# The Role of Legal Expertise in Interpretation of Legal Requirements and Definitions

David G. Gordon
Engineering and Public Policy
Carnegie Mellon University
Pittsburgh, USA
dggordon@cmu.edu

Travis D. Breaux
Institute for Software Engineering
Carnegie Mellon University
Pittsburgh, USA
breaux@cs.cmu.edu

*Abstract*—Government laws and regulations increasingly place requirements on software systems. Ideally, experts trained in law will analyze and interpret legal texts to inform the software requirements process. However, in small companies and development teams with short launch cycles, individuals with little or no legal training will be responsible for compliance. Two specific challenges commonly faced by non-experts are deciding if their system is covered by a law, and then deciding whether two legal requirements are similar or different. In this study, we assess the ability of laypersons, technical professionals, and legal experts to judge the similarity between legal coverage conditions and requirements. In so doing, we discovered that legal experts achieved higher rates of consensus more frequently than technical professionals or laypersons and that all groups had slightly greater agreement when judging coverage conditions than requirements, measured by Fleiss' K. When comparing judgments between groups using a consensus-based Cohen's Kappa, we found that technical professionals and legal experts exhibited consistently greater agreement than that found between laypersons and legal experts, and that each group tended towards different justifications, such as laypersons and technical professionals tendency towards categorizing different coverage conditions or requirements as equivalent if they believed them to possess the same underlying intent.

*Index Terms*—legal requirements, compliance, statutory interpretation, legal coverage

## I. Introduction

National and provincial laws are enacted in an attempt to address growing privacy and security concerns by restricting how and when personal data may be obtained, used, and protected. This may include obtaining consent from an individual prior to collecting data or after organizational policies have changed, encrypting or otherwise securing data in storage or in transit, and restricting the roles of parties with whom that information can be shared. Laws, including statutes, directives, and regulations, issued in the past five years include numerous U.S. data breach notification laws, updates to the Health Insurance Portability and Accountability Act (HIPAA), and revisions to the European Data Protection Directive, among others [2].

The cost and complexity of compliance with these laws – particularly in multi-jurisdictional contexts – is nontrivial [8, 31] and can have profound effects on system design and product strategy, such as influencing the attractiveness of a new market or the viability of collecting, storing, and securing data from certain jurisdictions or with certain resources, e.g. an overseas cloud storage service [19]. When addressing the issue of compliance with a new or updated law, an organization must interpret the text and determine whether or not it is covered by the law (i.e. does the law apply to the organization) and consequently what actions are permitted, required, or prohibited to achieve compliance, which we define as alignment of the organization's goals with those prescribed through legislation.

In large companies, these determinations are relegated to a legal department with attorneys and paralegals whose additional duties include discovering recently enacted laws, assessing the risk of noncompliance, and offering consultative advice to IT projects as needed. However, legal resources – be they in the form of an internal legal department, or external firm – may not be utilized in all applicable cases. For example, a 2012 survey of 352 mobile application developers found that 75% of mobile application companies have five or fewer employees, with 40% of those companies consisting of a single employee: the developer of the mobile application [10]. In this case, the mere presence of legal expertise is doubtful, and the organization may be unwilling, or financially unable, to retain legal advice. Even if the firm can afford legal advice, it would be desirable to guide the legal expert to assess specific technical problems relevant to their mobile market segment. In these scenarios, interpretation of the text may be consigned to individuals with little to no legal expertise.

In this paper, we investigate the within-group and between-group similarity with which individuals with no legal background, hereafter referred to as laypersons, technical professionals, and legal experts interpret legal texts. To answer this question, we conducted an empirical survey to evaluate how these groups (laypersons, technical professionals, and legal experts) make judgments based on their interpretation of legal texts. The survey consists of a legal knowledge test and questions regarding the relationship between similar legal definitions and requirements, and was administered to individuals with varying degrees of legal and IT training and

experience. The results show that legal experts achieve stronger majorities (e.g. 80% agreement vs. 60% agreement on a single interpretation) slightly more often than laypersons and technical professionals, and exhibit only slightly greater within-group agreement overall. Further, we discovered that technical professionals and legal experts achieve consistently higher rates of between-group agreement than laypersons and legal experts.

The remainder of this paper is organized as follows: in Section II, we review related work; in Section III, we present our model of legal decision making; in Section IV we introduce our statistics for measuring consensus; in Section V, we describe our survey design, including question selection; in Section VI we describe our summative results; in Section VII we discuss our findings, including differences in justifications provided by members of each group; in Section VIII we address threats to validity and how they were mitigated throughout the study; and in Section IX we conclude and discuss future work.

## II. Related Work

The significance of regulatory compliance in requirements engineering and system development has attracted increasing attention in academic research and industry [30, 12]. In this section, we review techniques and methods in requirements engineering to analyze and understand legal texts, including those that inspired this work; the role of statutory interpretation in determining legal coverage and standards; and studies regarding legal expertise.

Law has become an increasingly important consideration in the requirements engineering community. Research in this area has addressed a variety of issues, including extraction of legal requirements [4, 5], ambiguity detection and resolution [3, 33, 16], and compliance determination [23, 37]. However, these texts do not focus on the interpretive differences made by the analysts due to differences in background. Our approach in this work is driven by earlier studies pursued by the authors in requirements water marking and requirements coverage modeling, which involve determination of high legal standards in the presence of laws from multiple jurisdictions and differences in legal coverage in the presence of regulatory and environmental change, respectively [18, 19]. In the case of requirements water marking, differences in interpretation could affect what requirements and coverage conditions contribute to the high standard of care, potentially resulting in a different, though not necessarily incorrect, notions as to what as to what the high standard is. With respect to coverage modeling, differences in interpretation could affect how and when an organization responds to changes in its circumstances.

