA, there were no regulations requiring the disclosure of an institution's privacy practices in their privacy policy, nor were there regulations requiring that the privacy policies align with the privacy practices of the institution.

These findings suggest that HIPAA may have resulted in privacy policies that better represent the actual privacy practices of the institution. While HIPAA also likely influenced the privacy practices and business processes of these institutions, the verification of this claim is outside the scope of this thesis.

2.4.2 Categorical Observations

Vulnerability Categories

As previously mentioned, vulnerabilities reflect ways in which sensitive information may be susceptible to privacy invasions. As such, this study intended to evaluate the differences, if any, in the number and types of vulnerabilities found in the pre-HIPAA and post-HIPAA studies.

We found that the category with the greatest number of occurrences is the vulnerability category of *Information Transfer*. This increase in information transfer may suggest that the analyzed healthcare companies have begun disclosing information at an alarming rate. However, as mentioned previously, the increase can most likely be attributed to greater levels of transparency due to HIPAA.

Despite policies being more forthright, the number of occurrences is still alarming, as it illustrates that healthcare companies are disclosing personal healthcare information (PHI), a category of information the includes both PII and healthcare information, at a staggering rate.

The practice of disclosing information was also consistent across all organizations, with all but one of the organizations including at least one information transfer statement in their policies. The practice of disclosing information was also consistent across all three observed healthcare domains – health insurance, online drugstores, and pharmaceuticals. However, the health insurance companies and online drugstores policies observed the bulk of the information transfer occurrences, with pharmaceuticals observing only two information transfer occurrences.

The vulnerability categories of *Information Collection* and *Information Monitoring* approximately tripled (see Figure 2.1) from the pre-HIPAA study to the post-HIPAA study. Once again, this increase can most likely be attributed to greater levels of transparency due to HIPAA. However, in conjunction with the findings of the information transfer category, one can easily discern that companies are collecting consumer data, monitoring consumer activities, and then transferring this information to third parties.

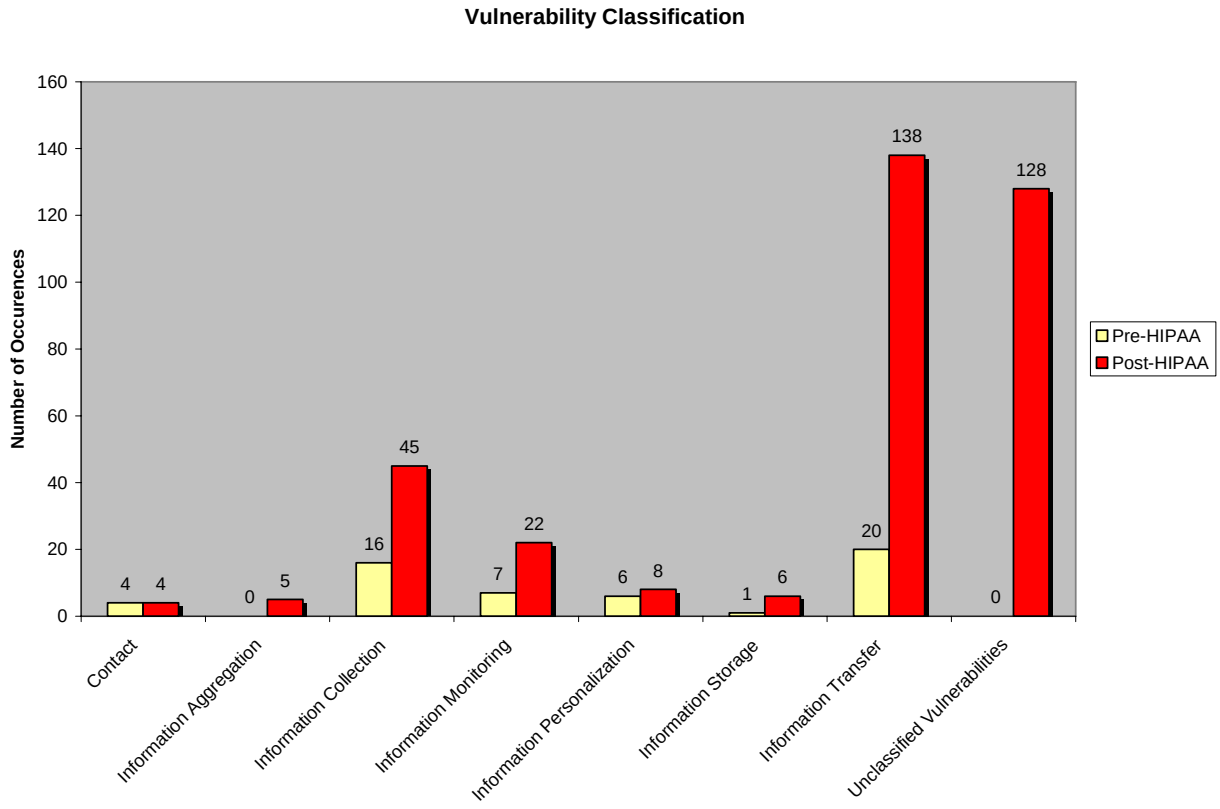

Figure 2.1 Comparison of the number of privacy vulnerabilities identified in privacy policy documents pre-HIPAA and post-HIPAA.

Protection Goal Categories

As previously mentioned, protection goals express ways in which sensitive information is protected. This study intended to evaluate the differences, if any, in the number and types of protection goals found in the pre-HIPAA and post-HIPAA studies.

We discovered that the category with the greatest overall increase from pre-HIPAA to post-HIPAA was the protection goals category of *Notice/Awareness*. The occurrences in the notice/awareness category increased from 7 (pre-HIPAA) to 59 (post-HIPAA) – an increase of over 800 percent (see Figure 2.2). This is due in no small part to the portion of the HIPAA Privacy Rule mandating that healthcare organizations notify individuals' about various procedures, including use and disclosure policies, procedures to access PHI, and procedures to amend one's own PHI [HIPAA].

The post-HIPAA occurrences in the protection goals categories of *Security/Integrity* and *Enforcement/Redress* approximately doubled (see Figure 2.2) the occurrences of the pre-HIPAA study. This notable increase can be attributed, at least in part, to the increased awareness of consumers. With the media increasing visibility of identity theft and data disclosures, consumers are taking steps towards safeguarding their PII. For example, a Pew Research Study [Fox05] found that 48 percent of online consumers are avoiding some Web sites because of a fear of spyware, and Entrust has just found that nearly 20 percent of those consumers who bank online are now doing so less [Entrust05].

The post-HIPAA occurrences of the *Choice/Consent* category did not increase as much as we initially anticipated. The total number of choice/consent occurrences is 33 compared to 59 in the notice/awareness category. This suggests that consumers are 1.79 more times likely to be told how their data will be used than asked for permission to use their data. There is an average of 1.27 choice/consent goals per privacy document.

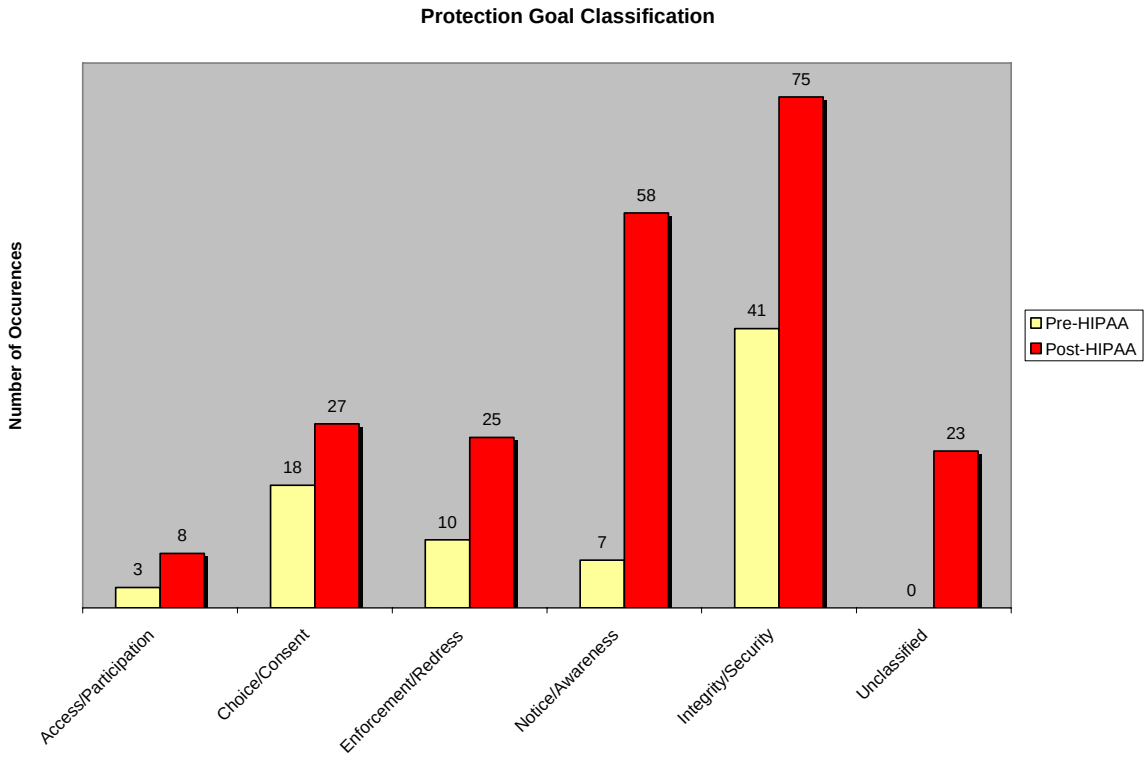

Figure 2.2 Comparison of the number of privacy protection goals identified in privacy policy documents pre-HIPAA and post-HIPAA.

Unclassified Goals

The post-HIPAA privacy documents have also evolved to include legal and liability disclaimers. Since these types of disclaimers do not directly affect the privacy of consumers' information, the taxonomy we used in this study did not account for them. This resulted in a large number of unclassified goals. Another way we distinguished unclassified goals from the rest of the goals in the privacy documents is by asking whether the goal is within the scope of what would be implemented as a security or privacy requirement.

It is also interesting to note that unclassified goals account for 36% of the total post-HIPAA vulnerabilities. This illustrates the vast variety of information consumers must navigate in order to understand the institutional privacy policies. This finding also indicates

that the privacy taxonomy we used may need to be augmented to account for these new types of vulnerabilities.

2.4.3 Organizational Self-Regulation

Romney [Romney04] states that "many privacy policies are often offensively paternalistic and one-sided," meaning that privacy policies often contain vernacular that seems ensuring, yet suspect. For example, a website policy might say that they are "compiling information for your benefit to improve the quality of services we provide." Markel and Perkins [PM04] argue that the consumer, not the organization, can make the most appropriate decision as to what is in the best interest of the consumer. Given that most websites are not subject to privacy legislation, consumers must trust that self-regulation will be sufficient to protect the privacy of their PII.

This notion of self-regulation may explain the great number of ambiguous statements found in many privacy policies today. Consider this following excerpt taken from a General Motors privacy policy [PM04]:

> *"When other information is collected from you, such as your name and e-mail address, we <u>generally</u> let you know at the time of collection how we will use the personal information. <u>Usually</u>, we use the personal information you provide only to respond to your inquiry or to process your request companies privacy policy we found in the privacy documents we examined post-HIPAA."*

The terms "generally" and "usually" are underlined to call attention to the variety of ambiguous vernacular that can be found in privacy policies. "Generally" and "usually" imply that sometimes, but not always, these actions are taken. The question one asks is: Where are the other disclosures, and why are we not told what they are?

Similar ambiguities were observed during our analysis of the post-HIPAA privacy documents. HIPAA requires organization to "make reasonable efforts to use, disclose, and request only the minimum amount of protected health information needed to accomplish the intended purpose of the use, disclosure, or request" [HHS03]. Unfortunately, as we observed in our study, many of the self-governed business transactions that result in the sharing of

sensitive data seem questionable. For example, Drugstore.com's Terms of Use document states that they "may disclose any content, records, or electronic communication of any kind ... if such disclosure is necessary or appropriate to operate the site." Because HIPAA allows organizations to self-govern what a reasonable effort is to protect data, a consumer would find it difficult, if not impossible, to discern what constitutes an "appropriate" disclosure [Drug03].

2.4.4 Policy Composition

Not only has the number of policy documents increased from nine (pre-HIPAA) to 24 (post-HIPAA), but their readability has also decreased. The documents in the post-HIPAA study were lengthier and contained more information about the privacy practices of the institutions than the pre-HIPAA documents. Thus, consumers are now burdened with having to read lengthier documents that are more difficult to comprehend in order to properly evaluate how an institution's privacy practices may affect him or her.

On average, an institution's main privacy policy document is now approximately two times the length of the same pre-HIPAA privacy document. For instance, Health Central's privacy policy contains 1,278 words compared to its pre-HIPAA version, which contained only 683 words. The total number of policy documents for the nine analyzed institutions also increased from nine (pre-HIPAA) to 24 (post-HIPAA). The addition of 15 new documents yielded a significant increase in goal occurrences from 153 goals (pre-HIPAA) to 609 goals (post-HIPAA). This increase suggests that the introduction of HIPAA has caused online healthcare organizations to be more comprehensive in describing their online privacy practices.

Another interesting metric to examine is the extent to which we were able to reuse goals from the pre-HIPAA study in this study. The ability to reuse goals indicates some stability in the way some privacy practices are expressed. In a sense, these goals survived the introduction of HIPAA. In contrast, when few goals are reused it suggests that the introduction of HIPAA required a fundamental change in the way privacy practices are now expressed or the need for organizations to express a broader range of privacy practices.

In this study, although the total number of goals increased, the percentage of goal reuse dropped from 69% (pre-HIPAA) to 43% (post-HIPAA) and the total number of unique/new goals increased from 133 (pre-HIPAA) to 366 (post-HIPAA). We observed that prior to HIPAA, healthcare privacy documents were strikingly similar, and in some cases several privacy policy sections were even identical across different organizations. Given that HIPAA now requires more detailed notices about an organization's privacy practices, it is reasonable to see more detailed, company-specific information expressed in these policy documents. To the benefit of consumers, this has resulted in more complete and unique privacy documents. At the same time, this increase in uniqueness makes it more difficult for consumers to compare the practices of different organizations as is the case with financial privacy documents [AEB04].

The types of vulnerabilities that appear to have survived HIPAA's introduction are those that reflect information transfer and general information collection (such as cookies). The types of protection goals that appear to have survived HIPAA's introduction are the inform goals (e.g., `INFORM customer of intended use of PII`) and the security goals (e.g., `MAINTAIN procedural safeguards to protect PII`).

The privacy documents have also evolved to include legal and liability disclaimers. Because these types of disclaimers do not directly affect the privacy of consumers' information, the taxonomy we used in this study did not account for them. This resulted in a large number of unclassified goals. It is also interesting to note that unclassified goals account for 36% of the total post-HIPAA vulnerabilities. This illustrates the vast variety of information consumers must navigate in order to understand the institutional privacy policies.

