

4. Automated classification of modes of moral reasoning in judicial decisions

Nischal Mainali, Liam Meier, Elliott Ash and Daniel Chen*

1. INTRODUCTION

The legal profession is undergoing a great transformation. The tools of machine learning and causal inference can be used to increase efficiency and fairness of the law. In this chapter, we discuss how these tools can also be used to detect how judges motivate their decisions. What is the role of moral reasoning in judicial decision-making?

Law is divided between two modes of moral reasoning. On the consequentialist view, optimal policy should be based on calculations of costs and benefits. The deontological view holds that policy should be determined from moral duties – what is right and just. One way of quantifying this divide empirically is to identify and measure the use of different modes of moral reasoning in judicial decisions. The goal of this project is to use computational techniques to automatically classify judicial decisions by the type of moral reasoning employed.

Computational linguistics, typically referred to by computer scientists as natural language processing (NLP), utilizes computational techniques to translate, make sense of, and produce material in human languages. Human language is complex, full of ambiguities, encoded knowledge, and variety. Though state-of-the-art NLP systems still struggle with tasks that are simple for humans (disambiguation, for example) statistical techniques that make use of large data sets have been highly successful in sentiment analysis, classification, and more (e.g. Hirschberg and Manning, 2015).

Two broad moral frameworks most often used by philosophers are

* These authors contributed equally.

consequentialism and deontology. For a consequentialist, what one ought to do is whatever brings about the best consequences. The canonical form of consequentialism is utilitarianism, according to which ethical agents are obligated to do whatever brings about the most utility. For hedonic utilitarians like J.S. Mill and Jeremy Bentham, pain is bad and pleasure good. While philosophers have debated about how to measure utility, the idea is broadly in line with preferring happiness over unhappiness (Sinnott-Armstrong, 2015).

In a deontological framework, an act is right if it conforms to a moral norm. For Kant, this is the categorical imperative, the first formulation of which holds that one should act only according to a maxim which they can will to be a universal law (Alexander and Moore, 2016). The idea is that morality is a matter of conforming to ethical law instead of acting in order to bring about certain consequences.

Applied ethicists argue about the morality of various issues from these positions. Take the permissibility of lying, for example. A consequentialist might argue that it would be permissible to lie if the consequences of telling the lie would be better than not telling the lie. In contrast, a deontologist instead might argue that it is immoral to lie because one could not will the acting of lying to be a universal law.

Law commonly operates in correspondence with morality. Some actions are made illegal not because of practical considerations, but because a society believes that the action is morally wrong. For example, laws against sex work are often justified on moral grounds.

In the legal system, we can observe when judges reason according to either of these moral frameworks. In the case of a contract, a deontological judge may hold that breach is always illegal, since it is a break of a promise. A consequentialist judge may reason that certain breaches of contract are acceptable, if, for example, a breach of contract would lead to a better outcome for the relevant parties. Such reasoning is recognized as “efficient breach” in law and economics.

In this chapter we build a classifier using tools from NLP and machine learning, trying to identify modes of moral reasoning. Moral reasoning is often studied in relation to political ideology (Fishkin et al., 1973; Graham et al., 2009; Emler et al., 1983). Recently researchers (Sagi and Dehghani, 2013) employed computational techniques to measure texts amongst the five moral dimensions posited by Jonathan Haidt (Haidt and Joseph, 2004). To the best of our knowledge, our work is the first attempt to computationally classify moral reasoning into the conventional theoretical categories of ethics.

Here we operate on the assumption that patterns in written language reflect, however imperfectly, different types of reasoning. For example,

a consequentialist would use language associated with ends. We might also expect language that attempts to infer the results of different types of actions and a higher frequency of words that designate consequence. In contrast, a judge reasoning deontologically might tend to use more absolute language and employ words like duty and obligation more often.

To train our classifier, we use a corpus of articles from the applied ethics literature. In these articles, philosophers argue about issues like abortion, vegetarianism, and war. The articles are drawn from an archive of philosophical literature, where they were categorized as either consequentialist or deontological. Our trained classifier assigns holdout samples to the proper class with over 90 percent accuracy.

We then apply the classifier to circuit court opinions, where judges outline their reasoning for the decision on a case. We use these opinions (dating back to 1883) to analyze trends in moral reasoning in the US legal system. We look at rates of consequentialist versus deontological reasoning over time, according to where the individual was born, where the individual attended school, their gender, and the party of the president under which they were nominated.