Legal experts consider a number of factors in addition to the meaning of the legal text when they advise a client about the legal standards they must adhere to. These factors generally relate to the text itself and the current legal landscape, including the novelty and public visibility of relevant laws, the strictness of their enforcement, and the severity of penalties and sanctions imposed. With regards to the text, the legal expert may take into account its legislative history, the intent surrounding its creation, and "lessons of common sense and policy" [13]. The expert will further develop her understanding of the text by taking into account interpretations made by governing bodies or legal authorities and how other members of the client's industry have chosen to interpret and respond to the text, among other factors. However, while many factors may be considered, the history, legislative intent, and landscape surrounding the text have little weight without first imparting some meaning to the text itself. Some legal experts, such as current Supreme Court Justice Antony Scalia, argue in favor of textualism: that statutory interpretation should begin and end with the apparent meaning of the statutory language [34]. The results of this study should be interpreted within this scope, elegantly encapsulated by Oliver Wendell Holmes' statement, "we do not inquire what the legislature meant; we ask only what the statute means." [22].

Individuals with little or no legal training frequently encounter legal texts, particularly contracts [32]. While it is commonly known that legal documents are difficult for laypersons to understand [21], to the best of our knowledge little research has been conducted investigating what laypersons believe the texts to mean with regards to legal coverage and compliance, much less comparing interpretations made by laypersons to those made by legal experts. Related is a study conducted by Christensen [7], who compared how legal experts and novices read judicial opinions. However, Christensen's focus was not on interpretation, but on reading processes (e.g. was the text read linearly or non-linearly), and she did not present her criteria for determining legal expertise. Of course, expertise can be difficult to define [11], despite being a topic of interest for over a century [35, 36]. Common measurements used to approximate expertise include the presence of experience, certifications, or social acclimation; the ability to make finely discriminable and consistent judgments, and performance on knowledge tests [36]. For example, legal expertise may be informed by possession of Juris Doctor or other certification (e.g. professional paralegal, certified legal assistant, etc.); bar certification in one or more jurisdictions; graded performance on commonly administered authoritative legal examinations, such as the Multistate Bar Examination [29]; years spent practicing law; or through self-attestation. We measured many of these variables in our study, which are detailed in Section V.

Thus, while several requirements engineering researchers aim to equip analysts with tools to demonstrate compliance, little is still known about the differences between experts and novices in how they interpret the law. In the next section, we describe our assumptions that underpin our study.

## III. Modeling Legal Coverage and Requirement Relationships

We inform our survey design by adopting a simple model for comparing legal coverage and requirements, which is used to evaluate the judgments obtained through our survey. Consider a small, specialized healthcare clinic based in New York that uses a custom mobile application for employees to record and view vitals (pulse, blood pressure, etc.) with their

own mobile devices. They consider opening another branch in California and need to know if California's laws are similar to New York's laws with regards to the types of entities, data, and circumstances covered. Second, they need to know if any new requirements imposed by California law exceed or fall short of those imposed by New York law. After obtaining relevant laws, they discover that they appear to apply to different types of data, as shown by the requirements in Figure 1.

| NEW YORK | CALIFORNIA |
|---|---|
| [the organization] must maintain a medical record for every person evaluated or treated as an inpatient, ambulatory patient, emergency patient or outpatient | [the organization] must maintain a medical record on all patients admitted or accepted for treatment |

Fig. 1. Medical record requirements from N.Y. Comp. Codes R. & Regs. tit. 10, § 405.10 and California Code Regs. tit. 22 § 70751

Believing the two requirements have some degree of similarity (i.e. they are not disjoint), with regards to data types alone, they consider the following possible coverage relationships, conceived as Venn diagrams:

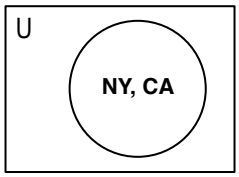
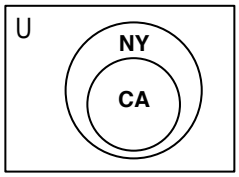
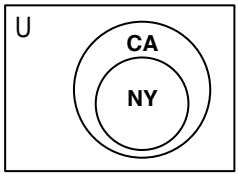
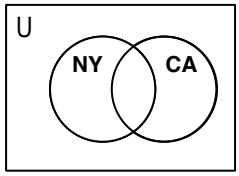
| COVERAGE RELATIONSHIP | VENN DIAGRAM *(U: DATA TYPES)* |
|---|---|
| **Equivalent**: the NY laws apply to the same data types as CA laws |   |
| **Subsumption**: the NY laws apply to all data types covered by the CA laws, as well as others (the NY laws *subsume* CA laws) |   |
| **Inclusion**: the CA laws apply to all data types covered by the NY laws, as well as others (the NY laws are *included* in CA laws) |   |
| **Partial**: the NY and CA laws apply to some but not all of the same data types |   |

Fig. 2. Possible Coverage Relationships between Data Types Described by NY and CA Law

Ideally, the coverage relationship determined by the clinic would be in agreement with that made by a legal expert. Unlike the clinic, the legal expert's decision would be informed by training and experience enabling her to provide more informed and consistent analyses and decisions than those without [36]. Disagreements about coverage could lead to needless system development constraints and resource expenditures on unnecessary compliance measures if the clinic believed itself to be covered when the expert believes it is not. Similarly, the organization could unknowingly neglect its obligations and partake in prohibited actions if the clinic believed itself not to be covered when the expert believes it is. In technical terms, such disagreements could include applying excessive data security protections, such as two-factor authentication, or sharing potentially sensitive information with unauthorized 3rd parties. Requirements can also be modeled in this fashion, with the standard of care imposed by one requirement being equal to that imposed by another, subsuming another, etc. Similarly detrimental outcomes could also occur if the clinic made incorrect assessments with regards to the relationship between the requirements as well: for example, if the clinic believed that a New York requirement matched or exceeded the standard of a similar California requirement when the expert believes the opposite, the clinic would fail to take the additional measures necessary to achieve compliance with the California standard.