2.4.5 Policy Readability and Comprehension

The introduction of HIPAA has made healthcare privacy policy documents more difficult to comprehend. The readability of the all the privacy documents within the examined organizations decreased as shown in Table 2.3. The average FGL for all privacy documents increased from 13.3 (pre-HIPAA) to 14.2 (post-HIPAA), and the average FRES decreased from 39.6 (pre-HIPAA) to 34.9 (post-HIPAA). The increase in the FGL is the equivalent of almost an entire grade level of education, making the already difficult to

understand documents less comprehensible to a large percentage of the general population. An FGL score of 14.2 is the equivalent of 2 years of college education or an associate's degree. Studies have shown that only 52.1% of the general population has obtained this level of education [NTI02].

The average FGL score for an organization's main "privacy policy" document increased from 13.3 (pre-HIPAA) to 13.87 (post-HIPAA) and the FRES also increased from 36.86 (pre-HIPAA) to 38.3 (post-HIPAA). However, the newly added privacy related documents (e.g. Terms of Use, Legal Disclaimers, Privacy Facts, etc.) received significantly higher FGL scores (see Table 2.3: Novartis's terms of use yields a FGL of 16.7 compared to a FGL of 13.8 for its privacy policy) and lower FRES scores (Novartis' Terms of Use FRES score is 27.2, whereas the Privacy Policy FRES score is 38.2), making them the most difficult to understand.

Interestingly, these documents also generally contain a higher ratio of vulnerabilities to protection goals. This is alarming because consumers should not be burdened with having to read complex documents to uncover these possible compromises to their sensitive information. Moreover, these documents seem to convey a misleading sense of protection to consumers with promising statements at the beginning of the document. For example, HealthCentral.com [HC03] stated, "HealthCentral is deeply committed to preserving your privacy" in the first paragraph of their Privacy Policy. However, their legal disclaimer contains only six protection goals and 18 vulnerabilities. Additionally, HealthCentral's Privacy Policy yielded a FGL score of 14.2, implying that only less than 30% of the general population can comprehend its content [NTI02]. This, combined with the fact that there are three vulnerabilities to each protection goal in this document, seems to contradict the organization's expressed commitment to their customers.

It is also important to note that the increase in the total number of privacy documents within these institutions, along with the increased difficulty of comprehension, may lead to information overload where consumers are presented with so much information that it will be difficult, if not impossible, to retain it in its entirety.

Table 2.3 Pre-HIPAA vs. Post-HIPAA Analysis of Privacy Policies: Goal Classification and Flesch Readability. Grey cells indicate cases in which there was no Pre-HIPAA data because the documents were introduced Post-HIPAA or cases in which the Pre-HIPAA document were replaced with newly-titled documents. Blank cells indicate null data.

	Policy Document	Protection Goals		Vulner-abilities		Unclass-ified		Total Goals		FRES		FGL	
Pre / Post HIPAA		**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**
Health Insurance													
AETNA	Privacy Policy	5	8	5	12			10		42.8	39.8	13.4	13.8
	Legal Statemnt		0		9			0	9		28.4		18
	Health		16		32			0	48				14.8
	Student		16		35			0	51		39.8		13.6
	Long Term		4		26			0	30		34.5		10.2
	Large Pension		5		11		1	0	17		28.7		14.8
	Life Disability		14		25			0	39		35.1		10.6
	Subtotal	5	63	5	150	0	1	10	214	42.8	34.3	13.4	13.7
AFLAC	Privacy Policy	1		1				2	0	30.4		15	
	Terms/Cond.		8		20			0	28		32.3		14.7
	Privacy Practic		6		0			0	6		42		13.4
	Privacy Notice		15		27		2	0	44		35.1		13.7
	Subtotal	1	29	1	47	0	2	2	78	30.4	36.5	15	13.9
CIGNA	Privacy Policy	6	18	5	11			11	29	43.9	43.6	10.9	11.4
	Legl Disclaim				6			0	6		27.8		16.6
	Subtotal	6	18	5	17	0	0	11	35	43.9	35.7	10.9	14
Online Drugstores													
DestinationRX	Privacy Policy	16	20	18	16		2	34	38	40	39	12.9	13.9
	Terms of Use		7		11		3	0	21		31.8		15.5
	Subtotal	16	27	18	27	0	5	34	59	40	35.4	12.9	14.7
Drugstore.com	Privacy Policy	15	19	14	36			29	55	39.1	40.9	13.6	13.7
	Terms of Use		8		18		3	0	29		35.8		15.1
	Subtotal	15	27	14	54	0	3	29	84	39.1	38.4	13.6	14.4
HealthCentral	Privacy Policy	13	12	12	12			25	24	39.5	41	12.5	13
	Terms of Use		6		15			0	21		24.9		16.3
	Subtotal	13	18	12	27	0	0	25	45	39.5	32.9	12.5	14.7
Pharmaceuticals													
Glaxo	Privacy Policy	5	9	7	6		2	12	17	39.5	41.9	12.5	12.7
	Legal Statemnt				5			0	5		28.8		16
	Subtotal	5	9	7	11	0	2	12	22	39.5	35.4	12.5	14.4
Novartis	Privacy Policy	18	23	5	9		4	23	36	27.4	41.4	16.7	13.1
	Legal Statemnt		3		11			0	14		27.4		16.7
	Subtotal	18	26	5	20	0	4	23	50	27.4	34.4	16.7	14.9
Pfizer	Privacy Policy	4	7	3	6			7	13	41.8	41.4	11.8	11.8
	Terms of Use	0			9			0	9		37.8		12.9
	Subtotal	4	7	3	15	0	0	7	22	41.8	39.6	11.8	12.4
Totals		**83**	**224**	**70**	**368**	**0**	**17**	**153**	**609**				
Averages										**38.3**	**35.8**	**13.3**	**14.1**

2.5 Chapter Summary

The intent of this research was to analyze the evolution of privacy documents in the presence of HIPAA legislation. What we discovered is that the privacy documents of healthcare organizations evolved to incorporate more information about their privacy practices. This study also shed light on the kinds of privacy practices that are most prevalent in online healthcare organizations. This decomposition of the privacy practices of each organization assists in comparing these practices to legislation to ensure compliance. This information can be of value to researchers, policy makers, as well as the healthcare organizations.

In addition to the privacy documents being lengthier and more numerous, we discovered that the documents were also more difficult to comprehend, and they contained a number of ambiguities that could potentially be exploited to the detriment of consumer privacy. The benefit of the privacy documents being more expressive and detailed with regards to the use of privacy-related information is offset by the lack of clarity and readability of the documents. The wealth of information contained in these documents has little meaning if consumers cannot comprehend the policies.

One variable this study did not investigate was the compliance of organizational business processes with the privacy practices outlined in their privacy policies. Given that these privacy documents are required by HIPAA to represent the privacy practices of the organization, for the purposes of this study, we assumed all privacy documents to be fully representative of each organization's actual privacy practices. And while HIPAA also likely influenced the privacy practices and business processes of these institutions, the verification of this claim is outside the scope of this thesis.

We believe the results of this analysis may be helpful for forecasting how future legislation will affect the state of online privacy in other domains. Law makers can utilize the findings of this study to substantiate the need for future privacy legislation. They can also use this study to identify key shortcomings in existing privacy law, such as readability concerns, to better protect consumer privacy in future privacy legislation. Moreover, the United States needs additional non-domain-specific legislation that broadly regulates online privacy and which better protects the consumer rather than institutions, and our findings

serve as validation that privacy law creates greater transparency and encourages better privacy practices within the governed domains.

The increase in the number and length of privacy documents post-HIPAA, along with the increase in difficulty of comprehending the content of the documents places added burden upon consumers. If a consumer attempted to read each privacy document of organizations they share their information with, their online experience would most likely prove to be difficult and unpleasant. If consumers are to make informed decisions regarding their online privacy, steps must be taken to alleviate this burden. The next chapter of this thesis makes an attempt towards this very goal.

Chapter 3

"I have not failed. I've just found 10,000 ways that won't work."
- *Thomas Edison*

EMPIRICAL SURVEY

Although Internet users are becoming increasingly concerned about their privacy on the Internet [ACR99, Westin], they still have few avenues to obtain information about the privacy practices of a given organization. Often times, privacy policies provide the only insight they have into the privacy practices of an organization. For this reason, it is imperative that privacy policies are easily accessible, clear and concise, and users are able to fully comprehend the their content.

Unfortunately, previous research has shown that privacy policies are neither clear nor concise [JP04, AEB04]. Moreover, the results of the HIPAA analysis in Chapter 1 of this thesis further demonstrate the need for clear, concise, yet comprehensive privacy policies.

The need for better privacy policy representations is more pressing than ever. If Internet users cannot understand the content of a privacy policy, they will not be equipped to make an informed decision about whether to share their personally identifiable information (PII) with an organization. This may lead to users sharing their information with organizations seeking to capitalize on user PII; and as a result, users may end up having their privacy compromised. Furthermore, usability has been identified as one of the grand challenges for security and privacy research [CRA03].

In an attempt to address the need for more comprehensible privacy policies, we conducted an experiment to gauge user perception and comprehension of various representation alternatives to natural language privacy policies. By comparing these results to the natural language policies, we were able to identify shortcomings in the design of existing privacy policies, as well as make several observations that may improve the perception and readability of privacy policies.

HIPAA specifically requires entities to provide a notice to consumers that is "plainly written." Despite given this requirement, the readability results from Chapter 2 of this thesis illustrate that healthcare privacy policies are anything but plainly written. We chose to analyze a subset of the privacy policies that were analyzed in Chapter 2 of this thesis in order to aid healthcare organizations with the requirement of providing a plainly written notice to consumers. The privacy policies used in this experiment were the original healthcare policies of the organizations we investigated in Chapter 2.

This chapter presents the results of our experiment and is organized as follows. Section 3.1 outlines the design of the experiment. Section 3.2 discusses our approach for data collection. Section 3.3 presents the results of the experiment. Section 3.4 discusses the observations we made based on the results. Section 3.5 outlines the limitations of this study and presents plans for future work.

3.1 Experimental Design

The objectives of this experiment were the following: (1) to gauge user perception of various alternatives to natural language privacy policies; (2) to measure user comprehension of the alternatives; and (3) to compare user perception with user comprehension in order to determine whether they are in alignment with one another.

To this end, we conducted an experiment in the form of an empirical survey instrument. The type of experimental design employed was a randomized complete block design. The study consisted of four ways to represent privacy policies; we refer to these as *variants* (treatments):

- ***Policy***. This variant is the original natural language privacy policy, This is the most common approach to privacy policy representation. Figure 3.1 shows an example of this type of variant. This is a portion of a natural language privacy policy found on a legitimate healthcare website.

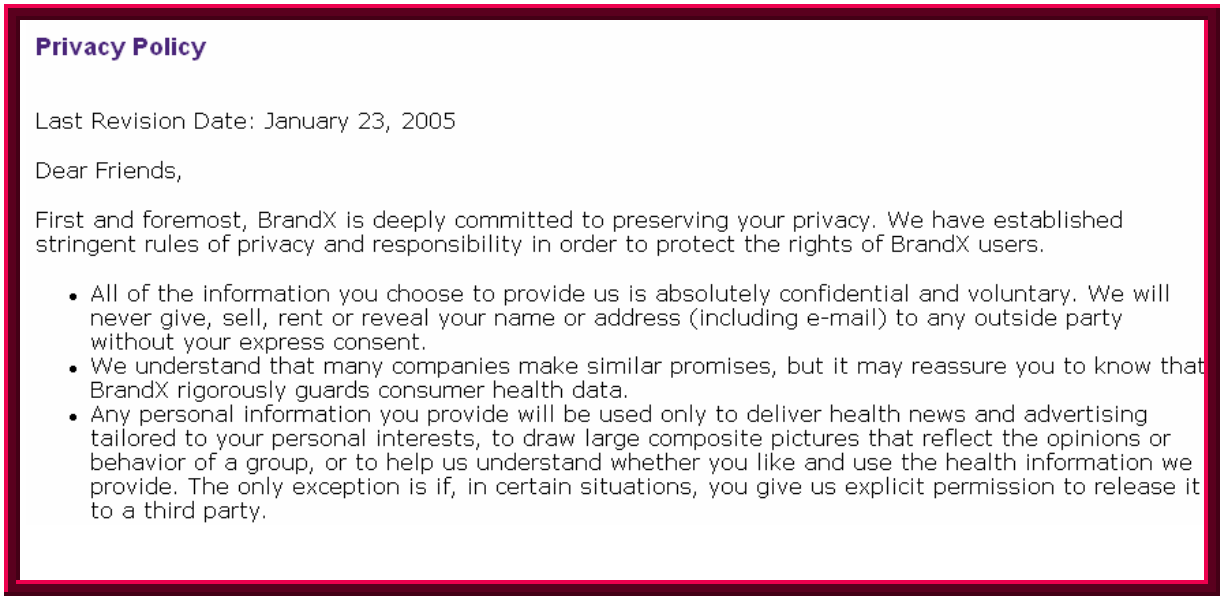

Figure 3.1 Policy Variant Illustration

- ***Goals***. In the goal variant, we expressed the policy as a list of privacy goals and vulnerabilities [AE04]. To create the list of privacy goals and vulnerability statements, we used a goal-mining approach (see Section 2.2) to distill natural language goals and warnings from stated (natural language) policies. This methodology uses a list of common words frequently used in policies. The purpose of doing this is to eliminate unnecessary text that can either mask the true meaning of a policy or cause the policy to be too complex for the general public to understand. Figure 3.2 presents an example of a goals variant. *Note: the goals in this example were mined from the natural language example in Figure 3.1.*

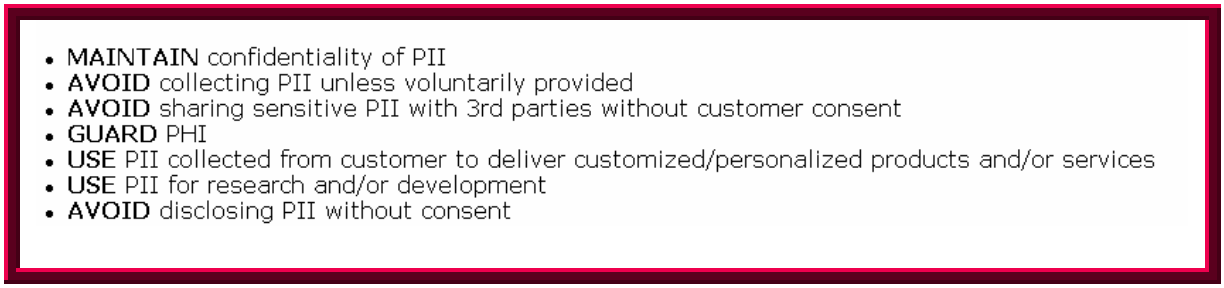

Figure 3.2 Goals Variant Illustration

- ***Categories***. In the categories variant, we expressed the policy as a list of goals that have been categorized. The goals were extracted from the original natural language policy, using goal-mining (see Section 2.2), and organized into categories based on the taxonomy [AE04] used to categorize extracted goals during the goal-mining process (e.g. information collection, notice/awareness, etc.). In this variant, subjects are first presented with a list of the 12 taxonomy categories. Subjects can then click on a category heading hypertext link to view a list of goals, presented in bulleted form, that are relevant to the given category of interest. Figure 3.3 illustrates how the categories are displayed to subjects (users), whereas Figure 3.4 illustrates what subjects would subsequently see when they click on one of the category headings.