2. AUTOMATED CLASSIFICATION OF MORAL REASONING

2.1 Training Data

The training corpus comprises all articles from the philosophy paper database PhilPapers.org tagged with “Applied Consequentialism” or “Applied Deontology”. We filtered out papers written in languages other than English, papers that had both consequentialism and deontology tags, and papers that obviously did not conform to either category. The resulting training corpus consists of 14 Consequentialist papers and 11 Deontology papers. These were converted to plain text, and artifacts from the PDF conversion were expunged when possible.

Our data set is composed of a few large texts. But most text classification techniques perform better with larger corpora. Accordingly, we separated each of our Consequentialist and Deontology corpa into 100 equally sized chunks. This introduces risk of overfitting, which we address by verifying performance over multiple random seeds that were used for separation of the data into the training and validation set.

Next we featurize the text. We tried many approaches, including bag-of-words, various lengths of n-grams, tf-idf, and so on, for featurizing

the text. After some hyperparameter tuning, we chose tf-idf with n-gram featurization, which we found performed best in the holdout test set. The tf-idf statistic is an indicator of the relative importance of a word or phrase to a specific document. We take phrases of up to 3-words-length and adjust them by their document frequency to create a list of tf-idf n-grams. Each document is represented as a sparse vector representing the frequency distribution over these n-grams. Because these vectors are high-dimensional, the algorithm proceeds by using singular value decomposition to reduce the dimensionality to a lower-dimensional manifold.

2.2 Model Training

The preprocessing step leaves us with a vector assigned to each of the texts to use for training the machine learning model. A natural choice for classification problems such as ours is Linear Support Vector Machine (SVM) or Naive Bayes. Initial performance of Naive Bayes was very poor, so we focused our efforts on implementation of the SVM, which is known to be highly effective at text classification (Joachims, 1998).

In broad terms, Linear-SVM tries to find a hyperplane that neatly separates the vectors with “Deontology” labeling and “Consequentialist” labeling. On one side of the hyperplane we find consequentialist text and on the other side we find deontological texts. The prediction is made by locating the vector of the test text in relation to the hyperplane. We use a multi class SVM, with an additional label “other”, trained on philosophy papers that did not have the consequentialism or deontological label. This measure reduces spuriously categorizing moral reasoning into the categories when in fact there is no moral reasoning being used at all.

During the initial steps in the training, we separate our data set into training and validation sets. After each training, the model predicts on the validation set and we tweak the training parameters to increase accuracy and get sensible features. Major tuning components in training the model were frequency threshold and stop words. A sensible frequency threshold prevents especially uncommon n-grams from influencing the classification. Stop words are common words filtered out of a dataset before introducing the data to a model.

Our model very quickly approached a 100 percent prediction accuracy in the validation set, so the main hyperparameter tuning step was finding junk features in the prediction function and removing them. For example, the papers were downloaded on a university license, so we removed the terms that came from the license watermark. There were other overfitting type terms (e.g. “fetus”, from the high-frequency of medical ethics papers),

and junk phrases from the PDF conversion such as “x-86”. These were also removed.

This was done through examination of the prediction function. Linear-SVM predicts by assigning weights (coefficients) to the n-grams. We looked at the top 50 predictors for both Consequentialist reasoning and Deontological reasoning and weeded out obviously irrelevant n-grams.

We also experimented with n-grams by adjusting the number of words per phrase. $N=3$ gave good results. Further increasing N helped very little and would have introduced risk of further overfitting.

Finally, we used Latent Dirichlet Allocation (a method frequently used to model topics) to check if any unobserved groups or topic bias are influencing the prediction. We found that the topics were varied and uniformly distributed, an indication that there was no such bias.

2.3 Feature Importance

The end product of the training is a weight assignment to each n-gram in the feature set. We rank the n-grams by their weights. In our model an n-gram with a positive weight means that it tends to occur in deontology articles, while a negative weight signals that the text is consequentialist.

The set of most predictive n-grams for each category are visualized in Figure 4.1’s word clouds. Some of the features are somewhat intuitive. For example, “pleasure” is a consequentialist value, while “duty” is a deontological value.