## IV. MEASURING BETWEEN-GROUP AGREEMENT

The extent to which individuals of a particular group agree to item categorization can be measured using an inter-rater reliability statistic, which is frequently used in medicine and psychology. Today, preference is given to statistics that account for the proportion of agreement corrected for chance agreement. Two of the most widely used statistics include Cohen's κ (Kappa) [9] and Fleiss' intra-class coefficients [15]: the former is use to calculate agreement between two raters for a fixed number of items across multiple categories; the latter is for more than two raters. In our survey, we are interested in comparing agreement within and between groups of raters. While methods exist to measure between-group agreement across ordinal data [27, 14], methods are only recently emerging to measure between-group agreement on categorical data [38]. In this study, we determine between-group agreement using a consensus-based Cohen's κ and Van Belle's Nominal Agreement Index (NAI).

The commonly used [38] consensus method involves determining the group consensus (i.e. the mode, or most common categorization) for each item, treating the group's consensus as an individual rater's categorization, and calculating agreement using a common agreement statistic – in this case, Cohen's κ. Though intuitive, by reducing each group to its consensus categorizations, the statistic is imperfect in that it treats the consensus categorization as a binary variable (agree or disagree) rather than a proportion, and does not reflect the *degree* of consensus each group achieves.

Recently developed in 2009, Van Belle's Nominal Agreement Index (NAI) is intended to be a "natural extension" of Cohen's κ for two groups of raters [38] and in the case that each group contains a single rater, reduces to Cohen's κ. The NAI is similar to Cohen's κ by taking into account the relative observed agreement between raters as well as the probability of

chance agreement, but with the following two modifications for groups: (a) unlike Cohen's κ, wherein agreement is binary – the two raters agree or do not agree – the NAI's agreement between groups is based on the *proportion* of each group that assigned a specific category to a specific item, which also allows comparisons between groups of different sizes; and (b) unlike Cohen's κ, wherein the best possible agreement for two raters is 1, the NAI defines the best possible agreement for each item as the category with the highest proportion of raters in agreement. For example, consider Table I that presents two five-member groups 1 and 2 who categorized 10 items into three categories A, B, or C. Each rater's categorizations and group consensus for each item are shown in Table I. As can be seen in the bottom two rows, the group's consensuses are never in agreement, resulting in a simple percent agreement of 0% and Cohen's κ of -.23, indicating the groups agree less than can be expected by chance alone. While the groups' consensus for each item is never the same, their responses are fairly similar, differing by a single rater in all cases. This similarity is accounted for by Van Belle's NAI, which is .78 for the data.

TABLE I. Artificial data illustrating differences in Consensus Cohen's Kappa and Van Belle's NAI

| | Rater | Item | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Group 1 | 1 | B | C | A | B | A | C | B | C | B | C |
| | 2 | B | C | A | B | A | C | B | C | B | C |
| | 3 | A | B | A | C | C | C | A | C | C | C |
| | 4 | A | B | B | C | C | A | A | B | C | B |
| | 5 | A | B | B | C | C | A | A | B | C | B |
| Group 2 | 6 | A | B | B | C | C | A | B | B | C | B |
| | 7 | A | B | B | C | C | A | B | B | C | B |
| | 8 | B | C | B | B | A | A | B | B | B | B |
| | 9 | B | C | A | B | A | C | A | C | B | C |
| | 10 | B | C | A | B | A | C | A | C | B | C |
| Group 1 Consensus | | A | B | A | C | C | C | A | C | C | C |
| Group 2 Consensus | | B | C | B | B | A | A | B | B | B | B |

Although this statistic has received some expert review, it remains to be empirically validated. Thus, we present our results with both the consensus-based Cohen's κ and the NAI.

## V. Research Method and Survey Design

We now discuss our study design, including research questions, survey construction, dataset selection process, participant recruitment methods, units of analysis, and analysis procedure. To guide our research, we proposed the following research questions:

**$RQ_1$**: Do laypersons, technical, and legal experts make internally consistent judgments regarding relationships about legal coverage and requirements?

**$RQ_2$**: Do laypersons, technical, and legal experts make similar judgments regarding relationships about legal coverage and requirements?

**$RQ_3$**: How do laypersons and technical experts justify their interpretations of legal coverage and requirement texts as compared to legal experts?

To address our research questions, we developed and administered an online survey to participants of varying legal and technical backgrounds.

The authors considered numerous research instruments prior to selection of the online survey. Although online surveys may restrict the freedom of participants' qualitative responses as well as the researcher's ability to explore these responses further, unlike interviews, we believed these consequences to be outweighed by the advantages of greater participant convenience and affordability. Legal experts' time is both highly limited and expensive, with even junior associates commanding rates upwards of $400/hour [20]. Unlike interviews, even if conducted remotely, the online survey gave legal experts the ability to participate in the study when and where they chose without any additional intervention on the part of the researchers. Even given these circumstances, the lack of legal expert availability we encountered motivated multiple means of recruitment as described in subsection C.

### A. Survey Process

The survey was administered online in two stages to avoid mental fatigue on the part of the participant and to reduce confounds due to priming effects from the knowledge test: in stage one, participants provide basic demographic information (age, gender), information regarding their technical and legal backgrounds (discussed in Section V) and then answer a number of multiple choice questions obtained from publicly available, multi-state bar examinations to provide basic validation regarding their professed legal background. At the end of stage one, participants are given the opportunity to participate in the second stage, consisting of twenty questions pertaining to requirements comparison and definition-based legal coverage, both framed using the model presented in Section III, and preceded by examples of both question types. Each stage was designed to take approximately 30-45 minutes, and all participants were reimbursed $10 per stage for a total of $20.

### B. Knowledge and Skills Question Selection

The knowledge questions in stage one were obtained from three publicly available, multi-state bar examination tests in the United States. The multi-state bar examination (MBE) is updated each year by the National Conference of Bar Examiners and is required for admission to the bars of all but two United States jurisdictions [29]. Topics covered include Constitutional law, contracts, criminal law and procedure, evidence, real property, and torts. Twenty questions were down-selected from the 600 available using the following process: (1) after consultation with a legal expert, real property and evidence were removed as categories due to domain specificity; and (2) only questions with one acceptable answer were retained, excluding questions with multiple valid answers. These two criteria yielded a total of 407 remaining questions, from which five were randomly selected from each remaining category (constitutional law, contracts, criminal law, torts) for a

total of 20 questions. Order of answer responses was randomized for each participant.