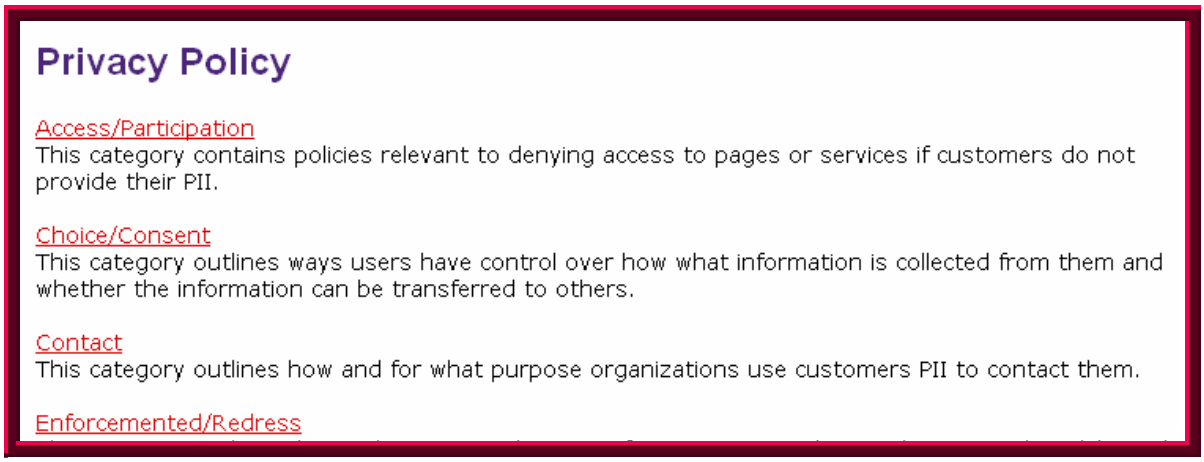

Figure 3.3 Categories Variant (list of categories)

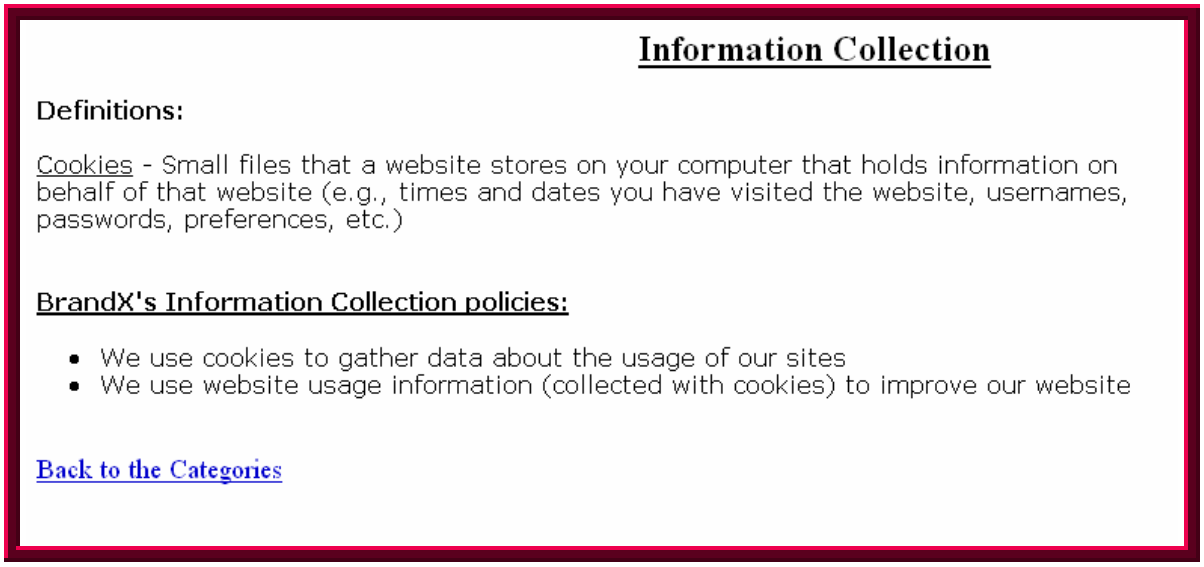

Figure 3.4 Categories Variant Illustration (goals within a category)

- ***Goals in Policy***. In this variant, subjects are presented with the original natural language privacy policy, but the format differs from the *policy* variant. Within the policy, statements that contain goals relevant to consumer privacy are bolded and highlighted. When a subject hovers their mouse over a statement, a popup window appears that contains the goal extracted from the statement. In this way, subjects are presented with both the natural language text, as well as its corresponding goal representation. Figure 3.5 illustrates what subjects see in the *goals in policy* variant. Notice the goal statements bolded and italicized within the policy, as well as the blue popup window containing the corresponding goal that appears when the subject hovers their mouse over a given statement.

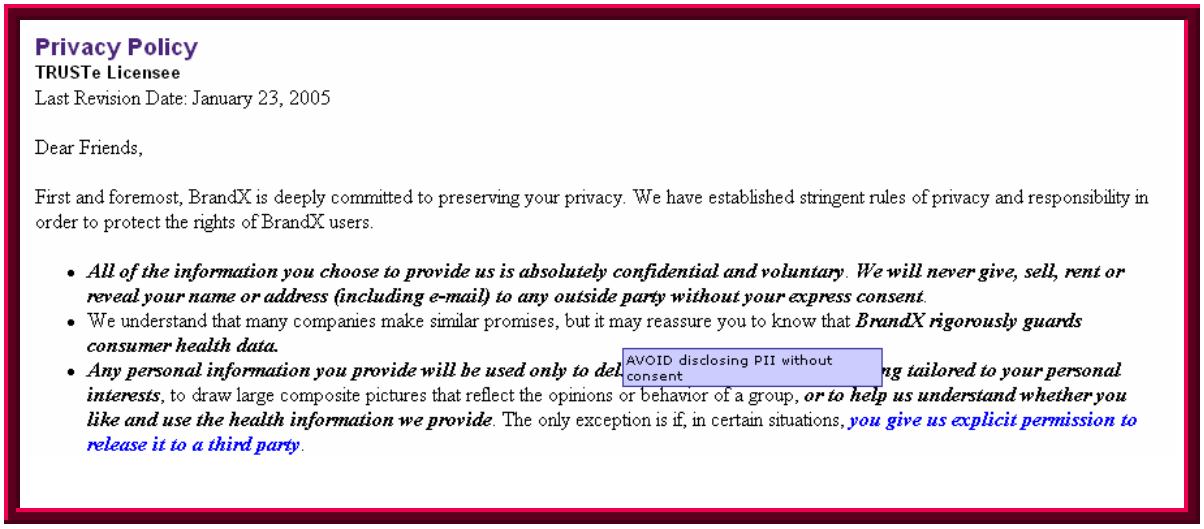

Figure 3.5 Goals in Policy Variant Illustration

In order to account for lurking variables and to incorporate replication, the experiment included three blocking factors. Each factor was a policy containing a different ratio of vulnerabilities to protection goals — Novartis.com (*good policy*), Drugstore.com (*bad policy*), and HealthCentral.com (*control*) — where good, bad and control reflects the number of privacy projection goals and vulnerabilities expressed in each policy. The Novartis.com policy contained more protection goals (23) than vulnerabilities (9), whereas the Drugstore.com policy contained more vulnerabilities (36) than protection goals (19). The HealthCentral.com policy served as the *control* factor because it contained the same number of vulnerabilities (12) as protection goals (12). The resulting experimental matrix is presented in Table 3.1.

Table 3.1 Experimental Treatments and Blocking Factors

Policy / Variant	*Policy*	*Goals*	*Categories*	*Goals in Policy*
Drugstore.com	Vulnerable / NL Policy	Vulnerable / Goals	Vulnerable / Categories	Vulnerable / NL Policy & Goals
HealthCentral.com	Control / NL Policy	Control / Goals	Control / Categories	Control / NL Policy & Goals
Novartis	Protective / NL Policy	Protective / Goals	Protective / Categories	Protective / NL Policy & Goals

When we refer to a policy in this experiment as *good* or *bad*, we mean only that the *good* policy reflects an organization that appears to be more trustworthy [than one with a bad policy] in that it contains more protection goals than vulnerabilities. In contrast, the *bad* policy contains more vulnerabilities than protection goals, and therefore reflects an enterprise that is inferior to the good policy in its articulated privacy practices from the general public's viewpoint. This classification does not take into consideration the reading difficulty (e.g. Flesch reading ease score [Flesch49]) or content organization, and it is possible that the organization with the *bad* policy is simply more forthright about their privacy practices than the organization with the *good* policy.

Each un-shaded cell in Table 3.1 represents one of the 12 possible policies (expressions) that could have been presented to subjects. Each column represents a different treatment, while each row represents a different blocking factor.

To prevent name-brand recognition bias from influencing the results of the experiment, all references to the names of the original policy authors were removed and replaced with "BrandX". To further prevent bias, subjects were randomly assigned to one of the 12 policy expressions. However, to maintain approximately the same number of responses for each of the 12 expressions, there was an attempt to balance the number of responses once they became unbalanced. This was accomplished by randomly assigning subjects to one of the 12 expressions until there were more responses for some expressions than others. For example, if the *Vulnerable/Goals* expression had several responses less than the other expressions, the next few subjects were assigned to the *Vulnerable/Goals* expression until the *Vulnerable/Goals* expression had an equivalent number of responses to the other expressions. Subsequently, subjects were again randomly assigned to one of the 12 expressions.

Before each policy expression, the following scenario was presented to the subjects:

> *Imagine you are in need of a medical service such as purchasing medications from an online pharmacy or retrieving medical information from a physician via the Internet. In this scenario, you are visiting BrandX's website on the Internet to obtain the healthcare related service you desire. However, in order to obtain this service, the website requires that you provide BrandX with some form of personally identifiable information (e.g., credit card number, address, phone number).*
>
> *In order to find out how your personally identifiable information (PII) will be collected, used, and transferred, you would need to read BrandX's privacy policy. A privacy policy is a document that contains the privacy practices of an institution, such as how the company collects and uses information, to whom they transfer information, how and when they may contact you, etc.*
>
> *We have provided you with BrandX's privacy policy below should you wish to consult it before deciding whether you wish to share your information with BrandX to obtain the service. Once you feel you have enough information to decide whether or not to share your PII with BrandX, you may proceed to the next step by clicking on the Proceed to Next Step link found at the bottom the page.*

Accompanying the scenario, before the expression of the privacy policy, subjects were presented with the necessary acronym definitions (e.g., PII, PHI, CI), keyword definitions (e.g., collect, guard, notify), and additional instructions for navigating the survey instrument.

Once subjects finished reading the policy, they were presented with a questionnaire based on the content of the policy. The question types can be classified into three categories: perception, comprehension, and demographic. The *perception* questions gauged how subjects felt about what they read and were presented in a 5-point Likert-scale form (i.e., Strongly Disagree to Strongly Agree). The *comprehension* questions measured how well subjects comprehended the content of the policy and were presented in the form of multiple choice quiz questions. The *demographic* questions measured the demographic makeup of the subject pool and were presented in the form of drop-down, multiple choice questions. A full list of the questions that were asked in the questionnaire is presented in Appendix A.

The same perception and demographic questions were asked of all subjects, regardless of which of the 12 policy expressions they received. The comprehension questions presented to subjects were determined by the policy, variant, and random chance as follows: For each policy, a single question was derived from each taxonomy category [AE04] if the policy contained a statement in this category. For example, if the Drugstore.com policy contained a statement that is categorized as an *information transfer* statement according to the privacy taxonomy, a question about the information transfer practices of Drugstore.com was included; otherwise, if no such policy statement existed, no question about the information transfer practices of Drugstore.com appeared. Each of these questions was accompanied by five possible answers, only one of which was correct. Each taxonomy category question was worded the same way for each policy. For example, the question "Which statement is TRUE regarding BrandX's information collection practices?" would be associated with the *information collection* category for each possible policy expression. However, since each policy (i.e., Drugstore.com, HealthCentral.com, Novartis.com) contained different content, the correct answer was different for each policy.

Asking subjects to answer a comprehension question for each category, in addition to the perception and demographic questions, would have placed too large a burden on the subjects. Instead, for each policy, the survey instrument randomly selected three questions from the set of all possible comprehension questions for that policy. Then for each variant within that policy, the instrument presented subjects with the same three questions. Once subjects in each variant responded to the three questions, the instrument randomly selected another three questions to present for each variant. For example, given the HealthCentral.com policy, the instrument randomly selected three questions from the set of HealthCentral.com comprehension questions. Let us assume that the instrument randomly chose *information transfer*, *notice/awareness*, and *information monitoring* as our categories. A single question would then be presented for each of these categories about the content of the HealthCentral.com policy. The instrument would present the same three questions to the next set of subjects who receive the *policy*, *goals*, *categories*, and *goals in policy* variants of the Healthcentral.com policy. Subsequently, the instrument would randomly select another three questions to be asked of each variant within the HealthCentral.com policy.

In addition to these three questions, subjects who receive the *categories* variant are also presented with a comprehension question based on the first and last category they chose to view, if these categories differed from the taxonomy categories of the original three questions that were selected.

3.2 Data Collection

Prior to distributing the survey, it was pilot tested using an online format. Based on some preliminary analysis and comments from the respondents, some items were deleted and others were reworded. The resulting instrument had seven scale items, eleven demographic items and between three and five others, depending on the variant being applied.

The final survey was linked from an NSF-sponsored website and was advertised to a variety of Internet users worldwide. Respondents were solicited through a variety of outlets, including links to the survey from various university web pages, general news sites, alumni mailing lists, professional mailing lists, email, electronic social networks, and word of mouth. To increase the variability in the data and generality of the survey, a marketing campaign was designed and launched to target all demographic audiences. Survey participants were offered an opportunity to enter into a prize drawing, to take place at the conclusion of the survey. The survey was available online from October 25, 2005 to December 10, 2005 via the Web at an NSF-sponsored project site.

The survey drew 1,215 total responses, but used some built-in mechanisms to distinguish between valid and invalid responses. Based on the pilot study, it was relatively certain that subjects could not read the directions, as well as the policy, in less than 30 seconds. Therefore, respondents who spent a combined total of 30 seconds or less reading the instructions and policy were eliminated, resulting in a reduction of 212 responses.

Additionally, 10 more responses were eliminated by analyzing answers to the perception questions: "*I believe BrandX will protect my personal information* ***more*** *than most other companies*" and "*I believe BrandX will protect my personal information* ***less*** *than most other companies*". These questions are binary opposites of one another and their answers should never be the same. If subjects gave the same answer to these two questions, the

subjects were assumed to not have carefully read the questions, and these responses were also eliminated from the dataset. This resulted in a total of 993 usable responses.

3.3 Results

This subsection presents the results of our experimental survey instrument. We conducted statistical analysis, including ANOVA, Tukey and Scheffe tests, to verify the statistical significance of all claims and observations made herein.

The remainder of this section is organized as follows. Section 3.3.1 presents the results related to the user perception questions. Section 3.3.2 presents the results related to the comprehension questions. Section 3.3.3 presents the results of other data collected during this experiment.