3. APPLICATION TO US CIRCUIT COURTS

The application corpus is the universe of US Circuit Court opinions from the years 1883 through 2013. Besides the text of the opinions, we have some relevant metadata, such as biographical details of the judges writing the opinions.

3.1 Prominent Paragraphs

To understand better how our classifier works in the judicial corpus, we ranked the paragraphs in the corpus by the mode of moral reasoning used. We list some of those paragraphs here. We can read these paragraphs and see how our classifier works in the new context.

duties and determined that 201 duties are “special dumping duties” because they are “more like AD [antidumping duties] in purpose and function than they are like ordinary customs duties.” *Id.*

[I]f an insurer who refuses to defend were estopped from asserting the lack of coverage as a defense in a subsequent action, then the insurer’s duty to indemnify would be coextensive with its duty to defend. [The Maine Law Court], however, ha[s] repeatedly stated that an insurer’s duty to indemnify is independent from its duty to defend and that its duty to defend is broader than its duty to indemnify.

By reason of its nature as a public institution St. Elizabeth’s Hospital owes a duty to the public in carrying out its difficult responsibilities. We have no occasion now to decide, however, whether its public duty included an entirely separate duty to Mrs. Morgan. There was a particular duty to the Court of General Sessions, and in the circumstances of this case it was intertwined with a duty to her. See *infra*, note 12.

3.2 Strongly Consequentialist Prediction

The following paragraphs are considered consequentialist by the prediction function with high confidence:

No rule of bankruptcy practice and procedure is designed to be considered in isolation. Each rule is to be considered in conjunction with every other rule. What the entire body of rules makes available to the practitioner and the bankruptcy judge is a gestalt designed to constitute a functional whole. The rules are not a melange of independent parts. The Advisory Committee alluded to this in its preface to the rules: “The proposed rules are not divided into chapters related to the different types of debtor relief chapters in the Code. These rules apply in all chapter cases except as a particular rule otherwise provides.” Preface to Rules and Forms, 11 U.S.C. XXI. One of the trustee’s most vigorous arguments, therefore, is explicitly contradicted by the Preface to the Rules. He argues that Rule 1019(4) does not apply to Chapter 7 cases. This is the rule that specifies that claims filed in the superseded case shall be deemed filed in the Chapter 7 case. The Advisory Committee’s preface clearly reveals that the rules apply in all chapter cases “except as a particular rule provides otherwise.” Rule 3002(a), requiring the filing of a claim in a Chapter 7 case, does not “provide[] otherwise.” So construed, there is but one proper resolution of this case.

Besides the nonrestrictive nature of the ordinary meaning of the claim term “code,” the doctrine of claim differentiation provides a powerful argument against construing the term “code” restrictively, to mean “spreading [**1615] code.” Independent claim 1 of the ‘966 patent uses the term “code,” and dependent claim 5 recites, in full, “The subscriber unit of claim 1 wherein the same code is a spreading code.” The clear implication of narrowing the term “code” in dependent claim 5 by limiting the claim scope to cases in which the

claimed code “is a spreading code” is that the term “code” in the independent claim is not limited to a spreading code.

However, the evidence submitted by Clark does not stand alone. Critically, Jerlene Bush and Mildred Bobo, both within job code 1433, testified that at the time of the RIF, they trained two employees, Carolyn Muse and James Russell Hunter, who then replaced them in their positions. See Appellants’ App. at 496-98. Yet Hunter was placed in job code 1432, see *id.* at 171, and Muse in 1402, see *id.* at 170.10 Muse’s placement in another job code generally raises questions about the claimed functional differences between job codes. But Hunter was placed into the very job code that plaintiffs argue was pretextually separated from their own code for purposes of discrimination. On the basis of personal observation, two witnesses from job code 1433 testified that they trained someone to perform their jobs and that the person they trained was placed in job code 1432. That is sufficient for a reasonable jury to find Seagate’s differentiation of the two codes pretextual.

First, we see that the statements considered deontological all contain the word “duty,” which is intuitive. The statements categorized consequentialist include phrases like “Each rule is to be considered in conjunction with every other rule”, emblematic of context-sensitive consequential reasoning.

Time series for consequentialist reasoning

Next we examine trends in consequentialist versus deontological reasoning over time (Figure 4.2). We see there is a discrete jump in consequentialism in the 1930s, indicating a major switch in thought at the time.