The questions for stage two regarding requirements and legal coverage comparison were constructed by the investigators using requirements and definitions (hereafter referred to collectively as terms) selected from laws analyzed in prior work [17, 18, 19] and chosen because they cover multiple jurisdictions, address the increasing modernization of the healthcare industry, and reflect the most recent developments in data protection, as presented in Table II.

TABLE II. Recent Developments in Data Protection Law

| NAME | YEAR(S) | JURISDICTION |
|---|---|---|
| Data breach notification laws | 2003 – 2011 | US; state-level |
| Data destruction laws | 2003 – 2011 | US; state-level |
| Medical record retention laws | 1971 – 2011 | US; state-level |
| Health Insurance Portability and Accountability Act *(HIPAA)* | 1996 | US; national |
| Health Information Technology for Economic and Clinical Health Act *(HITECH)* | 2009 | US; national |
| Information Technology Rules | 2011 | India; national |

The terms from these prior studies have undergone extensive analysis, including demarcation, modal classification (e.g., as obligations, permissions, or prohibitions), categorical annotation (e.g. data breach criteria, medical record storage, medical record transfer), identification of internal and external cross-references, and alignment of fully and partially equivalent pairs of requirements or definitions [17]. Each requirement and definition pair was assigned a unique identifier and then randomly sampled for use in the survey. After sampling was completed, terms were reviewed to ensure they were understandable when taken out of context; terms requiring additional context were discarded and replaced.

Importantly, this study is not intended to address an individual's ability to decompose and comprehend elaborately structured definitions (e.g. HIPAA's definition for "covered entity") or ornate requirements with overlapping, clause-heavy exceptions: it is intended to address differences in individual interpretations for short terms or phrases. Even taken in isolation, the requirements and definitions used in this study are verbose, dense, and filled with legalese and jargon. To address this issue, the investigators reviewed the selected terms and conservatively omitted or generalized extraneous text in order to focus the participant's attention on the interpretation of specific words or phrases for each comparison. Details and validity concerns regarding this process are addressed in Section VIII. While some might assume there is a "correct" relationship for each pair of terms that can be treated as a golden standard, we did not design our survey with this assumption.

Finally, terms were framed as comparison questions (example in Figure 3, below) using the relationship types described in Section III, with full, plain-language explanations of each relationship type offered at the top and bottom of each page.

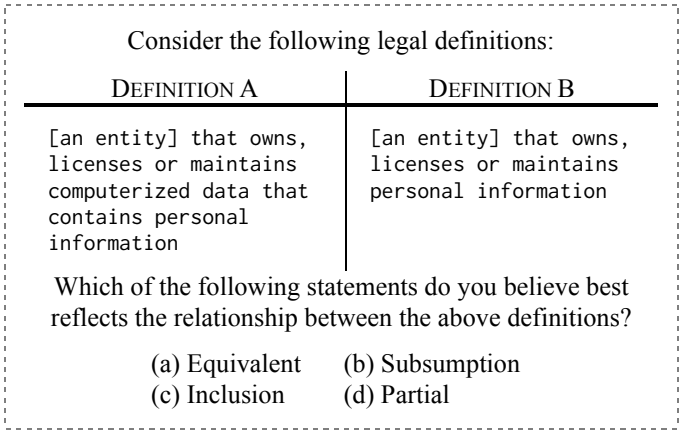



Fig 3. Example Question Comparing Two Definitions

### C. Participant Recruitment

Due to the difficulty of recruiting legal experts, participants were recruited through a variety of media, including human intelligence tasks (HITs) on Amazon Mechanical Turk, e-mail solicitations of graduate technical departments of major academic institutions, posts to online message boards (e.g. craigslist), and distribution of fliers at the 2014 IAPP Privacy Summit in Washington, D.C.

## VI. SUMMATIVE FINDINGS

Responses were collected over an eight-week period beginning in January 2014. Table III presents the demographic, education and training background of the participants, including the number in each category (Count), their average age, gender as a proportion of females to the total count and education level achieved as a proportion of number of participants in the group possessing the degree to the total number of participants in that group (e.g. 33% of laypersons held bachelors degrees). The training is a self-reported measure of the number of years of experience in information technology, software engineering, computer science, or computer engineering (for technical training) or law (for legal training). Of the 53 participants with complete responses, 25 (47%) came from technical backgrounds with an average of 6.6 years of training and 5.0 years of experience. Technical professionals were on average younger and more educated than laypersons. Legal experts had an average of 5.3 years of legal training and 11.3 years of experience, and all possessed Juris Doctors, except for one English resident possessing an LLB (Bachelor of Law), which is often used to practice law outside the United States.

TABLE III. Demographics, Education, and Training/Exp. by Group

| | | LAY | TECH | LEGAL | ALL |
|---|---|---|---|---|---|
| Count | | 21 | 25 | 7 | 53 |
| Avg. Age (*years)* | | 39 | 32 | 42 | 36 |
| Gender *(% female)* | | .52 | .28 | .29 | .38 |
| EDU. | Associates *(%)* | .19 | .24 | .14 | .21 |
| | Bachelors *(%)* | .33 | .80 | 1.00 | .64 |
| | Masters *(%)* | .00 | .32 | .86 | .26 |
| | Doctorate *(%)* | .00 | .08 | .00 | .04 |
| Tech. Training *(years)* | | 0.0 | 6.6 | 2.4 | 3.4 |
| Tech. Experience *(years)* | | 0.0 | 5.0 | 1.4 | 2.5 |
| Legal Training *(years)* | | 0.0 | 0.3 | 5.3 | 0.9 |
| Legal Experience *(years)* | | 0.0 | 0.0 | 11.3 | 1.5 |

* Three participants encountered computer issues during the survey; their survey times have been excluded from the average.