3.3.1 User Perception

The user perception questions were presented to the subjects in a 5-point Likert-scale form. Each answer was assigned a numeric value: strongly disagree = 1, disagree = 2, unsure = 3, agree = 4, and strongly agree = 5. Formulating the answers in this manner allows the average response for each question to be calculated. Comparisons using these values permitted us to make important observations about the ways in which the independent variables (e.g., policy, variant, and demographic variables) affected the dependent variables (e.g., comprehension and perception variables).

Users feel more secure with natural language policies.

When asked whether they feel secure sharing personal information with BrandX, after viewing their privacy policy, users tend to feel more secure with the natural language privacy policies than the privacy policies expressed using goal statements ($F_{98,894} = 7.69$; $p < 0.0001$). The average scores for each expression, including the average values for each variant are presented in Table 3.2 below.

Table 3.2 Average response to "I feel secure sharing my personal information with BrandX after viewing their privacy practices" for each policy representation. The table also includes the average values by policy (last column) and by variant (bottom row).

	Variant				
Policy	*Policy*	*Goals*	*Categories*	*Goals/Policy*	**Policy Averages**
Drugstore.com (bad)	2.59	2.13	2.48	2.83	**2.51**
Novartis (good)	2.75	2.72	2.64	2.79	**2.72**
HealthCentral (neutral)	2.8	2.48	2.62	2.93	**2.71**
Variant Averages	**2.72**	**2.44**	**2.58**	**2.85**	**2.65**

The scores listed in Table 3.2 illustrate how users associate the same degree of security between the policies that contain goals, as well as between the policies that are expressed in natural language.

Users feel less secure with the privacy policy that contains the most vulnerabilities.

A noteworthy observation in the experiment results ($F_{98,894} = 4.56$; $p < 0.01$) is that users felt less secure with the Drugstore.com privacy policy than the other two policies (see Table 3.3). This is encouraging because the Drugstore.com privacy policy contained the largest number of vulnerabilities, and also had a higher ratio of vulnerabilities to protection goals than the other two policies. This may indicate that users are, indeed, aware of how to distinguish between policies that protect their privacy and which ones may cause their privacy to become vulnerable to invasion or exploitation.

Different expressions had no affect on whether users felt the policy would protect their information more than other websites..

In this experiment, two questions that were expected to have binary opposite results were asked: "*I believe BrandX will protect my personal information* ***more*** *than most other*

companies" and "*I believe BrandX will protect my personal information **less** than most other companies*". The results) illustrate that neither the variant nor the policy had an affect on whether users felt more secure sharing information with the BrandX website than other websites (see Table 3.3). Users seem to feel equally uncertain about each variant ($F_{98,894}$ = 2.75; $p < 0.04$), as well as each different policy ($F_{98,894}$ = 3.63; $p < 0.03$). This seems to indicate that users may lack the tools and/or confidence necessary to make informed comparisons between different organizational privacy policies.

Table 3.3 Average response to "I believe BrandX will protect my personal information more than other companies" for each policy representation. The table also includes the average values by policy (last column) and by variant (bottom row).

	Variant				
Policy	Policy	Goals	Categories	Goals/Policy	**Policy Averages**
Drugstore.com (bad)	2.62	2.38	2.57	2.7	**2.57**
Novartis (good)	2.66	2.91	2.55	2.83	**2.74**
HealthCentral (neutral)	2.83	2.51	2.68	2.88	**2.73**
Variant Averages	**2.71**	**2.6**	**2.6**	**2.8**	**2.68**

Users feel that natural language privacy policies are explained more thoroughly than alternative expressions.

The results (see Table 3.4) illustrate that users feel natural language privacy policies are explained more thoroughly than goal-based privacy policies ($F_{98,894}$ = 19.41; $p < 0.0001$). This result was to be expected because natural language privacy policies are generally more verbose than goal-based policies. For example, Drugstore.com's natural language privacy policy contained 2,099 words compared to the 671 words of Drugstore.com's goal-based privacy policy. This seems to be further illustrated by the fact that users feel Drugstore.com's privacy policy (2,099 words) is explained more thoroughly than Novartis.com's privacy policy (1,350 words). Users seem to associate the length of the policy with its thoroughness, possibly to their own detriment. Associating policy length with thoroughness could be dangerous. Organizations may exploit this result by creating wordy privacy policies that contain unnecessary or irrelevant text that has nothing to do with the privacy practices of the institution. Users may be mislead into believing the organization is

thorough in its explanation of their privacy practices, thereby creating false trust in an otherwise deceptive organization.

Table 3.4 Average response to "I feel that BrandX's privacy practices are explained thoroughly in the policy I read" for each policy representation. The table also includes the average values by policy (last column) and by variant (bottom row).

	Variant				
Policy	Policy	Goals	Categories	Goals/Policy	**Policy Averages**
Drugstore.com (bad)	3.34	2.74	3.39	3.49	**3.24**
Novartis (good)	3.23	2.96	2.59	3.05	**2.96**
HealthCentral (neutral)	3.16	2.63	2.7	3.44	**2.99**
Variant Averages	**3.24**	**2.77**	**2.88**	**3.33**	**3.06**

Users feel confident in their understanding of categorized policies.

Despite not feeling secure sharing their information with companies that utilize goal-based policies (i.e., *categories* and *goals*), the results ($F_{98,894} = 6.47$; $p < 0.0002$; see Table 3.5) illustrate that users feel as confident in their understanding of the *category* policies as they do with either of the natural language policies (i.e., *policies* and *goals in policy*). It is interesting to note that although users felt less secure and protected with *category* policies, they felt as confident in what they read in the *category* policies than either of the natural language-based policies. This may indicate that users do, in fact, comprehend the policies better when they have been categorized a priori, and they notice that the policies of the organizations are not as secure as they would like.

Users did not feel as confident in what they read when presented the *goals* variant. We suspect that this is due to users being intimidated by a privacy policy presentation that they are unfamiliar with and is less than user-friendly. HCI issues may be to blame for this statistical observation.

Table 3.5 Average response to "I feel confident in my understanding of what I read of BrandX's privacy policy" for each policy representation. The table also includes the average values by policy (last column) and by variant (bottom row).

	Variant				
Policy	Policy	Goals	Categories	Goals/Policy	**Policy Averages**
Drugstore.com (bad)	3	2.71	3.31	3.17	**3.05**
Novartis (good)	3.34	3.01	3.38	3.22	**3.23**
HealthCentral (neutral)	3.44	3.11	2.97	3.55	**3.27**
Variant Averages	**3.26**	**2.95**	**3.22**	**3.3**	**3.19**

Users also felt the least confident with what they read in the Drugstore.com policy ($F_{98,894} = 4.81$; $p < 0.008$). This may be due to Drugstore.com having the longest policy, which may have intimidated users into believing they could not comprehend the policy in its entirety.

3.3.2 User Comprehension

The user comprehension questions were presented to subjects in a multiple choice format. For each user response, a score was calculated based on the number of questions the user answered correctly. The formula for the score was as follows: score = (questions correct / total questions asked) * 100. This yielded a numeric value between 0 and 100. This score was then used to calculate the average score by variant and policy.

The results ($F_{108,884} = 12.60$; $p < 0.0001$) clearly illustrate that the *policy* variant is the worst privacy policy variant for comprehension (see Table 3.6). In contrast, the *categories* variant was the easiest to comprehend. The difference between the *policy* and *goals in policy* variants is that, in the *goals in policy* variant, pop-ups are provided that contain the goals associated with each privacy-related statement in the policy. This simple addition of goal pop-ups resulted in a significant increase in comprehension scores (see Table 3.6). These results illustrate that goal-based policies are easier to comprehend than natural language policies. Furthermore, goal-based policies are easier to comprehend when they are presented in a categorized, easy to understand format.

Table 3.6 Comprehension Score by Variant (all responses)

Variant	Score
Policy	35.70
Goals	43.82
Categories	52.14
Goals in Policy	43.27
Overall Average	**43.74**

One of the questions presented to users within the questionnaire was: "Why didn't you read the entire privacy policies of the website?" One of the responses to this question was: "I read the entire privacy policies of the website". By eliminating all users who did not read the entire set of privacy policies, we could further analyze the readability of the policy variants.

The average score by variant for users who read the entire policy is presented in Table 3.7. Based on these scores, we made three important observations ($F_{11,521} = 18.27$; $p < 0.0001$). First, based solely on the quiz questions, it did not matter whether users read the entire policy or not – the *categories* variant was always the easiest to comprehend and the *policy* variant was always the most difficult to understand. Second, we observed a greater increase in comprehension in the goal-based policies (13.5% for *categories*, 11.6% increase for *goals*) than the increase in comprehension of the natural language based policies (4.3% for *policy*, 6.1% for *goals in policy*). Lastly, even among users who read the entire privacy policy, they answered only about half of the comprehension questions correctly.

Table 3.7 Comprehension Score by Variant (responses of subjects who read the entire policy)

Variant	Average
Policy	40.00
Goals	55.46
Categories	65.67
Goals in Policy	49.38
Average Overall	**52.88**

3.3.3 Other Observations

As mentioned in Section 3.3.2, we asked users, "Why didn't you read the entire privacy policies of the website?". The possible answers, along with the respondent results are presented in Table 3.8 below.

Table 3.8 Why Users Did Not Read The Policy (percentage by variant)

	By Variant				
	Policy	*Goals*	*Categories*	*Goals/Policy*	**Average**
The policies were too hard to understand	1.69	5.69	4.64	2.34	3.59
The policies were too long	35.86	34.15	26.16	42.97	34.23
The policies were not organized well	4.66	13.01	5.49	3.13	6.39
I have no interest in the privacy policies of institutions I share my personal information with	0.42	2.85	2.11	1.95	1.8
I read the entire set of privacy policies of the website	56.36	44.31	61.6	49.61	53.49

Each cell in Table 3.12 contains the percentage of users who gave the answer of its row when presented with a policy expressed in the variant of its column. In each variant, the majority of users read the entire set of privacy policies. The most common reason why users did not read the entire set of privacy policies was that the given policy was too long.

One important observation is that users most often read the entire set of privacy policies when given the *categories* variant. As a result, there was a smaller percentage of users who felt that the *categories* policies were too long.

The survey instrument kept track of the number of hits received on each page. Of the 2,608 visits to the policy pages, only 1,791 of these users proceeded to the survey page. Of the users that left the policy page, 15.08 percent were viewing the *policy* variant, 23.55 percent were viewing the *goals* variant, 30.17 percent were viewing the *categories* variant, and 31.2 percent were viewing the *goals in policy* variant. Based on these results, we may conclude that HCI factors may deter users from attempting to read privacy policy formats that they are unfamiliar with.

3.4 Discussion

The objective of this study was to glean valuable information about the user perception and comprehension of alternatives to natural language privacy policies. What we discovered was that there is a significant disparity between the user perceptions of the various policy expressions, as well as the user comprehension of the various policy expressions.

Users feel most secure and protected by natural language privacy policies.

The results presented in Section 3.3 illuminate several issues worthy of discussion. One of the more compelling observations is that users tend to feel more secure and protected by policies that are expressed using natural language than with goal-based policies. This may be due, at least in part, to the level of familiarity associated with natural language. It is unlikely that many of the subjects were familiar with goal-based policies, yet it is very likely that the majority, if not all, of the subjects were familiar with policies expressed using natural language. Subjects may have been wary of a policy format that they had not previously had any experience with, which would be reflected in the results. It may also be that it is not the format, but rather the unfamiliarity with the taxonomy of categories and the impersonal and harsh nature of the goals. This is supported by noting that even within a policy containing the same privacy statements (blocking factor), subjects were always more comfortable the natural language policies than with the goal-based policies.

The previous observation may be further supported by analyzing which policies were abandoned most often by users that did not proceed to the subsequent questionnaire. The results show that of the users who left the policy page, the *policy* variant was abandoned the least often. The implication is that there was less of a learning curve and a greater sense of familiarity with the natural language policy.

Users also felt more secure and protected by the *good* and *control* policies than they did with Drugstore.com's *bad* policy. From a consumer point of view, this is encouraging because it implies that users are able to identify which policies are most likely to violate their privacy. However, these results may discourage companies who would otherwise be

forthright about their privacy practices in their privacy policy. Since there is little legislation that govern privacy policy content outside of the healthcare and financial domains, companies may use these results to justify a lack of transparency in their privacy policies for fear users will identify their policy as a *bad* policy when they are simply trying to be honest about their practices.

Users comprehend the goal-based policies better than the natural language-based policies.

This study investigated how well users would comprehend alternative expressions to natural language privacy policies. What we discovered is that a simple natural language expression of a privacy policy was the most difficult to comprehend. In fact, users only answered a third of the privacy related questions correctly when given the *policy* variant.

The comprehension of the natural language privacy policy increased when privacy statements were highlighted within the policy and accompanied by goal statements, as in the *goals in policy* variant. This illustrates that the natural language policy alone may not be sufficient for conveying policy information to the average Internet user.

The comprehension scores increased when users were presented with the policy expressed as a list of goal statements, as in the *goals* and *categories* variants. This is fairly intuitive, given that goal statements are uniform and eliminate extraneous and unnecessary information. For example, in the Drugstore.com policy, the *goals* variant contained one fourth of the number of words contained in the *policy* variant. With less information to read and try to retain, users do not experience information overload and can retain essential information regarding the uses and disclosures of their PII.

The comprehension scores were the highest for the *categories* variant. This policy variant organizes the goal statements found in the *goals* variant into the privacy taxonomy category it belongs to. This variant presents users with a list of the different categories and allows them to choose which category of policies they wish to view.

Since this variant is comprised of the same goals that are found in the *goals* variant, we expected to see similar comprehension results. Instead we found that the comprehension of the policies actually increased when they were categorized. This may suggest that users

will retain information better when it is presented to them in an organized manner and within categories that they may associate with the policies.

Demographics had little influence on the results of the study.

The results of the study also verify a common stereotype – that older generations are not as technologically savvy as younger generations. All age groups scored the same on the comprehension quiz questions, with the exception of users ages 57 and higher ($F_{108,884} = 2.5$; $p < 0.008$). This group received lower comprehension scores than the other age groups.

After analyzing the data, we conclude that demographics made no difference in the user perception of the various variants and policies with which they were presented. Furthermore, with the exception of a single age group, demographics had no affect on the comprehension scores of the various variants and policies.

The analysis of the demographic data provided additional insight into the makeup of the subject pool. The subject pool did not precisely represent the Internet population in that 75.38 percent of the subjects had at least a bachelor's degree or higher, yet we know that only 26.9 percent of the general population over 25 years old has obtained this level of education [NTI02]. This means that our subject pool was more educated than the average Internet community. Moreover, 24.82 percent of the subjects listed their occupation as *Information Technology*.

Given these observations, one might expect this subject pool to be familiar with online privacy policies and their esoteric language and terminology. As a result, we might also reasonably anticipate this subject pool to receive higher comprehension scores than the average Internet user. Following this logic, the low comprehension scores observed in this study are even more disconcerting because they demonstrate just how difficult these policies would be for the average user to comprehend.

A complete list of demographic frequencies is included in Appendix B.