There could be many factors driving this change. One possibility is that the hardships of the depression brought disenchantment with prevailing norms; so moral attitudes shifted toward a focus on better outcomes, rather than strict adherence to laws regardless of outcomes. Another potential factor is the legal realism movement of the 1920s and 1930s, which viewed law as a means toward an end, rather than an end in itself (Posner, 1986).

Another explanation for the trend is that over time the language of circuit court opinions has shifted to that of the consequentialist articles. For example, words that are weighted as consequentialist like “code” and “rule” might be appearing more often now compared to the past because there are more cases being decided on statutory grounds instead of common-law grounds.

In the Appendix, we report rankings of legal topics by moral reasoning. Opinions written in estate law and family law have a deontological bent, while administrative law tends to be more consequentialist. As the

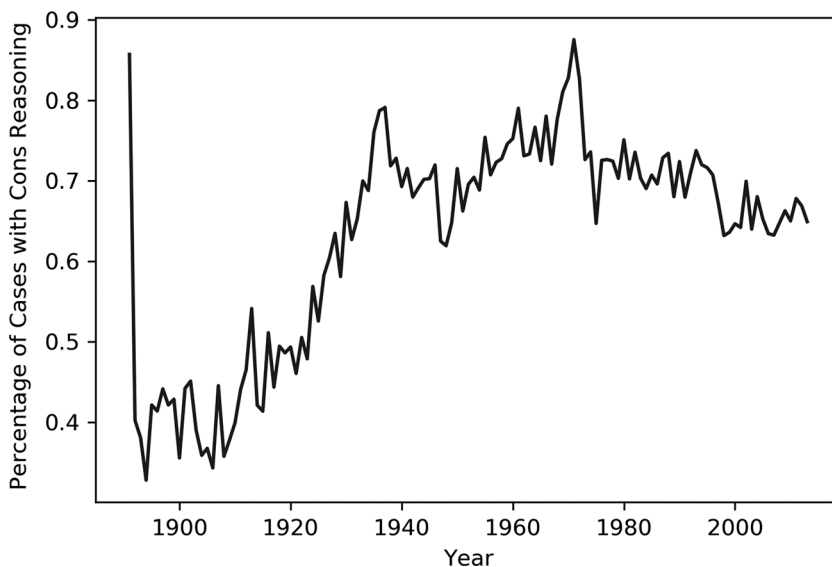


Figure 4.2 Consequentialist reasoning in the Circuit Courts, 1891–2013

popularity in topics have changed over time, some of the rankings may reflect when opinions in that topic were being written, because reasoning has become more consequentialist (e.g. the fact that there has been more administrative law in more recent times explains why more opinions in this topic are consequentialist). This effect could be disentangled in future work.

3.3 Judge Rankings

It might also be illuminating to learn which judges are most consequentialist or most deontological. For example, we might expect pragmatist judges such as Richard Posner to be consequentialist. To ensure a large enough sample size, we filter out all judges who published fewer than 50 opinions in our dataset.

Table 4.1 has the most deontological judges. Table 4.2 has the most consequentialist judges. We can see that this appears to be driven by judge cohorts and time effects. Interestingly, the most consequentialist judge, Neil Gorsuch, was recently promoted to the US Supreme Court.

Table 4.1 Top ten deontological judges

Judge	Percent Consequentialist
Trieber, Jacob	21.2
Van Devanter, Willis	21.9
Cotterall, John H.	24.4
Reed, Henry Thomas	24.6
Kenyon, William	25.3
McDowell, Henry Clay	25.8
Hawley, Thomas Porter	26.6
Booth, Wilbur F.	27.5
Philips, John Finis	27.6
Pritchard, Jeter C.	28.1

Table 4.2 Top ten consequentialist judges

Judge	Percent Consequentialist
Gorsuch, Neil M.	92.4
Moore, Kimberly Ann	91.3
Martin, Beverly B.	90.2
Tatel, David Stephen	88.8
Madden, Joseph	88.1
Clevenger, Raymond Charles, III	87.9
Wilkey, Malcom R.	87.8
Rich, Giles Sutherland	87.6
Starr, Kenneth W.	87.5
Patterson, Robert P.	87.3
Chambers, Richard H.	87.0

3.4 Moral Reasoning and Judge Characteristics

Here we look at what biographical characteristics of judges are associated with the use of more consequentialist versus deontological language.