Table IV describes the MBE test results, including mean, max, and minimum scores by group and subject, as well as the average total time members of each group spent on the text. Because each question had only one correct answer, the score is computed as a simple ratio of correct answers to total number of questions. While the legal experts on average outperformed laypersons and technical professionals on the small sample of MBE questions, the top performers in the layperson and technical professional groups achieved scores higher than the lowest-performing legal expert. All groups performed best on questions regarding tort law. There are a number of potential explanations for this result, such as individuals lacking legal training or experience having established some degree of competence through other means, or specializations among legal experts that depart from areas of law relevant for the MBE. We intend to further explore these results as well as alternative means to measure legal expertise through knowledge testing.

TABLE IV. MBE Performance by Subject and Group

| | | LAY | TECH | LEGAL | ALL |
|---|---|---|---|---|---|
| Mean | | .39 | .39 | .52 | .41 |
| Max | | .55 | .60 | .70 | .70 |
| Min | | .20 | .15 | .30 | .15 |
| SUBJ. | Tort Law | .52 | .52 | .66 | .54 |
| | Contracts | .38 | .42 | .63 | .43 |
| | Constitutional Law | .33 | .29 | .46 | .33 |
| | Criminal Law | .32 | .35 | .34 | .34 |
| Average Time* *(min:sec)* | | 26:07 | 29:59 | 19:50 | 27:10 |

Figures 3 and 4 indicate participants' self-ratings regarding technical and legal knowledge and expertise. Despite having more domain experience than technical professionals, no legal expert claimed to be an expert in their field, whereas many technical professionals did[1]. As there are accepted but not definitive criteria for expertise, we wished to acknowledge the possibility, albeit remote, that an individual may be an expert without this status being reflected in other measurements, such as training. In the event this occurred, the individual would be contacted to validate their claim and obtain further information regarding their background. This did not occur in our study.

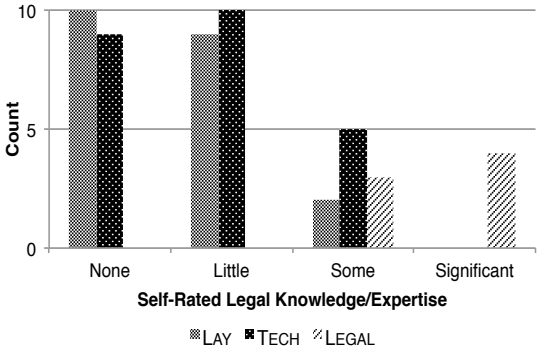



Fig 3. Self-Rated Legal Knowledge/Expertise by Group

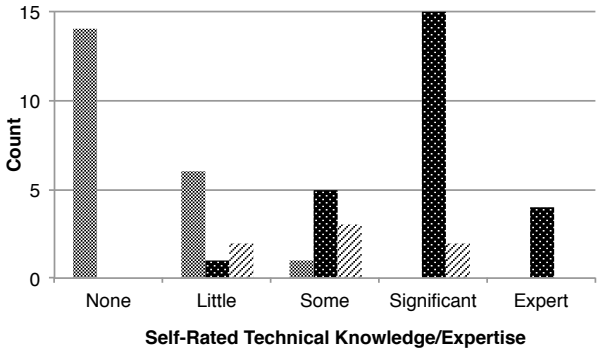



Fig 4. Self-Rated Technical Knowledge/Expertise by Group

Additionally, participants were asked how frequently they read laws and statutory texts, ranging from multiple times a day to never; results are shown in Figure 5. This was done to check for the possibility that an individual may consider herself to have no legal experience but spend a significant amount of time working with such texts, which could also contribute to their ability to function as a legal expert. Laypersons that claimed to read laws and statutory texts "a few times a week" all self-rated themselves as "Little" or "None" on the legal knowledge and expertise scale shown in Figure 3.

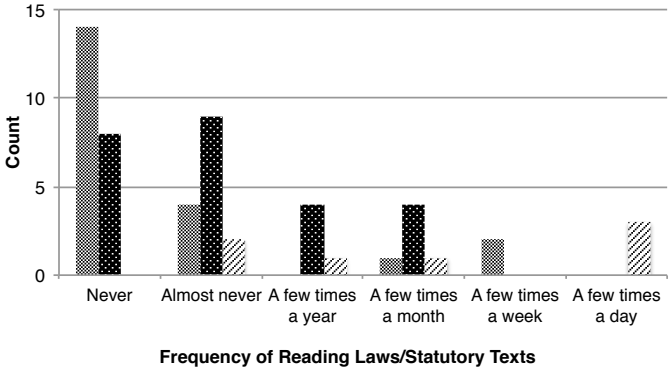



Fig 5. Self-Attested Frequency of Reading Laws and other Statutory Texts by Group

[1] We light-heartedly refer to this phenomena as "nerd arrogance".

## VII. Quantitative Results

Regarding $RQ_1$, we first assessed each group's within-group consistency by determining how frequently the group achieved a certain consensus percentile; results are shown in Table V. Each cell in this table indicates the number of comparison questions out of 20 answered by that group that achieved a consensus within the range in the leftmost column: e.g., legal experts achieved a consensus between 60-50% on 6 (30%) of the questions. Legal experts were the only group to achieve perfect consensus, doing so for 3 (15%) questions, and in general achieved higher consensuses more often than the other groups, indicating a greater uniformity in interpretation. The final column is not a total for each row.

TABLE V. Percent of Questions in Consensus Percentile by Group

| Consensus Range | Comparison Questions *(count; out of 20)* | | | |
|---|---|---|---|---|
| | Lay | Tech | Legal | All |
| > 90% | 1 | 1 | 3 | 0 |
| 90 - 80% | 1 | 1 | 0 | 2 |
| 80 - 70% | 2 | 3 | 5 | 3 |
| 70 - 60% | 6 | 4 | 0 | 1 |
| 60 - 50% | 4 | 2 | 6 | 3 |
| 50 - 40% | 4 | 6 | 5 | 8 |
| < 40% | 2 | 3 | 1 | 3 |

In Table VI, we show the inter-rater agreement calculated by Fleiss' κ, which differs from consensus measures used in Table V by factoring in the frequency of all responses for each question rather than just the mode. The possible range of values for Fleiss' κ is from 0 to 1, where 0 means no agreement beyond that expected by chance and 1 means complete agreement. Each row reflects the type of comparison made (between coverage conditions or requirements) with the latter broken down into obligations and prohibitions. All groups achieved fair to poor agreement [1] for each question type, with legal experts achieving greater agreement than laypersons and technical professionals regarding coverage questions (κ = .22, .27, and .34 for laypersons, technical professionals, and legal experts, respectively) but not requirements. In general, groups achieved equal or greater agreement on coverage questions than requirements questions. It should be noted that not all disagreements are qualitatively equivalent, as described in Section VIII.