User perception was not in alignment with user comprehension.

Though users felt secure with and protected by the natural language privacy policies, their comprehension of the content was poor. Even among users who read the entire policy,

users who read the *policy* variant only answered a little over a third of the questions correctly. However, despite not feeling as secure or protected by the goal-based policies, user comprehension scores were much greater when given goal-based policies.

This misalignment of user perception with user comprehension is disconcerting because users may be more inclined to trust a company with a policy that lacks clarity and readability. This leads to less informed decisions that could result in the increase of unanticipated and unwanted uses and disclosures. Deceitful organizations could even exploit these results to garner user confidence and then proceed to violate their privacy.

It is worth noting that even though users perceived Novartis.com's *good* policy to be less secure and provide less protection than Drugstore.com's *bad* policy, based on the results of this study, the *good* policy was the easiest to comprehend. This is most likely due to its length being much less than the *bad* policy, preventing information overload and allowing users to absorb more of the policy at once.

Users seem to prefer longer policies. This seems to suggest that they equate quantity with quality. However, as our results show, this equivalence does not exist and this assumption by users may, once again, lead them to trust an organization without being fully aware of how their information will be used and disclosed.

Users are inclined to read content in the order in which it is presented.

Given a simple natural-language privacy policy, it is difficult to determine which portions of the policy users are most interested in reading. However, we were able to track this information when subjects were given the *categories* variant. Since goals were designated into their appropriate privacy taxonomy page, we tracked when a subject chose to view a categories' goals. By doing so, we were able to observe the order in which users prioritize the information they are interested in.

The subjects were presented with a list of the privacy taxonomy categories for which goals existed for that particular policy. For example, if there existed a goal in the policy related to information monitoring, we would present the subject with the opportunity to click on the *Information Monitoring* category to view a list of goals classified by this category. If no *information monitoring* goals existed in the policy, the subject was not given the

opportunity to choose this category. Since each policy contained different goals, the list presented to each user was different and depended on which policy they were given. The list of categories was presented to subjects in alphabetical order (e.g., *contact* before *information transfer*). Each category was also accompanied by a description of the types of goals it included.

We found that, despite given the opportunity, the majority of subjects did not prioritize the information they are interested in. Table 3.10 below outlines the order with which subjects chose to read the policy. Each row represents the frequency in which subjects chose that particular category. For example, 79.61 percent of the time, the first category subjects chose to click on was the first category in the list the they were presented with; 81.5 percent of the time, the second category subjects clicked on was the second category in the list they were presented with; and 83.7 percent of the time, the seventh category subjects clicked on was the seventh category in the list that they were presented with.

Table 3.9 Category Viewing Frequencies: The percentage of users who clicked on the category presented in the order that they were presented.

Order	Percentage of Users
1	79.61
2	81.5
3	78.9
4	80.3
5	81.9
6	82
7	83.7
8	90.9
9	88.24
10	96.77

These results illustrate that, despite the option to view information relevant to specific categories of policies, users chose, instead, to view the policy in the order it was presented to them. While this may seem intuitive and to be expected, this study confirms that users will generally read a policy from top to bottom.

The implications of this confirmation can be to the detriment of consumers who do not read an organization's privacy policy in its entirety. As this study has also shown, even when users do read privacy policies, a great many of them do not read it in its entirety. If

organizations choose to exploit this fact, they may concentrate the most important and privacy-relevant information towards the end of their privacy policy. This would impair users' ability to make informed decisions about sharing information with the organization.

The increase in percentage correlating with the order is due to the fact that there are more users that chose to view only one category than there are users who chose to view two categories, more users that chose to view two than three, and so on. These percentages demonstrate that the users who chose to view upwards of ten categories are, in all likelihood, the same users who chose to view the categories in the order with which they appear. So, while they may be viewing the categories in order, they are at least viewing an increasing amount of policy. This indicates that the users who do not read the entire policy are viewing the policy out of order, thus they are the users that are likely to prioritize the information they read when given the option. This also indicates that if these same users were presented with a regular natural language privacy policy, they most likely would try and scan the document for the privacy-relevant information they are interested in rather than read the entire document. However, due to the difficulty of parsing natural language privacy policies, it is unlikely the user would find all of the information relevant to the categories of interest. As a result, they may not have all the information necessary to make an informed decision.

Users spent less time reading goal-based policies than natural language policies.

Users spent less time reading the *goals* variant than either the *policy* or *goals in policy* variants, despite comprehending the *goals* variant better ($F_{11,974} = 4.69$; $p<0.003$). Users spent, on average, 210 seconds reading the *goals* variant, while they spent 269 seconds reading the *policy* variant and 288 seconds reading the *goals in policy* variant. This result substantiates the claim that goal-based policies are easier to read and comprehend than the existing natural language policies.

Users spend 241 seconds, on average, reading the *categories* variant. There is no statistically significant difference between the *categories* variant and any of the other variants. This may be because users are presented with a description of each taxonomy category, in addition to the list of goals to read.

The results of this study can pave the way for machine readable privacy policies that users can also comprehend.

There has been significant research into the development of machine-readable privacy policies that can be parsed at runtime and checked against user specified privacy preferences [SDP05, CLM02, EPAL06]. Such technology would allow users to specify how they wish their information to be used and disclosed. Then, while browsing the web, users could be notified when they visit a web site where the privacy practices do not align with their specified privacy preferences.

From an information assurance perspective, this technology would allow users to make informed decisions about who they share their information with, and also ensure that they do not disclose their PII to organizations that do not comply with the user's privacy preferences.

Unfortunately, implementations of this technology, such as P3P, have not been widely accepted by the web community at large. One reason for this is the ambiguous nature of P3P. In fact, a quote from CitiGroup's position paper [Citi02] states, "the same P3P policy could be represented to users in ways that may be counter to each other as well as to the intent of the site . . . This results in legal and media risk for companies implementing P3P that needs to be addressed and resolved if P3P is to fulfill a very important need." Furthermore, one report stated that P3P does not adequately address user privacy [Epic00]. No P3P successor will be satisfactory unless it is unambiguous, indisputable, and accepted by both online entities and Internet users.

It is essential that any machine-readable policy implementation be free of ambiguity. The mere existence of ambiguities would allow organizations to exploit the ambiguities in their policies in order to avoid legal action or other penalties when they violate their policy. Users require an indisputable policy in order to protect their privacy. Moreover, it would be in an organization's best interest to use an implementation that is free of ambiguities to avoid misinterpretation and resulting legal action.

Another reason that current implementations of machine-readable privacy policies are insufficient is because they need to better account for the *purposes* and *instruments* of privacy statements. For example, a user may not mind their information being disclosed *for*

the purpose of treatment, but they may not want their information disclosed *for the purpose of marketing*. Since there are an infinite number of possible purposes for which actions may be taken, users could not enumerate them all in their privacy preferences.

So, what happens when a user is faced with a machine-readable privacy policy that contains a purpose that they have not specified in their privacy preferences? And how could they agree to the policy if it is only machine readable? The results of our study answer this question. We have proven that users not only understand goal-based policies as well as natural language policies, in the proper context, they comprehend them better.

The goals we used are structured in a form called a restricted natural language statement (RNLS). Prior research [BA05a, BA05b, BA05c] has illustrated how RNLS(s) can be parameterized and modeled using the Knowledge Transformation Language (KTL) [BA05b]. After an RNLS has been parameterized and modeled in KTL using a process called Semantic Parameterization [BA05a, BA05b], it becomes machine-readable. At this point, comparisons can be made between RNLS(s), ambiguities can be identified, and queries can be utilized to identify information within statements [BVA06]. Semantic models are also able to identify and classify *purposes* and *instruments*.

If policies were expressed using these parameterized models, exceptions can be identified and presented to the user in a way that they can comprehend. For example, if an exception occurred because there exists a purpose that the user had not previously accounted for, the original policy RNLS can be reconstructed from the model and presented to the user. This should prove effective since, as a result of this study, we now know that users would be able to comprehend the RNLS, since the goals used in our study are in RNLS form, well enough to make an informed decision about whether this policy meets their privacy preferences.

3.5 Limitations and Future Work

In this study, we sought to understand how average Internet users perceive and comprehend various alternatives to natural language privacy policies. Despite our best efforts to target a subject pool that precisely represents the Internet community, the subject pool was more educated and technologically savvy than the average Internet user. This was

reflected in the data collected during the experiment. However, we were still able to draw very important conclusions as to the perception and comprehension of alternatives to natural language privacy policies.

As with any survey, there was concern that subjects would not be completely honest in their responses. Several measures were taken to avoid incorporating these user's responses into the respondent dataset, including: preventing users from revisiting the policy to look up answers to comprehension questions; requiring that the questionnaires be completed before submission; ensuring the anonymity of subjects would be preserved; and identifying and removing responses from the dataset that were identified as being invalid.

As a result of this study, we discovered that there is a serious disparity between user perception of the various privacy policy expressions and how well they comprehend each of the various policies. Even though users felt more secure with, protected by, and comfortable with the natural language privacy policies, they did not comprehend them as well as the goal-based privacy policies. Now that we know that users comprehend goal-based policies better than natural language policies, researchers need to find ways to facilitate users feeling comfortable with them.

The disparity between user perception and comprehension may be due to Human Computer Interaction (HCI) factors, in which the users are simply not comfortable with the manner in which the goal-based policies are presented to them. To support this claim, one need only note the marked improvement in comprehension scores between the *goals* variant and the *categories* variant. The categories variant does nothing more than present the goal statements to the user in an organized fashion. If research efforts were invested in addressing the HCI issues surrounding these policies, the misalignment between user perception and user comprehension may be rectified.

There are many implications for the development of machine-readable policies that are associated with the results of this study. As previously discussed in Section 3.4, the goals in the goal-based policies are a relaxed form of restricted natural language statements (RNLS). RNLS(s) can be parameterized and modeled in such a way that they become machine-readable. If research were invested in the reconstruction of RNLS(s) from models, as well as the specification of user preferences and the runtime interpretation of these

semantic models, we might very well accomplish machine-readable, user-comprehendible policies.

One of the discoveries of this study was that users tend to read policies in the order in which they are presented. Given that this study also illustrates that many users, when presented with a privacy policy, do not read it in its entirety, we suggest that research be conducted to discover whether organizations are already exploiting this fact. If the most important and privacy-relevant information is being discussed near the end of privacy policies, users may be at a serious disadvantage when trying to make an informed decision about whether to share their information with a given organization.

Chapter 4

"All sorts of computer errors are now turning up. You'd be surprised to know the number of doctors who claim they are treating pregnant men."
- *Isaac Asimov*

ANALYSIS OF HIPAA COMPLIANCE

In the United States, federal and state regulations prescribe stakeholder rights and obligations that must be satisfied in order for organizations to be compliant with legislation. Failing to be in compliance with such legislation can result in severe, and costly, penalties. For example, violating the HIPAA by obtaining and disclosing personal health information (PHI) for commercial advantage, personal gain, or harm carries a potential $250,000 in fines and a ten year prison term [FA03]. Even simple non-compliance with the HIPAA can cost an organization $25,000 annually.

Legislation governing privacy is being increasingly introduced at both the federal and state levels. In addition to the HIPAA, Children's Online Privacy Protection Act (COPPA), and Gramm-Leach-Bliley Act (GLBA), the Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM), the Fair Credit Reporting Act, the national Do Not Call Registry, and the Freedom of Information Act (FOIA) are all examples of recent federal efforts to protect the rights and privacy of American citizens. Furthermore, many states have enacted their own privacy protection laws in response to the increase in consumer concern.

Unfortunately for organizations governed by newly introduced legislation, new legislation creates additional requirements. In order to ensure that an organization is compliant, one must be able to parse both the legislative texts, as well as the policies of the organizations subject to the legislation. By doing so, a comparative analysis can be

conducted, either by the organization or compliance and government officials, to ensure that the organization is in compliance with legislation.

In Chapter 2 of this thesis, we explored how the privacy policies of healthcare organizations evolved with the introduction of the Health Insurance Portability and Accountability Act (HIPAA). What Chapter 2 did not address was whether or not the policies expressed in these privacy documents were compliant with the HIPAA. In this chapter, we present a methodology for extracting rights and obligations from legislative text, our results from applying the methodology to the HIPAA Privacy Rule, and a preliminary comparative analysis of the set of rights and obligations that were obtained from the Rule to goals that were extracted during our prior analysis, discussed in Chapter 2 of this thesis.

This chapter presents the results of our analysis as follows. Section 4.1 presents a list of keys terms that we use throughout this chapter. Section 4.2 outlines the methodology we used to extract rights and obligations from regulatory text. Section 4.3 presents the results of applying this methodology to the HIPAA Privacy Rule. Section 4.4 illustrates how the results of applying the methodology can be used to analyze regulatory compliance. Section 4.5 discussions the limitations of our work and discusses future work.

4.1 Terminology

We define the following key terms:

- A *stakeholder* is an entity afforded rights and/or obligations by the HIPAA Privacy Rule.
- A *right* is an action that a stakeholder is conditionally permitted to perform.
- An *obligation* is an action that a stakeholder is conditionally required to perform.
- A *rule statement* is the regulation text that includes the right or obligation and any constraints.
- A *constraint phrase* is the part of a rule statement that describes a single precondition.
- A *normative phrase* contains words that indicate what “ought to be” as rights or obligations.

4.2 Methodology for Analyzing Regulations

In this study, we adopt a process called Semantic Parameterization, in which rights and obligations from regulation texts are restated into restricted natural language statements (RNLS), to describe discrete activities [4]. RNLS(s) can be mapped into semantic models that are amenable to formal analysis [3, 4] for several purposes, including reasoning about conditional relationships. Semantic Parameterization was developed using Grounded Theory, in which theory that is systematically obtained from a dataset is valid for that dataset [11].

Our methodology begins by extracting rule statements using a relaxed form of Semantic Parameterization. The relaxed form uses only two RNLS patterns to separate the right or obligation phrase(s) from relevant constraint phrase(s). Constraint phrase(s) restrict the scope of actors and objects that already appear in the right or obligation phrase. Separated constraint phrase(s) are used to construct pre-conditions in the form of logical expressions.

The relaxed form of Semantic Parameterization uses only two RNLS patterns: (1) activities that distinguish subjects and objects [3, 4]; and 2) activities following condition keywords (*if, unless, except*) identified in an earlier pilot study [5].

In order to extract rights and obligations, we first identify the normative phrase of a statement. The normative phrase indicates what a stakeholder is *required* or *permitted* to do. Extracting rights and obligations based solely on the normative phrase helps us to distinguish between *expressed* and *implied* rights and obligations.

Implied rights and obligations are those that are not explicitly delegated by the Rule. For example, consider unrestricted natural language:

> **UNLS$_1$**: A covered entity ***must permit*** an individual to request access to their protected health information.

We first identify that UNLS$_1$ contains an expressed obligation, indicated by the normative phrase “must permit.” Notice, however, that UNLS$_1$ implies that the individual has a right to request access to their protected health information.

To say that a party has a right is a way of talking about the counterparty’s implied obligation. For example, if a healthcare patient has a right to access their health records, then their physician’s office has an obligation to provide access. It is necessary to identify

implied rights and obligations to increase requirements coverage, since implied obligations derived from rights may be operationalized as requirements. Reformulating rights in terms of the implied rights and obligations of counterparties will enable stakeholders and compliance officers to ensure full compliance with regulation requirements. However, for the purposes of study, we forego the identification of implied rights and obligations until we perform the comparative analysis.