First, we show in Figure 4.3 that there seems to be very little difference across gender in moral reasoning. While it appears that females are more likely to reason consequentially, the difference is not statistically significant. What difference we do see is likely due to more female judges being in office in later years, after the upward shift.

Next, we ask in Figure 4.4 whether political party affiliations matter for modes of reasoning. There appears to be no significant difference in

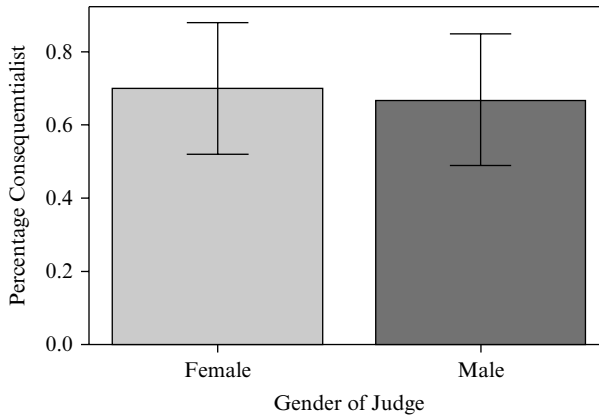


Figure 4.3 Consequentialist reasoning by gender

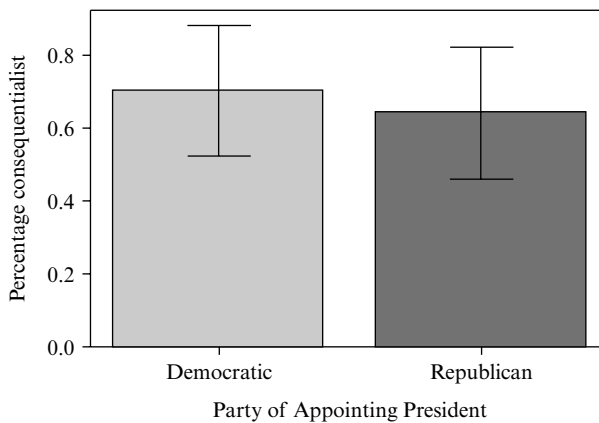


Figure 4.4 Consequentialist reasoning by political affiliation

consequentialist reasoning across party affiliations of the President who appointed the judge.

Another factor determining differences across judges in the mode of reasoning is their legal training. Are there differences between judges that come from different law schools? To answer this question we took the average use of consequentialism for the judges trained at each law school and then ranked them. We include all law schools where at least 1000 opinions by attendees are in the data set. This ranking is reported in Table 4.3, along with the percentage difference of that school from

Table 4.3 Ranking of judge consequentialism by law school attended

School	Percentage
Washington and Lee University School of Law	-19.6
University of North Carolina School of Law	-9.9
University of Wisconsin Law School	-9.5
University of Oxford	-3.5
University of Nebraska College of Law	0.5
St. Louis University School of Law	1.2
University of California, Berkeley	1.7
New York University School of Law	2.3
Columbia Law School	2.8
Cornell Law School	3.1
Syracuse University College of Law	3.6
Fordham University School of Law	4.0
University of Arkansas School of Law	4.3
University of Alabama School of Law	4.5
Harvard Law School	5.9
George Washington University Law School	5.9
Notre Dame Law School	6.0
Northwestern University School of Law	6.1
University of Utah College of Law	7.2
University of Washington School of Law	7.4
University of Southern California Law School	7.6
University of Virginia School of Law	7.8
Louisiana State University Law School	8.4
Yale Law School	8.4
University of Minnesota Law School	8.5
University of Chicago Law School	9.6
Tulane University Law School	9.6
University of Texas School of Law	11.1
University of Mississippi School of Law	11.7
University of Montana School of Law	11.7
Stanford Law School	12.1
Georgetown University Law Center	17.2

Note: Rankings of judge law schools by consequentialism

the global average. So, for example, judges who attended law school at Washington and Lee are on average reasoning consequentially about 20 percent less often than the average judge. There are large differences by law school attended.