TABLE VI. Within-Group Agreement (Fleiss' κ) by Comparison Type and Group

| Question Type | Agreement *(Fleiss' Kappa)* | | | |
|---|---|---|---|---|
| | Lay | Tech | Legal | All |
| Coverage | .22 | .27 | .34 | .23 |
| Requirements | .23 | .20 | .19 | .19 |
| - Obligations | .25 | .19 | .12 | .18 |
| - Prohibitions | .17 | .20 | .28 | .18 |

When addressing the extent of agreement between groups ($RQ_2$), we first examined how often each group assigned each relationship category across all questions (see Table VII) as a means to describe high-level agreement. The frequencies of each category are similar between groups, with subsumption the greatest and partial the least. Technical professionals assigned the partial category more so than other groups, particularly laypersons (20% compared to 13%). We verified that participants comprehended the directed relationships (e.g. subsumption, inclusion) by checking that their supplemental explanations did not conflict with their given categorical assignment.

TABLE VII. % of Relationship Type by Group

| Relationship | Lay | Tech | Legal | All |
|---|---|---|---|---|
| Equal *(%)* | .24 | .23 | .24 | .24 |
| Subsumption *(%)* | .35 | .32 | .35 | .34 |
| Inclusion *(%)* | .27 | .25 | .23 | .26 |
| Partial *(%)* | .13 | .20 | .18 | .17 |

While Table VII indicates that the groups assigned each relationship with similar frequency, the consensus-based Cohen's κ and Van Belle's NAI shown in Table VIII both offer a more detailed look at the agreement between groups, as they account for the co-occurrence of these assignments. Question types are the same as those found in Table VI but have been abbreviated for space: COV corresponds to Coverage, $REQ_{All}$ to all requirements, $REQ_O$ to Obligations, and $REQ_P$ to Prohibitions. Measurements ranges are identical to Fleiss' κ and interpreted similarly, with a measurement of 0 for no agreement beyond that expected by chance to 1 for complete agreement. Each column in Table VII corresponds to a group (e.g., LA~L compares laypersons to legal, etc.) with the rightmost column comparing laypersons to technical professionals, collectively, and legal experts. Results using a consensus-based Cohen's κ show that for all question types there is greater agreement between technical professionals and legal experts than laypersons and legal experts, with all group pairs showing higher between-group agreement on coverage questions compared to requirements questions. With regards to requirements questions, we discovered that, although there is low agreement between laypersons and legal experts, technical professionals agree with both laypersons and legal experts equally, indicating that they agree with each differently and function as intermediaries. The results using Van Belle's NAI slightly higher on prohibitory requirements. This is because the statistic accounts groups with low within-group agreement for this question type (see Table VI).

TABLE VIII. Between-Group Agreement by Group and Relationship Type using a Consensus-based Cohen's κ and VanBelle's Nominal Agreement Index (NAI)

| | Q. Type | Group Pair (*LA: Laypersons; T: Tech; L: Legal*) | | | |
|---|---|---|---|---|---|
| | | LA ~ L | T ~ L | LA ~ T | (LA,T) ~ L |
| Consensus Cohen's Kappa | COV | .42 | .72 | .43 | .57 |
| | $REQ_{ALL}$ | .19 | .46 | .46 | .32 |
| | - $REQ_O$ | .17 | .33 | .38 | .36 |
| | - $REQ_P$ | .27 | .64 | .56 | .27 |
| VanBelle's NAI | COV | .44 | .62 | .74 | .55 |
| | $REQ_{ALL}$ | .42 | .56 | .73 | .51 |
| | - $REQ_O$ | .36 | .59 | .83 | .48 |
| | - $REQ_P$ | .71 | .75 | .88 | .77 |

## VIII. Qualitative Responses

In order to answer $RQ_3$, participants were required to provide open-text explanations for half of the comparison questions. Although the explanations given by laypersons and technical professionals were consistently shorter than those provided by legal professionals (averaging 148, 113, and 192 characters, respectively) some responses focus on the same words or phrases in a question (e.g. "'paper-based method' in requirement A could mean...") and provided similar justifications. In this section we provide a brief overview of the notable differences in explanations offered by our legal experts as well as those by laypersons and technical professionals.

### A. Legal Expert Explanations

Most distinctive of the explanations given by legal experts were those provided when claiming two terms had a partial relationship. Unlike non-legal participants, whose explanations tend to identify specific words that justify the partial relationship, the explanations offered by legal experts not only identify such words but also provide hypothetical examples to illustrate. Legal experts' explanations frequently recognize or hypothesize that words were "terms of art," such as the various patient classifications referenced in Section III, which we reproduce here:

| Requirement A | Requirement B |
|---|---|
| [the organization] must maintain a medical record for every person evaluated or treated as an inpatient, ambulatory patient, emergency patient or outpatient | [the organization] must maintain a medical record on all patients admitted or accepted for treatment |

In this and similar instances, legal experts were reluctant to claim equal, subsumptive, or inclusive relationships without specifying the additional information needed; for example, in reference to the term "*reasonable belief* of unauthorized acquisition of personal information", one legal expert stated comparison depends on whether an "accusation of a breach that never actually occurred" would count as a reasonable belief. Some laypersons and technical professionals handle these cases more simply, such as by claiming that requirement B applied to a broader class of patient due to having more cases present.