Consider the unrestricted natural language statement $UNLS_1$ summarized from §164.522(a)(1)(iii) in the HIPAA Privacy Rule and parameterized as RNLS(s):

> **$UNLS_2$**: A covered entity that agrees to a restriction ***may not*** use OR disclose protected health information, ***except if*** the individual who requested the restriction is in need of emergency treatment.
>
> **$UNLS_3$**: A covered entity that agrees to a restriction ***may not*** use AND disclose protected health information, ***except if*** the individual who requested the restriction is in need of emergency treatment
>
> **$RNLS_1$**: The covered entity who (**$RNLS_2$**) ***may not*** disclose protected health information, ***except if*** (**$RNLS_3$**).
> **$RNLS_2$**: The covered entity agrees to a restriction.
> **$RNLS_3$**: The individual who ($RNLS_4$) needs emergency treatment.
> **$RNLS_4$**: The individual requests the restriction.

RNLS(s) are restricted to expressing one discrete activity. However, $UNLS_2$ describes two separate activities, *use* and *disclose*; the analyst must identify that, though there is a single normative phrase (*may not*), the disjunction (or) indicates two separate activities: *may not use* and *may not disclose*. The English conjunctions (and, or) are not necessarily equivalent to logical-and and logical-or. For example, $UNLS_3$ is identical to $UNLS_2$, except that $UNLS_3$ uses the English conjunction "and" instead of "or." The analyst would need to decide whether, in the context of the text, to interpret this statement as a single obligation (*may not use and disclose*) or as two separate, discrete obligations (*may not use* and *may not disclose*). The burden is on the analyst to determine whether or not the English conjunction is dependent (logical-and) or independent (logical-or).

Constraints on obligations are rights are identified by observing condition keywords (e.g. *if*, *unless*, *when*). RNLS(s) are separate constraints on a right or obligation if they distinguish subjects and objects (e.g., $RNLS_2$ and $RNLS_4$) or follow condition keywords (e.g., $RNLS_3$). Consider $RNLS_1$, where $RNLS_2$ distinguishes the covered entity as one who

agrees to a restriction, whereas $RNLS_3$ distinguishes the individual as one who requests a restricting by using the nested $RNLS_4$. We do not classify other RNLS(s) such as transitive verbs followed by verb phrases, instruments or purposes [BA05a, BA05b] as separate constraints. Rather, these phrases remain nested in the right, obligation or constraint statement for the purpose of this study.

Regulation texts are often highly segregated and indexed by stakeholder and process, and the text is loaded with references to subsections within the text or external references to other legislation. In order to maintain traceability, the rights, obligations, and constraints extracted from the text are indexed by their corresponding subsection in the original text. Constraints are then organized into a logical expression based on their dependence. If constraints are independent of one another when invoking the right or obligation, they are in a disjunction; otherwise, they are dependent and are in a conjunction. The RNLS(s) from the example above are indexed and organized as follows:

Constraints:

A. The covered entity agrees to a restriction. 164.522 (a)(1)(iii)
B. The individual needs emergency medical treatment. 164.522 (a)(1)(iii)
C. The individual requests a restriction. 164.522 (a)(1)(iii)

Obligation:

1. The covered entity **may not** disclose protected health information. 164.522 (a)(1)(iii) [$A \wedge \neg B \wedge C$]

The constraints *A*, *B*, and *C* were derived from $RNLS_2$, $RNLS_3$, and $RNLS_4$, respectively. The obligation and each constraint are indexed by and labeled with their original subsection in the original regulation text. The obligation is labeled with the logical expression, in this case a conjunction, in square brackets. The keywords “except if,” in the original text from which this constraint was extracted indicates that this constraint is an exception. The exception is handled by negating the constraint in the conjunction.

Regulation texts tend to be laden with cross-references to different sections within the given regulation. Analysts must follow each cross-reference to evaluate the impact of the referenced content on each right or obligation statement. Organizations can use the indices to help maintain traceability as constraints are incorporated from different sections.

4.3 Analyzing Healthcare Regulations

Privacy is primarily concerned with preserving the confidentiality of information, but users are also concerned with the integrity and availability of their information. In the context of the Rule, we strove to investigate the various aspects of privacy, such as uses and disclosures, notice and awareness, and access to information. For this reason, we focused on the following nine sections in the Rule:

§164.506: Consent for uses or disclosures to carry out treatment, payment, and health care operations.
§164.508: Uses and disclosures for which an authorization is required.
§164.510: Uses and disclosures requiring an opportunity for the individual to agree or to object.
§164.512: Uses and disclosures for which consent, an authorization, or opportunity to agree or object is not required.
§164.514: Other requirements relating to uses and disclosures of protected health information.
§164.520: Notice of privacy practices for protected health information.
§164.522: Rights to request privacy protection for protected health information.
§164.524: Access of individuals to protected health information.
§164.526: Amendment of protected health information.

The Rule is comprised of two parts, numbered 160 and 164, and contains a total of 33,500 words. The nine sections we analyzed contain a total of 17,728 words or 52.9% of the Rule. Each part is sub-divided into subparts and each subpart is divided into sections. Sections §160.103 and §164.503 contain definitions necessary to distinguish the entities governed by the Rule. From the definitions, stakeholders can be organized into a class hierarchy. The class hierarchy in Figure 4.1 shows the stakeholders from the nine Rule sections we analyzed. The arrows in the hierarchy represent sub-class relationships; for example, group health plans (GHP) and health maintenance organizations (HMO) are both types of health plans (HP), and an HP is a type of covered entity (CE).

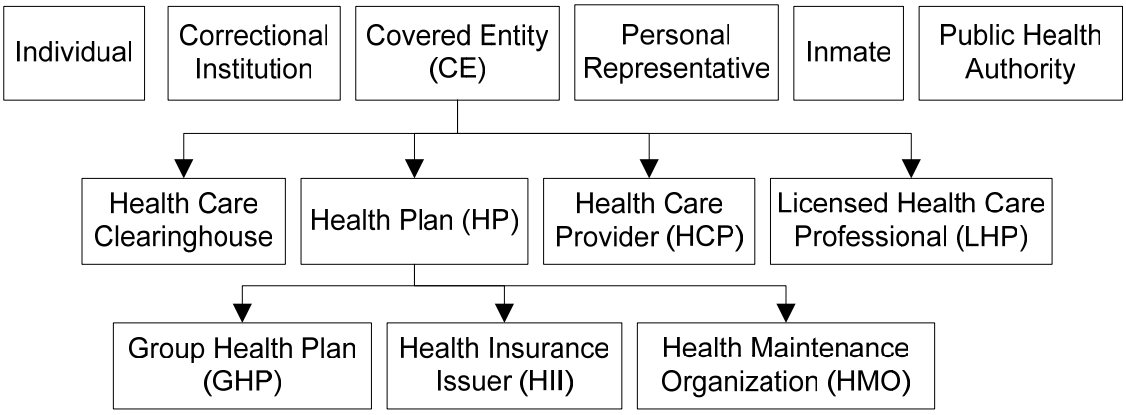

Figure 4.1 Stakeholder Class Hierarchy

In addition to §160.500 on Applicability, the definitions and corresponding class hierarchy are important when deciding which sections of the Rule apply to which stakeholders.

Each Rule section is divided into sub-sections that contain standards and implementation specifications, often with a separate emphasis on unique stakeholders. All of the sub-sections we analyzed were deemed to contain potential compliance requirements. The sub-sections are also sub-divided in ways that balance between conciseness and readability. Below, we summarize text from §164.520(c)(2) and (c)(3) as a model standard we simply index as (a) to illustrate the effect of applying the methodology to the regulation text. The normative phrase (must) and condition words (if, unless) are **bold**, the constraint phrases are underlined and the obligation phrases are *italicized*.

> (1) *A covered entity* who has a direct treatment relationship with an individual ***must*** …
>
> (A) *Provide notice* no later than the first service delivery;
>
> (B) **If** the covered entity maintains a physical delivery site:
>
> *i. Have the notice available for individuals to take.*
>
> *ii. Post the notice in a clear and prominent location.*
>
> (2) For the purposes of paragraph (a)(1), *a covered entity* who delivers service electronically ***must*** *provide electronic notice* **unless** the individual requests to receive a paper notice.

Applying our methodology, we derive constraints A–E and obligations 1–4, below. The *italicized* phrases in obligations 2, 4, and 5 are ambiguities resolved using the regulation text.

Constraints:

A. The CE has a direct treatment relationship with the individual. (a)(1)
B. The notice is provided no later than the first service delivery. (a)(1)(A)
C. The CE maintains a physical delivery site. (a)(1)(B)
D. The CE delivers service electronically. (a)(2)
E. The individual requests to receive a paper notice. (a)(2)

Obligations:

1. The CE must provide notice to the individual (a).
2. The CE must provide notice *to the individual.* (a)(1)(A) [A ∧ B]
3. The CE must have the notice available for individuals to take. (a)(1)(B)(i) [A ∧ C]
4. The CE must post the notice in a clear and prominent location *for the individual to read*. (a)(1)(B)(ii) [A ∧ C]
5. The CE must provide electronic notice *to the individual.* (a)(2) [A ∧ B ∧ D ∧ ¬E]

Right, obligation and constraint statements are often distributed across several sub-sections in the Rule. For example, the subject *CE* is specified in sub-section (a)(1); however, the obligation phrases *provide notice*, *have notice available*, and *post the notice* each appear separately in sub-sections (a)(1)(A), (a)(1)(B)(*i*), and (a)(1)(B)(*ii*), respectively. The constraint *A* appearing in sub-section (a)(1) is applied across each of these obligations as well as the obligation in sub-section (a)(2) due to a cross-reference back to (a)(1). Cross-references to other sections in the Rule pose the greatest challenge to analysts, since each section is written from a different viewpoint; this makes the relevance of constraints from other sections subtle and uncertain.

4.3.1 Analysis Results from HIPAA

We identified 193 rights and 145 obligations in §164.506–§164.516 and §164.520–§164.526. Table 1 summarizes the total number of rights ***(R)***, obligations ***(O)***, constraints ***(C)*** and cross-references ***(CR)*** identified per section. From a compliance perspective, each constraint within a system introduces complexity that makes compliance checking more difficult.

Table 4.1 Number of Rights, Obligations, and Constraints in HIPAA §164.506–§164.526

Section	*R*	*O*	*C*	*CR*
164.506	24	12	25	13
164.508	6	5	15	42
164.510	44	8	12	16
164.512	67	10	154	59
164.514	7	35	48	21
164.520	9	17	54	37
164.522	7	19	24	9
164.524	20	26	69	29
164.526	10	18	36	23

Table 2 summarizes the total number of unique normative phrases ***(N)*** we identified as well as the modality. In the table, *anti-rights* refer to activities the regulation explicitly exempts from stakeholder rights, but does not require the stakeholder to avoid (e.g., does not have a right to*).* Similarly, *anti-obligations* are activities that the regulation explicitly exempts from stakeholder obligations, but does not require the stakeholder to avoid (e.g., is not required to). We distinguish both anti-rights and anti-obligations from activities that are disallowed, such as obligations using the phrase "may not." Also in Table 2, normative phrases with an asterisk (*) indicate rights and obligations assigned through delegation. In delegation, a stakeholder is permitted or required to assign other stakeholders specific rights and obligations.

Table 4.2 Normative Phrases in HIPAA Sections §164.506–§164.526

Phrase	***N***	***Modality***
does not have a right to	1	Anti-Right
has a right to	9	Right
is permitted to	1	Right
is not required to	5	Anti-Obligation
may	173	Right
may deny*	3	Right
may not	26	Obligation
may not require*	1	Obligation
may require*	6	Right
must	100	Obligation
must deny*	1	Obligation
must permit*	16	Obligation
must request*	1	Obligation
retains the right to	1	Right

4.4 Comparative Analysis

Once rights and obligations have been extracted from regulatory texts, a comparative analysis can be performed by the organization to determine whether or not an organization is in compliance with regulations. Organizations can identify and resolve ambiguities, as well as balance *expressed* rights and obligations to obtain *implied* obligations [BVA06].

Unfortunately, a comparative analysis performed by external stakeholders (e.g., compliance officers and government officials) would not be feasible to ensure compliance by organizations. This is due to the large number of the constraints governing the rights and obligations. Since a stakeholder is not obligated until the constraints the of that obligation have been satisfied, it would be difficult, if not impossible, for external parties to determine which constraints have been satisfied, and therefore, which obligations an organization must fulfill. Furthermore, in the case of *implied* obligations, an organization is only obligated when an *expressed* right has been exercised. Since rights are also subject to having constraints placed on them, an organization would not be assigned an *implied* obligation until the constraints of the *expressed* right had been satisfied, and subsequently the right had been

assigned to a stakeholder. For example, consider the following right extracted from the Privacy Rule, along with its constraints, as well as a goal extracted from an AETNA policy:

Constraints:

A. The CE has an indirect treatment relationship with the individual. (a)(2)(i)
B. The individual is an inmate. (a)(2)(ii)
C. The HCP created or received the PHI in the course of providing health care to the individual (a)(2)(ii)

$R_{6.1}$: The HCP may use or disclose PHI to carry out treatment, payment, or health care operations (without consent). [A $\wedge$ B] or [A $\wedge$ C]

G_{1383}: DISCLOSE PII to business associates of covered entities for payment.

Goal G_{1383} states that AETNA will disclose PII (a subset of PHI) to their business associates for the purpose of performing payment services. Right $R_{6.1}$ extracted from the Privacy Rule states that PHI can be disclosed to carry out payment services. However, the constraints [A $\wedge$ B] or [A $\wedge$ C] must be satisfied in order for right $R_{6.1}$ to be assigned to the HCP. Only AETNA would be able to determine whether constraints A or C have been satisfied, so an external party would not be able to determine whether AETNA had been assigned right $R_{6.1}$.

Though external entities cannot ensure full compliance with regulations, a comparative analysis can be conducted by entities outside of the organization to detect *potential* non-compliance. In the context of this study, one could determine the *potential* obligations of an organization by observing the rights and obligations extracted from the Rule and comparing them to the privacy policies of the institution.

Since we already analyzed the policies of various healthcare organizations in Chapter 2 of this thesis, we compare these results to the rights and obligations extracted from applying our methodology to the HIPAA Privacy Rule. Consider right $R_{6.1}$ and goal G_{1381} from the previous example. Notice that the right is only assigned to a health care provider (HCP); However, AETNA is a not a HCP, but rather a health plan (HP). This indicates a potential conflict. Upon further evaluation of the rights and obligations extracted, we notice the following relevant right, along with its constraints.

Constraints:

A. The individual is informed in advance of the use or disclosure. (510)
B. The individual has the opportunity to agree to or prohibit or restrict the disclosure. (510)

$R_{10.1}$: The CE may use or disclose PHI without the written consent or authorization of the individual. [A $\wedge$ B]

In this example, since a HP is a type of CE, this right would be assigned to AETNA upon satisfaction of constraints *A* and *B*. The challenge, as an external party, is to determine whether or not the constraints have been satisfied. In this case, upon evaluating the rest of the goals in the policy, it appears the individual has not been provided an opportunity to restrict or prohibit the disclosure intended for payment services in goal G_{1381}. This may be identified as a potential conflict.