Perhaps moral reasoning style is determined by where one grew up as well as where one attended law school. In Figure 4.5, we look at

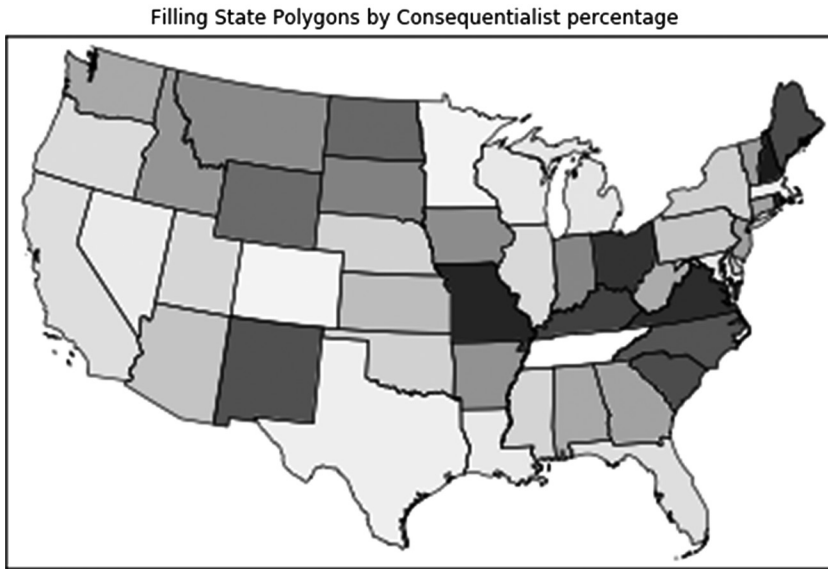


Figure 4.5 Consequentialist reasoning by judge birth state

differences by birth state. We find that judges from coastal states seem to have a relatively deontological leaning.

Red corresponds to highly consequentialist states, blue (darker) corresponds to highly deontological states. We refer readers of the text edition to the web edition to see the colors.

4. CONCLUSION

Utilitarian versus deontological modes of reasoning is a classic divide in moral philosophy and in economics and law. To understand human values, artificial intelligence (AI) systems will likely need to be able to detect and annotate when an argument is utilitarian or deontological. This chapter has demonstrated the use of computational linguistics and machine learning techniques for the problem of classifying moral reasoning in written texts. We show that even a small corpus of training articles can be utilized to reasonably infer patterns of moral reasoning. Such tools can then be used to help understand judging and the practice of law.

That said, we must caution against interpreting these results too strongly. How people reason, morally and otherwise, is a complex psychological and behavioral phenomenon. Oftentimes the reasoner herself cannot

correctly identify how they came to some conclusion. To do so from an outside perspective presents even more of a challenge.

Machine learning techniques are far from able to engage in the sophisticated and robust patterns of inference and reasoning that humans exhibit. Current machine learning techniques should be understood as effective but limited, essentially statistical methods that can be used to draw insights from data. This work demonstrates how these techniques might be used to try and answer especially difficult questions in law and elsewhere. We think that the major conclusions suggested by our classifier are reasonable. It is difficult to assess how accurate the more specific findings might be.

There is room for plenty of future work. One approach might be to classify assignments of obligations and authority: for example, “the right is vested in party 1”, “party 1 has the right”, “the duty is assigned to party 1”, “party 1 must”, and so on, or to classify conditional language: “if A, then B” constructions. Another could be to observe whether economics-trained judges use different features of defendants consistent with cost-benefit analysis when obtaining sentences. A larger corpus of training data could be obtained, for example, by hand labeling instances of judges’ reasoning in a specific way in court opinions. Topic modeling and dimensionality reduction should be further pursued in order to better understand how different populations reason morally.

REFERENCES

- Alexander, Larry and Michael Moore (2016), “Deontological Ethics”, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, accessed 1 March 2018 at <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>.
- Emler, Nicholas, Stanley Renwick, and Bernadette Malone (1983), “The Relationship Between Moral Reasoning and Political Orientation”, *Journal of Personality and Social Psychology*, **45** (5), 1073–80.
- Fishkin, James, Kenneth Keniston, and Catherine McKinnon (1973), “Moral Reasoning and Political Ideology”, *Journal of Personality and Social Psychology*, **27** (1), 109–19.
- Graham, Jesse, Jonathan Haidt, and Brian A. Nosek (2009), “Liberals and Conservatives Rely on Different Sets of Moral Foundations”, *Journal of Personality and Social Psychology*, **96** (5), 1029–46.
- Haidt, Jonathan and Craig Joseph (2004), “Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues”, *Daedalus*, **133** (4), 55–66.
- Hirschberg, J. and C.D. Manning (2015), “Advances in Natural Language Processing”, *Science*, **349** (6245), 261–6.
- Joachims, Thorsten (1998), “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, in C. Nédellec and C. Rouveirol (eds), *Machine Learning: ECML-98*, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998, Lecture Notes