### B. Non-Legal Explanations

Non-legal participants were more likely to interpret different words or phrases as having equivalent meanings, if they believed these words to have the same underlying intent. For example, when considering the two requirements below, which more than half of non-legal participants claimed were equivalent, many participants explicitly stated the two phrases mean "the same thing", one even stated "I imagine lawyers get a kick out of constructing arguments for why the phrases 'in the most expedient time possible' and 'without unreasonable delay' might mean different things. I am not one of those people":

| Requirement A | Requirement B |
|---|---|
| [the organization] shall make the disclosure in the most expedient time possible | [the organization] shall make the disclosure without unreasonable delay |

Perhaps most interestingly, some explanations offered by technical participants demonstrated the participant's reliance on formal logic to make comparisons. For example, when comparing definitions that used conjunctions, participants referenced "the lack of associativity of mixed Boolean operators" and set theory in their explanations. One participant went so far as to explain his answer symbolically as follows: "1) x or (y and z) = (x or y) and (x or z) 2) (x or y) and z = (x and z) or (y and z)".

## IX. Threats to Validity

We now discuss threats to validity and our mitigations.

### A. Construct Validity

Construct validity reflects whether the measurements actually measure that which they are intended to measure [39]. Requirements and definitions used in comparison questions were obtained using previously validated methods for legal requirements extraction [6], have appeared in prior published work [17, 18, 19], and were reviewed by the authors before final inclusion in the survey as described in Section V. As there is no widely agreed definition for legal or technical expertise, we drew on multiple measurements, including experience, training, certifications, and knowledge testing. Similarly, within- and between-group agreement were measured using multiple statistics, with the consensus-based approach in particular chosen to reflect how small firms absent legal resources may attempt to mitigate this shortcoming by combining interpretations from multiple individuals.

### B. Internal Validity

Internal validity is the extent to which a causal relationship exists between two variables or whether the investigator's inferences about the data are valid [39]. In this study, the technical participants recruited averaged approximately 10 years younger than the participants in other groups, which

could serve as an alternate explanation for differences in agreement. As the younger average for technical professionals was due primarily to a much higher representation of individuals in the 21-26 age range than in layperson and legal expert groups, we recalculated all statistics using a subset of participants that excluded this age group. In so doing, we found that all conclusions presented in this work remain the same regardless of the presence of age as a factor.

### C. External Validity

External validity is the extent to which findings generalize [39]. In this study, we used data obtained from a single domain of data privacy and security, which is less settled than other legal domains, such as tax law [24]. The data sanitization process mentioned in Section V involves omission and generalization of words or phrases in legal definitions and requirements in order to focus our analysis on fewer, direct comparisons, which may have oversimplified the space of interpretation. In all instances, this causes the resultant terms to become more similar, meaning that unsanitized data may show fewer equal categorizations than exhibited in our dataset. In future work we hope to explore differences in interpretations in other legal domains, as well as into the methods used to clarify and subsequently align unfamiliar terms, such as the use of a medical dictionary when analyzing laws regarding medical records.

Legal experts represent only 13% of our sample, compared to 47% and 40% for technical professionals and laypersons, respectively. We believe this is largely due to the value of their time: recruiting 25 experts for a more representative sample using their average rate of pay of $400 per hour would cost $10,000 [20]. This is a major concern for research going forward, and motivates alternative means for gaining access to legal experts or finding adequate experimental proxies, such as paralegals or senior law students. While our study is qualitative in nature, we believe the findings offer deeper insight than what could have been obtained using a different study design.

## X. Discussion and Summary

In this paper, we present the results of an exploratory study investigating the role of legal expertise in the interpretation of legal texts governing data privacy and security. To do so, we administered a survey to laypersons, technical professionals, and legal experts in which they were asked to categorize the relationship between related pairs of legal definitions and legal requirements extracted from domestic and foreign laws regarding data protection, medical record privacy, and data breach notification. We discovered that while legal experts achieved greater degrees of within-group consensus (e.g. 70% agreement or greater) on their categorizations more often than laypersons or technical professionals, their overall agreement was only marginally higher whether categorizing definitions (Fleiss' K: .34, .27, .22) or requirements (Fleiss' K: .28, .20, .17). More significantly, we also found that technical professionals and legal experts show considerably greater between-group agreement than laypersons and legal experts, particularly when categorizing definitions (consensus-based Cohen's K: .72 v. .42). These results are encouraging, suggesting that there may be value in technical professionals performing basic preliminary analysis of laws that affect their system design and development decisions.

In evaluating the explanations that participants offered for their responses, the authors discovered differences in approaches used by each group that merit further investigation. For example, legal experts often justify differences through case-based reasoning, providing instances of each term as evidence to support similarity and dissimilarity. This is in contrast to technical professionals, one of whom employed Boolean logic-based reasoning, which involves treating terms as abstractions. Should further study discover that the instance-based approach is common among legal experts, technical professionals may aim to achieve greater agreement by adopting this approach. Alternatively, legal experts and law firms working with technical clientele may better justify and communicate their opinions by mapping instances to abstract categories. The presence of mixed approaches supports the need for individuals with inter-disciplinary backgrounds that are capable of using both to serve as intermediaries between technical and legal functions.

The findings of this study provide a number of directions for the research community to pursue, including (i) identification of suitable proxies for legal experts in experimental settings, (ii) deeper investigation into the characteristics of statements and laws that produce substantial disagreement within and between different groups, (iii) assessment of weights and considerations given to external or non-textual matters affecting interpretation (e.g. legislative history, case law), (iv) determination of the collective effects different interpretations have on subsequent system design, and (v) further validation of existing between-group reliability statistics.

## Acknowledgment

We would like to thank a number of individuals for their advice and guidance regarding legal matters, survey design, and statistics, including Melanie Teplinsky, Baruch Fischhoff, and Will Frankenstein, among others. We would also like to thank the International Association of Privacy Professionals (IAPP) for allowing us to recruit survey participants through their Global Privacy Summit. This research was supported by the Hewlett-Packard Labs Innovation Research Program (Award #CW267287) and National Science Foundation (NSF) IGERT Award #0903659.

## References

[1] D. Altman, Practical Statistics for Medical Research, Chapman and Hall, 1991.