4.5 Discussion and Future Work

The objective of this research was to develop a methodology for extracting rights and obligations from regulatory texts in order to investigate compliance with legislation. We conducted a case study, using the HIPAA Privacy Rule, to extract rights and obligations of healthcare stakeholders. Additionally, we illustrate how a comparative analysis can be conducted, using the extracted rights and obligations, to check for compliance by organizations and for *potential* non-compliance by external parties.

For the purposes of this study, the comparative analysis proved sufficient for analyzing simple rules. However, we foresee this process becoming more cumbersome for anything larger than a simple compliance check. Given that there were total of 193 rights, 145 obligations, and 437 constraints, determining which constraints have been satisfied to assign rights and obligations may prove to be a difficult task, even for such a small portion of legislation. For larger legislative texts, the task may prove impossibly daunting. For these reasons, we propose further investigations into the automation of the compliance checking and compliance monitoring tasks, in order to extend our methodology to be able to handle large regulatory texts.

We observed that the legislative text often contained ambiguous language that requires further refinement. For example, there were 16 mentions of “professional

judgment" and 17 uses of the term "reasonably," yet no definition for either of these terms was provided. The inherent ambiguities in these terms create difficulty in the task of compliance checking. How can a system measure or implement "reasonably"? If you make any decision in the capacity of your profession, is it considered "professional judgment"? And is there a distinction between *good* and *bad* professional judgment? These are questions that are introduced when using such ambiguous vernacular. They are difficult to answer because, from the point of view of the stakeholder, answering them incorrectly can result in non-compliance. Legislators need to begin using unambiguous terminology or methods to formalize their meaning are needed to avoid non-compliance.

It is reasonable to expect that our methodology can be generally applied to legislation, other than HIPAA, to ensure compliance. However, the methodology is also expected to continue expanding. Though we only identified 14 separate phrase heuristics, we expect that this list will evolve to incorporate new phrases as additional legislative texts are analyzed.

In summary, we have formalized a methodology for extracting the rights and obligations of various stakeholders from regulatory texts and law. These rights and obligations can be used to reason about conditional relationships, clarify ambiguities, as well as performing a comparative analysis to determine compliance. However, further research needs to be conducted to automate the process of compliance checking and compliance monitoring for large-scale systems.

Chapter 5

"A conclusion is the place where you got tired of thinking."
- *Arthur Bloch*

CONCLUSION

The objective of this thesis is to advance privacy preservation and legal compliance in software-based healthcare systems. To this end, we conducted three separate studies: (1) evaluation of the evolution of privacy policies in the presence of the HIPAA; (2) an experiment, using an empirical survey instrument, to gauge user perception and comprehension of alternatives to natural language privacy policies; and (3) the development of a methodology for analyzing regulatory texts to extract stakeholder rights and obligations, which can be used to verify compliance with regulation and laws.

During our first study, we found that the privacy policies of healthcare organizations evolved significantly with the introduction of the HIPAA Privacy Rule. As a result, privacy policies seem to provide more insight into the actual privacy practices of the healthcare organizations. The policies are more descriptive, and contain much more privacy-related information than they did prior to the enactment of the HIPAA Privacy Rule. However, the total number of privacy related policies increased, as well as the average length of each individual policy. Moreover, the readability of each document decreased, placing undo burden on the consumer to parse these longer, more complex privacy policies [AEV06].

Based on the results of the first study, we acknowledge the need to provide consumers with more clear and concise privacy policies so that they can make informed decisions with regard to with whom they share their sensitive information. To address this need, we conducted an experiment to gauge user perception and comprehension of alternatives to natural language privacy policies. We expect that legislators will be able to utilize the results of our first study to substantiate the need for future legislation governing

the protection of privacy in domains other than healthcare. We found that users tend to perceive natural language privacy policies as being the more secure and protective of their personal information than the other alternatives with which they were presented, even when the natural language policies are the ones that are most likely to exploit or disclose their information. We also found that users do not comprehend natural language privacy policies as well as the other alternatives presented to them. Given these results, it is obvious that there are several alternatives to natural language privacy policies available to organizations that are interested in providing the consumer with the most comprehensive, readable policy. We also expect that HCI experts will be able to employ the results of our empirical survey as a basis for research into more effective alternatives to natural language privacy policies that could further increase the readability and user-friendliness of website privacy policies [EV06].

In the last study, we developed a methodology for analyzing regulatory texts to extract stakeholder rights and obligations, along with the constraints that must be satisfied before rights and obligations are designated to a stakeholder. Using this methodology, organizations would be able to perform a comparative analysis to determine whether or not they are compliant with regulations. External parties could also use this method to identify potential non-compliance issues [BVA06].

Significant research challenges remain to be addressed in this particular area. The comparative analysis process is sufficient for minor compliance checks, but we foresee the need to automate the processes of compliance checking and compliance monitoring when implementing systems based on such regulatory texts.

The results of these studies offer an insightful foundation for future research. Though many challenges still remain if we are to make further progress towards preserving privacy and ensuring legislative compliance in healthcare systems. The results of these studies can aid researchers, legislators, compliance officers, and organizations.

REFERENCES

[ACR99] Mark S. Ackerman, Lorrie Faith Cranor, Joseph Reagle. "Privacy in e-commerce: examining user scenarios and privacy preferences," *Proceedings of the 1st ACM conference on Electronic Commerce***,** pp. 1-8. November 1999.

[AE01] A.I. Antón and J.B. Earp. Strategies for Developing Policies and Requirements for Secure Electronic Commerce Systems. in *E-Commerce Security and Privacy*, ed. by A.K. Ghosh, Kluwer Academic Publishers, pp. 29-46, 2001.

[AEB04] A.I. Antón, J. B. Earp, D. Bolchini, Q. He, C. Jensen and W. Stufflebeam. "The Lack of Clarity in Financial Privacy Policies and the Need for Standardization," *IEEE Security & Privacy*, 2(2), pp. 36-45, March/April 2004.

[AER02] A.I. Antón, J.B. Earp and A. Reese, "Analyzing Web Site Privacy Requirements Using a Privacy Goal Taxonomy, *10th Anniversary IEEE Joint Requirements Engineering Conference (RE'02)*, Essen, Germany, pp. 605-612, 9-13 September 2002.

[AHB04] A.I. Antón, Q. He and D. Baumer. "The Complexity Underlying JetBlue's Privacy Policy Violations," *IEEE Security & Privacy*, 2(6), pp. 12-18, November/December 2004..

[AE04] A.I. Antón and J.B. Earp. "A Requirements Taxonomy to Reduce Web Site Privacy Vulnerabilities," *Requirements Engineering Journal*, Springer Verlag, 9(3), pp. 169-185, August 2004.

[AEL02] W.F. Adkinson, J.A. Eisenach and T.M. Lenard. "Privacy Online: A Report on the Information Practices and Policies of Commercial Web Sites," Progress and Freedom Foundation, Washington DC. March 2002

[AEL05] W.F. Adkinson, J.A. Eisenach and T.M. Lenard. *Privacy online: A Report on the Information Practices and Policies of Commercial Web Sites.* Washington, DC: Progress & Freedom Foundation, 2002. Downloaded January 19, 2005: http://www.pff.org/issues-pubs/books/020301privacyonlinereport.pdf

[AEV06] A.I. Antón, J. B. Earp, M. W. Vail, N. Jain, J. Frink and C. Gheen. "An Analysis of Web Site Privacy Policy Evolution in the Presence of HIPAA," To Appear: *IEEE Security & Privacy*, NCSU Technical Report #TR-2004-21, 24 July 2004.

[Ant96] A.I. Antón. Goal-Based Requirements Analysis, *2nd IEEE Int'l Conf. on Requirements Engineering (ICRE '96)*, Colorado, pp. 136-144, 15-18 April 1996.

[AP98] A.I. Antón and C. Potts. The Use of Goals to Surface Requirements for Evolving Systems, *IEEE International Conference on Software Engineering*, pp. 157-166, April 1998.

[BA05a] T.D. Breaux, A.I. Antón. Deriving Semantic Models from Privacy Policies. *IEEE 6th Workshop on Policies for Distributed Systems and Networks*, Stockholm, Sweden, pp. 67-76, 2005.

[BA05b] T.D. Breaux, A.I. Antón. Analyzing Goal Semantics for Rights, Permissions, and Obligations. *IEEE 13th Requirements Engineering Conference*, Paris, France, pp. 177-186, 2005.

[BA05c] T.D. Breaux, A.I. Antón. Mining Rule Semantics to Understand Legislative Compliance. *ACM Workshop on Privacy in Electronic Society*, Alexandria, Virginia, USA, pp. 51-54, 2005.

[BEP00] D. Baumer, J.B. Earp and F.C. Payton. Privacy of Medical Records: IT Implications of HIPAA. *ACM Computers and Society*, 30(4), pp.40-47, December 2000.

[BVA06] T.D. Breaux, M.W. Vail, A.I. Antón. Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. Technical Report TR-2006-6, North Carolina State University, Department of Computer Science, Raleigh, North Carolina, USA, February 2006.

[Citi02] *CitiGroup P3P Position Paper*, http://www.w3.org/2002/p3p-ws/pp/citigroup.html, September 2002.

[Clarke99] R. Clarke, "Internet privacy concerns confirm the case for intervention," *Communications of the ACM*, 42(2). pp. 60–67, 1999.

[CLM02] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle. "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification". W3C Recommendation. http://www.w3.org/TR/P3P/ 16 April 2002.

[CNN04] "Google adds 1 billion pages to search," 2004. http://www.cnn.com/2004/TECH/internet/02/18/google.expands.ap/

[CNN05] "Info on 3.9M Citigroup customers lost," 2005. http://money.cnn.com/2005/06/06/news/fortune500/security_citigroup/

[CP06] Federal Trade Commission: For the Consumer. ChoicePoint Settles Data Security Breach Charges; to Pay $10 Million in Civil Penalties, $5 Million for Consumer Redress. http://www.ftc.gov/opa/2006/01/choicepoint.htm January, 2006.

[CRA03] Computer Research Association (CRA) (2003) Conference on "Grand Research Challenges in Information Security and Assurance." http://www.cra.org/Activities/grand.challenges/security/. November 16-19, 2003.

[Drug03] "Drugstore.com Privacy Policy". http://www.drugstore.com. Downloaded on September 19, 2003.

[EASS05] J. B. Earp, A. I. Antón, L. A. Smith, and W. Stufflebeam. Examining Internet Privacy Policies Within the Context of User Privacy Values. *IEEE Transactions on Engineering Management.* 52(2), pp. 227 – 237, May 2005.

[EB03] J. B. Earp and D. Baumer, "Innovative web use to learn about user behavior and online privacy," *Communications of the ACM*, 46(4), pp. 81–83, April 2003.

[Elb05] Elbirt, A.J **"**Who are you? How to protect against identity theft," IEEE Technology and Society Magazine, 24(2), pp. 5–8, Summer 2005

[Entrust05] Entrust. Consumers and Regulators Call for Additional Security to Protect Identities and Reduce Fraud. http://www.entrust.com/news/2005/6126_6342.htm November, 2005.

[EPAL06] P. Ashley, S. Hada, G. Karjoth, C. Powers, M. Schunter. Enterprise Privavy Authorization Language (EPAL 1.2). http://www.w3.org/Submission/EPAL/ November, 2003.

[Epic00] Electronic Privacy Information Center (EPIC). Pretty Poor Privacy: An Assessment of P3P and Internet Privacy. http://www.epic.org/reports/prettypoorprivacy.html

[EStats05] United States Census Bureau. Measuring the Electronic Economy. http://www.census.gov/estats/ May, 2005.

[EV06] J. B. Earp, M. W. Vail. "Privacy Policy Representation in Web-based Healthcare," Submitted to: *12th Americas Conference on Information Systems*. March 2006.

[FA03] First Administrators, Inc. Penalties for Noncompliance. http://www.firstadministrators.com/hipaa/penalties.asp Accessed on February 25, 2006.

[Fle49] R. Flesch, *The Art of Readable Writing*, Macmillan Publishing, 1949.

[Fox03] S . Fox, "Internet Health Resources," Pew Internet and American Life Project, Washington D.C., July 2003.

[Fox05] S . Fox, "The Threat of Unwanted Software Programs is Changing the Way People Use the Internet," Pew Internet and American Life Project. Washington D.C. July 2005.

[FTC03] "Federal Trade Commision – Identity Theft Survey Report," 2003. http://www.consumer.gov/idtheft/pdf/synovate_report.pdf

[FTC98] *Privacy Online: A Report to Congress*, http://www.ftc.gov/ reports/privacy3/, Federal Trade Commission, June 1998.

[GHS00] J. Goldman, Z. Hudson and R.M. Smith. *Privacy Report on the Privacy Policies and Practices of Health Web Sites*, Sponsored by the California Health care Foundation, January 2000.

[Gno04] "Usability and Readability Considerations For Technical Documentation," http://developer.gnome.org/ documents/usability/usability-readability.html. Accessed June 2004.

[HC03] "HealthCentral.com Privacy Policy". http://www.healthcentral.com. Downloaded on September 19, 2003.

[HHS03] United States Department of Health and Human Services. http://www.hhs.gov/news/facts/privacy.html. Posted April 2003. Accessed May 2004.

[HIPAA] United States Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information. http://www.hhs.gov/ocr/hipaa/finalreg.html December, 2000.

[HIPPA05] "HIPAA Primer". http://www.hipaadvisory.com/regs/HIPAAprimer.htm.

[Jup02] Jupiter Research, "Security and Privacy Data." FTC Security Workshop, May 20, 2002.

[JP04] C. Jensen and C. Potts. Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. *Proceedings of ACM Conference on Human Factors in Computing Systems: CHI 2004*. pp. 471-478, April 2004.

[LTLW04] Lau, G. T., Kerrigan, S., Law, K. H., and Wiederhold, G. An e-government information architecture for regulation analysis and compliance assistance. In *Proceedings of the 6th international Conference on Electronic Commerce.* pg. 461-470, October 2004.

[Mason86] R. Mason, "Four ethical issues of the information age," *MIS Quarterly*, vol. 10, pp. 4–12, 1986.

[MH01] M. Hochhauser. Lost in the Fine Print: Readability of Financial Privacy Notices. *Privacy Rights Clearinghouse*. July 2001. http://www.privacyrights.org/ar/GLB-Reading.htm

[NTI02] National Telecommunications and Information Administration. *A Nation Online: How Americans Are Expanding Their Use of the Internet*, http://www.ntia.doc.gov/ntiahome/dn/ Washington, D.C. February 2002.

[RR04] V. Romney and G. Romney. Neglect of Information Privacy Instruction – A Case of Educational Malpractice? *Proceedings of the 5th conference on Information technology education*. Pages: 79 – 82, October 28-30, 2004.

[PM04] E. Perkins and M. Markel. Multinational Data-Privacy Laws: An Introduction for IT Managers. *IEEE Transactions on Professional Communication*, 47(2), pp. 85-94, June 2004.