- in Computer Science (Lecture Notes in Artificial Intelligence), vol. 1398. Berlin and Heidelberg: Springer, pp. 137–42.
- Posner, Richard A. (1986), “Legal Formalism, Legal Realism, and the Interpretation of Statutes and the Constitution”, *Case Western Reserve Law Review*, **37** (2), 179–217.
- Sagi, Eyal and Morteza Dehghani (2013), “Measuring Moral Rhetoric in Text”, *Social Science Computer Review*, **32** (2), 132–44.
- Sinnott-Armstrong, Walter (2015), “Consequentialism”, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2015 ed.), accessed 1 March 2018 at <https://plato.stanford.edu/archives/win2015/entries/consequentialism/>.

APPENDIX

We can rank different legal topics by the percentage of opinions written within the topic that are labeled consequentialist:

Table 4A.1 Topics with consequentialism score

Topic	Percentage
Wills, Trusts and Estates	44.7
Negotiable Instruments	50.2
Torts	52.6
Real Property	54.0
Admiralty and Maritime	54.9
Native Peoples	56.7
Judicial Ethics and Conduct	57.1
Natural Resources	57.7
Mortgages and Liens	58.0
Real Estate Investment Trust	58.6
Damages and Remedies	58.8
Landlord and Tenant	59.1
Debtor Creditor	60.1
Bankruptcy Law	60.4
Alcohol and Beverage	61.2
Motor Vehicles and Traffic Law	61.5
Legal Malpractice	61.9
Personal Property	62.3
Eminent Domain	62.6
Corporate Law	62.7
Entertainment Law	63.3
Contracts	63.8
Professional Responsibility	64.9
Civil Procedure	66.1
Civil Rights	66.3
Medical Malpractice	66.6
Agency	66.7
Criminal Law	66.8
Mergers and Acquisitions	67.1
Class Actions	67.6
Transportation Law	68.5
International Trade Law	68.5
International Law	68.8
Art Law	68.9
Appellate Procedure	69.0
Government	69.5
Constitutional Law	69.7

Table 4A.1 (continued)

Topic	Percentage
Prisoners' Rights	69.9
Partnerships and Non-Corporate Business Entities	70.1
Insurance Law	70.5
Employee Benefits	70.6
Postal Service Law	71.6
Alternative Dispute Resolution	72.0
Habeas Corpus	72.8
Gambling and Lotteries Law	72.9
Workers' Compensation	73.2
Elections and Politics	73.7
Securities Law	74.1
Employment Law	74.5
Franchise Law	74.5
Banking and Finance	74.6
Land Use Planning and Zoning	74.8
Uniform Commercial Code	75.2
Evidence	75.2
Environmental Law	75.7
Products Liability	75.7
Agricultural Law	75.9
Government Employees	76.0
Copyright Law	76.1
Military Law	76.2
Consumer Law	76.5
Patent Law	76.7
Education Law	76.8
Construction Law	77.0
Tax and Accounting	77.5
Professional Corporations	77.8
Antitrust and Trade	77.8
Immigration and Naturalization	77.9
Trade Secrets	78.6
Trademark Law	78.8
Intellectual Property Treaties and Conventions	79.2
Conflict of Laws	79.2
Executive Compensation	80.0
Communications and Media	80.3
Religious and Non-Profit Organizations	80.9
Hazardous Material Law	81.3
Health Law	81.7
Government Contracts	81.9
Privacy and Information Law	82.1

Table 4A.1 (continued)

Topic	Percentage
Sports Law	82.6
Homeland Security	83.3
Social Security	83.4
Administrative Law	84.6
Labor Law	84.6
Energy Law	92.3
Technology Law	93.9

Note: Scores are a percentage of the opinions in that category that were classified consequentialist.