[2] Baker & McKenzie, *Global Privacy Handbook*. 2013.

[3] D.M. Berry, E. Kamsties, M.M. Krieger. “From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity,” Univ. of Waterloo Technical Report, November 2003.

[4] T.D. Breaux, M.W. Vail, and A.I. Anton, “Towards Regulatory Compliance: Extracting Rights and Obligations to Align

Requirements with Regulations," Proc. 14th IEEE Int'l Requirements Eng. Conf., pp. 46-55, 2006.

[5] T.D. Breaux, A.I. Anton, K. Boucher, M. Dorfman, "Legal requirements, compliance and practice: an industry case study in accessibility." IEEE 16th Int'l Req'ts Engr. Conf., pp. 43-52, 2008.

[6] T.D. Breaux, *Legal Requirements Acquisition for the Specification of Legally Compliant Information Systems.* Ph.D. Thesis, North Carolina State University, 2009.

[7] L.M. Christensen, "The Paradox of Legal Expertise: A Study of Experts and Novices Reading the Law," B.Y.U. EDUC. & L.J., issue 53, 2008.

[8] L. Christensen, A. Colciago, F. Etro and G. Rafert, "The Impact of the Data Protection Regulation in the E.U," International Think-tank on Innovation and Competition, 2013.

[9] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement , issue 20, 37-46.

[10] A. Cravens, "A Demographic and Business Model Analysis of Today's App Developer," GigaOm, 2012.

[11] K. Anders, Ericsson et al., *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, 2006.

[12] Ernst & Young. "2006 Global Information Security Survey," November 2006.

[13] W.N. Eskridge and P.P. Frickey, "Statutory Interpretation as Practical Reasoning," Stanford Law Review, vol. 42, 1990, pp. 321-384.

[14] P.D. Feigin and M. Alvo, "Intergroup diversity and concordance for ranking data: an approach via metrics for permutations," The Annals of Statistics issue 14, 1986, pp. 691–707.

[15] J.L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, issue 76, 378-382.

[16] S. Ghanavati, D. Amyot, and L. Peyton, "Compliance Analysis Based on a Goal-oriented Requirements Language Evaluation Methodology," 17th IEEE International Requirements Engineering Conference, 2009, pp. 133-142.

[17] D.G. Gordon and T.D. Breaux, "Reconciling multi-jurisdictional legal requirements: a case study in requirements water marking," 20th IEEE International Requirements Engineering Conference, 2012, pp. 91-100.

[18] D.G. Gordon and T.D. Breaux, "A Cross-domain empirical study and legal evaluation of the requirements water marking method." 20th IEEE Requirements Engineering Journal, 2013, pp. 1-27.

[19] D.G. Gordon and T.D. Breaux, "Assessing Regulatory Change through Legal Requirements Coverage Modeling," 21st IEEE Int'l Req'ts Engr. Conf., 2013, pp. 145-154.

[20] S.J. Harper, "The Tyranny of the Billable Hour," The New York Times, 2013.

[21] J. Hartley, "Legal Ease and 'Legalese'," Psychology, Crime, & Law, volume 6, issue 1, 2000, pp. 1-20.

[22] O.W. Holmes, "The Theory of Legal Interpretation," Harvard Law Review, volume 12, No. 7, 1899, pp. 417-420.

[23] S. Ingolfo, A. Siena, and J. Mylopoulos, "Establishing Regulatory Compliance for Software Requirements,"" 30th IEEE International Requirements Engineering Conference, pp. 47-61, 2011.

[24] B. Johnston, G. Governatori. "Induction of Defeasible Logic Theories in the Legal Domain," Proc. of the 9th Int'l Conf. on AI and Law, pp. 204-213, June 2003.

[25] L. Kaufman, P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley Publishing, 1990.

[26] S. Kerrigan, K.H. Law. "Logic-Based Regulation Compliance-Assistance," Proc. of the 9th Int'l Conf. on AI and Law, pp. 126-135, June 2003.

[27] H.C. Kraemer, "Intergroup concordance: definition and estimation," Biometrika, issue 68, 1981, pp. 641–646.

[28] H.C. Kraemer, S.P. Vyjeyanthi, and A. Noda "Dynamic Ambient Paradigms," Tutorial in Biostatistics vol 1., pp. 85–105, 2004.

[29] National Conference of Bar Examiners, "Multi-State Bar Exam," Online: http://www.ncbex.org/about-ncbe-exams/mbe/, 2014.

[30] P.N. Otto, A.I. Anton, "Addressing legal requirements in requirements engineering." 15th IEEE Int'l Req'ts Engr. Conf., pp. 5-14, 2007.

[31] Ponemon Institute LLC, "Third Annual Benchmark Study on Patient Privacy and Security," 2012.

[32] J. Redish, "How to Draft more Understandable Legal Documents," from Drafting Documents in Plain Language. 1979.

[33] A. Rifaut, S. Ghanavati, "Measurement-oriented comparison of multiple regulations with GRL," IEEE 5th Workshop on Requirements Engineering and Law (RELAW), pp. 7–16, 2012.

[34] A. Scalia, K-Mart Corp. v. Cartier, Inc., 108 S. Ct. 1811, 1831-34, 1988.

[35] J. Shanteau, "Decision making by experts: The GNAHM effect," Decision Science and Technology: Reflections on the Contributions of Ward Edwards, 1999, pp. 105-130.

[36] J. Shanteau, D.J. Weiss, R.P. Thomas, J.C. Pounds, "Performance-based Assessment of Expertise: How to Decide if Someone is an Expert or not," European Journal of Operations Research, 136, 2002. 253–263.

[37] A. Siena, I. Jureta, S. Ingolfo, A. Susi, A. Perini, J. Mylopoulos. "Capturing Variability of Law with Nomós 2," 31st Int'l Conf. Conc. Mod., pp. 383-396, 2012.

[38] S. Vanbelle. "Agreement between raters and groups of raters." Doctoral Dissertation, Universite de Liege. 2009.

[39] R. K. Yin, Case Study Research: Design and Methods, 4th edition, Sage Publications, 2009.