[SDP05] R. Senanayake, G. Denker, and J. Pearce. Towards Integrated Specification and Analysis ofMachine-Readable Policies Using Maude. *4th International Semantic Web Conference*. November, 2005.

[TFM05] J. Turows, L. Feldman and K. Meltzer. *Open to exploitation: American shoppers online and offline*. Technical Report., Annenberg Public Policy Center, University of Pennsylvania, June 2005. http://www.annenbergpublicpolicycenter.org/04_info_society/Turow_APPC_Report_WEB_FINAL.pdf.

[Westin] Louis Harris & Associates and Dr. Alan Westin, "E-Commerce & Privacy: What Net Users Want," Technical Report. 1998.

[WHIR05] "Yahoo! to Digitize Library Contents," 2005**.** http://www.thewhir.com/find/articlecentral/story.asp?recordid=1460&page=1

[WW05] Princeton Survey Research Associates International. *Leap of Faith: Using the Internet Despite the Dangers*. http://www.consumerwebwatch.org/pdfs/princeton.pdf Yonkers, New York. October, 2005.

APPENDIX

APPENDIX A

User Perception Questions:

Why didn't you read the entire privacy policies of the website?

- The policies were too hard to understand
- The policies were too long
- The policies were not organized well
- I have no interest in the privacy policies of institutions I share my personal information with
- I read the entire set of privacy policies of the website

I feel secure sharing my personal information with BrandX after viewing their privacy practices.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

I believe BrandX will protect my personal information more than most other companies.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

I believe BrandX will protect my personal information less than most other companies.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

I feel that BrandX's privacy practices are explained thoroughly in the policy I read.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

I feel confident in my understanding of what I read of BrandX's privacy policies.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

I generally read the privacy policies of websites that I share my personal information with.

- Strongly Disagree
- Disagree
- Unsure
- Agree
- Strongly Agree

Demographic Questions:

How old are you?

- less than 18
- 18-21
- 22-28
- 29-35
- 36-42
- 43-49
- 50-57
- 57+
- Rather Not Say

What is your gender?

- Female
- Male

- Rather Not Say

What is your level of education?

- Some High School
- High School Graduate
- Some College
- College Graduate
- Some Graduate School
- Master's Degree (or equivalent)
- Doctorate Professional Degree (MD/JD)
- Other
- Rather Not Say

What is your occupation?

- Accounting
- Consulting
- Education
- Finance
- Government
- Health
- Information Technology
- Manufacturing
- Marketing
- Nonprofit
- Research and Development
- Retail
- Student
- Other
- Rather Not Say

What is your ethnic background?

- African
- African American
- Asian or Pacific
- Hispanic
- Native American
- White/Caucasian
- Other
- Rather Not Say

What is your nationality?

- United States
- African
- Asian
- Canadian
- European
- Mexican
- Other
- Rather Not Say

What is your web usage per week on average?

- 0-4 hours
- 5-9 hours
- 10-19 hours
- 20-29 hours
- 30+ hours
- Rather Not Say

How frequently do you make purchases online?

- Less than once each month
- About once each month
- Several times each month
- About once each week
- More than once each week
- At least once each day
- Never
- Rather Not Say

When was the last time you made a purchase online?

- Within the past 30 days
- 2-3 months ago
- 3-6 months ago
- 6 months to a year ago
- Longer than a year ago
- Not Applicable
- Rather Not Say

How often do you purchase healthcare related products online?

- Less than once each month
- About once each month
- Several times each month
- About once each week
- More than once each week
- At least once each day
- Never
- Rather Not Say

How often do you research health related topics online?

- Less than once each month
- About once each month
- Several times each month
- About once each week
- More than once each week
- At least once each day
- Never
- Rather Not Say

Comprehension Questions:

DRUGSTORE.COM

Which statement is TRUE regarding BrandX allowing or restricting the ability to access/modify Personally Identifiable Information (PII)?

- Users are allowed to access their PII, but not modify it
- Users must submit a request in writing to access their PII
- Users cannot access or modify their PII
- Users are allowed to access and modify their PII
- The policy does not address access and participation

Which statement is TRUE regarding BrandX ensuring that users are given the option to decide what Personally Identifiable Information is collected about them and how it can be used?

- BrandX does not give users this option
- BrandX allows consumers to opt-out of sharing information with 3rd parties
- BrandX does not share any information except at the consumer's request
- BrandX allows only one change to a user's PII usage preferences per year
- The policy does not address choices and consent

Which statement is TRUE about the mechanisms BrandX has in place to enforce privacy?

- BrandX limits employee access to your PII
- BrandX obligates subsidiaries to comply with their institution's privacy policy
- BrandX maintains secure financial services
- BrandX obligates 3rd parties to agree not to connect aggregate info with PII to identify customers
- The policy does not address enforcement mechanisms

Which statement is TRUE regarding BrandX's information aggregation practices?

- BrandX aggregates information by zip code
- BrandX aggregates information about average visit length
- BrandX aggregates information about user browsing patterns
- BrandX aggregates information about advertising effectiveness on their websites
- The policy does not address aggregation practices

Which statement is TRUE regarding BrandX's information collection practices?

- BrandX collects Personal Health Information (PHI) in order to fill a prescription
- BrandX collects information about your browser type
- BrandX collects information from customer emails sent to their institution
- BrandX collects information from forms filled out on their site
- The policy does not address how information is collected

Which statement is TRUE regarding BrandX's information monitoring practices?

- BrandX monitors statistics about visits to their website
- BrandX uses session cookies to help users navigate their website efficiently
- BrandX allows 3rd parties to use Web bugs
- BrandX monitors customer activities relating to their online promotions
- The policy does not address how information is monitored

Which statement is TRUE regarding BrandX using information collected to customize users' experiences?

- The policy does not address how information is used to customize my experience
- BrandX uses Personally Identifiable Information (PII) to improve their site
- BrandX uses information collected from aggregation services to provide targeted advice to help you achieve your health goals
- BrandX uses session cookies to provide personalized service
- BrandX uses cookies to store customer preferences for marketing offers

Which statement is TRUE regarding BrandX storing information?

- BrandX stores all information collected from international users according to US law
- BrandX stores credit card information securely in a separate, encrypted database
- BrandX stores customer email responses to meet legal requirements
- BrandX stores IP addresses to track a connection to point of origin for security reasons
- The policy does not address how or what information will be stored

Which statement is TRUE regarding BrandX transferring information?

- BrandX only shares information internally
- BrandX shares information with 3rd parties
- BrandX has a strict policy of not sharing or transferring any information
- BrandX will only share information after receiving consent from the consumer
- The policy does not address how or what information will be transferred/shared

Which statement is TRUE regarding BrandX's efforts to ensure the integrity and/or security of information?

- BrandX prevents anyone from accessing customer access codes
- BrandX avoids selling customer lists
- BrandX prevents unauthorized access by using a firewall
- BrandX protects computer systems by detecting computer viruses
- The policy does not address any efforts to ensure integrity/security

Which statement is TRUE regarding BrandX notifying consumers and making them aware of how and what information is collected and/or used?

- BrandX helps customers understand how they protect Customer Information
- BrandX informs customers how to limit the amount of marketing the customer receives
- BrandX will notify customers of changes to our Privacy Policy
- BrandX informs customers of contact info for their privacy office
- The policy does not address any efforts towards notification/awareness

NOVARTIS.COM

Which statement is TRUE regarding BrandX allowing or restricting the ability to access/modify Personally Identifiable Information (PII)?

- Users are allowed to access their PII, but not modify it
- Users must submit a request in writing to access their PII
- Users cannot access or modify their PII
- Users are allowed to access and modify their PII
- The policy does not address access and participation

Which statement is TRUE regarding BrandX ensuring that users are given the option to decide what Personally Identifiable Information is collected about them and how it can be used?

- BrandX does not give users this option
- BrandX allows parents opt out of the collection of their child's PII
- BrandX does not share any information except at the consumer's request
- BrandX allows only one change to a user's PII usage preferences per year
- The policy does not address choices and consent

Which statement is TRUE regarding BrandX using Personally Identifiable Information to contact you?

- BrandX contacts customers via email
- BrandX uses your PII for marketing & promotions
- BrandX contacts customers only to resolve problems
- BrandX uses regular mailings to update customers
- BrandX will contact customers when cards have been lost or stolen

Which statement is TRUE about the mechanisms BrandX has in place to enforce privacy?

- The BrandX obligates subsidiaries to comply with their institution's privacy policy
- BrandX limits employee access to your PII
- BrandX maintains secure financial services
- BrandX obligates 3rd parties to agree not to connect aggregate info with PII to identify customers
- The policy does not address enforcement mechanisms

Which statement is TRUE regarding BrandX's information collection practices?

- BrandX uses cookies to collect information about site usage
- BrandX collects information about your browser type
- BrandX collects information from customer emails sent to their institution

- BrandX collects information from forms filled out on their site
- The policy does not address how information is collected

Which statement is TRUE regarding BrandX storing information?

- BrandX stores all information collected from international users according to US law
- BrandX stores PII long enough to meet legal requirements
- BrandX stores customer email responses
- BrandX stores IP addresses to track a connection to point of origin for security reasons
- The policy does not address how or what information will be stored

Which statement is TRUE regarding BrandX transferring information?

- BrandX only shares information internally
- BrandX shares information with 3rd parties to perform services on their behalf
- BrandX has a strict policy of not sharing or transferring any information
- BrandX will only share information after receiving consent from the consumer
- The policy does not address how or what information will be transferred/shared

Which statement is TRUE regarding BrandX notifying consumers and making them aware of how and what information is collected and/or used?

- BrandX helps customers understand how they protect Customer Information
- BrandX informs customers how to limit the amount of marketing the customer receives
- BrandX will notify parents of information collected from a child
- BrandX informs customers of contact info for their privacy office
- The policy does not address any efforts towards notification/awareness

HEALTHCENTRAL.COM

Which statement is TRUE regarding BrandX allowing or restricting the ability to access/modify Personally Identifiable Information (PII)?

- Users are allowed to access their PII, but not modify it
- Users must submit a request in writing to access their PII
- Users cannot access or modify their PII
- BrandX limits licensing of content to 3rd parties
- The policy does not address access and participation

Which statement is TRUE regarding BrandX ensuring that users are given the option to decide what Personally Identifiable Information is collected about them and how it can be used?

- BrandX does not give users this option
- BrandX avoids collecting PII without the customer's consent
- BrandX does not share any information except at the consumer's request
- BrandX allows only one change to a user's PII usage preferences per year
- The policy does not address choices and consent

Which statement is TRUE regarding BrandX's information collection practices?

- BrandX collects responses to promotional emails
- BrandX collects information about your browser type
- BrandX collects information about children that visit the site
- BrandX collects information from forms filled out on their site
- The policy does not address how information is collected

Which statement is TRUE regarding BrandX's information monitoring practices?

- BrandX monitors statistics about visits to their website
- BrandX uses session cookies to help users navigate their website efficiently
- BrandX collects PII to provide customers with a personalized/customized experience
- BrandX monitors customer activities relating to their online promotions
- The policy does not address how information is monitored

Which statement is TRUE regarding BrandX using information collected to customize users' experiences?

- The policy does not address how information is used to customize my experience
- BrandX uses session cookies to provide personalized service
- BrandX uses information collected from aggregation services to provide targeted advice to help you achieve your health goals
- BrandX uses Personally Identifiable Information (PII) collected from customers deliver customized services
- BrandX uses cookies to store customer preferences for marketing offers

Which statement is TRUE regarding BrandX transferring information?

- BrandX only shares information internally
- BrandX shares information with 3rd parties to perform marketing services on our behalf
- BrandX has a strict policy of not sharing or transferring any information
- BrandX will only share information after receiving consent from the consumer

- The policy does not address how or what information will be transferred/shared

Which statement is TRUE regarding BrandX's efforts to ensure the integrity and/or security of information?

- BrandX prevents anyone from accessing customer access codes
- BrandX prevents unauthorized 3rd party access to customer PII
- BrandX protects PII by encrypting it
- BrandX protects computer systems by detecting computer viruses
- The policy does not address any efforts to ensure integrity/security

Which statement is TRUE regarding BrandX notifying consumers and making them aware of how and what information is collected and/or used?

- BrandX helps customers understand how they protect Customer Information
- BrandX informs customers how to limit the amount of marketing the customer receives
- BrandX informs customers they cannot protect PII posted in chat rooms
- BrandX informs customers of contact info for their privacy office
- The policy does not address any efforts towards notification/awareness

APPENDIX B

Demographic Data

Age	**Percentage**
Less Than 18	0.92
18-21	12.41
22-28	21.13
29-35	16.31
36-42	14.15
43-49	11.59
50-57	11.79
57+	9.85
Rather Not Say	1.44
No Response	0.41
Gender	**Percentage**
Female	29.74
Male	66.87
Rather Not Say	2.67
No Response	0.72
Nationality	**Percentage**
United States	72.72
African	0.31
Asian	4.41
Canadian	4.10
European	4.31
Mexican	0.41
Other	11.18
Rather Not Say	1.95
No Response	0.62
Ethnicity	**Percentage**
African	0.10
African American	2.77
Asian or Pacific	7.59
Hispanic	4.82
Native American	1.13
White/Caucasian	71.28
Other	3.28
Rather Not Say	8.10
No Response	0.92

Occupation	Percentage
Accounting	0.92
Consulting	5.44
Education	13.95
Finance	1.03
Government	3.90
Health	2.26
Information Technology	24.00
Manufacturing	0.92
Marketing	1.44
Nonprofit	1.44
Research and Development	7.90
Retail	0.72
Student	24.82
Other	8.10
Rather Not Say	2.77
No Response	0.41

Education	Percentage
Some High School	1.64
High School Graduate	1.95
Some College	19.69
College Graduate	20.82
Some Graduate School	10.87
Master's Degree (or equivlaent)	24.62
Doctorate	14.67
Professional Degree (MD/JD)	3.79
Other	0.62
Rather Not Say	0.92
No Response	0.62

Web Usage (per week)	Percentage
0-4 hours	5.03
5-9 hours	15.49
10-19 hours	25.95
20-29 hours	19.18
30+ hours	32.51
Rather Not Say	1.44
No Response	0.41

Online Purchases	**Percentage**
Less than one each month	27.79
About one each month	30.05
Several times each month	25.54
About once each week	6.36
More than once each week	2.87
At least once each day	0.21
Never	4.82
Rather Not Say	1.64
No Response	0.72
Last Online Purchase	**Percentage**
Within the past 30 days	69.44
2-3 months ago	14.97
3-6 months ago	4.51
6 months to a year ago	3.38
Longer than a year ago	1.74
Not Applicable	3.90
Rather Not Say	1.54
No Response	0.51
Healthcare Product Purchases	**Percentage**
Less than one each month	27.08
About one each month	3.59
Several times each month	0.62
About once each week	0.10
More than once each week	0.10
At least once each day	0.21
Never	66.77
Rather Not Say	1.13
No Response	0.41
Healthcare Research	**Percentage**
Less than one each month	45.33
About one each month	22.77
Several times each month	10.46
About once each week	3.59
More than once each week	3.38
At least once each day	1.13
Never	11.59
Rather Not Say	1.33
No Response	0.41