What's the matter with 'reasonable'?

Lucien Baumgartner (lucien.baumgartner@philos.uzh.ch)
Department of Philosophy, Zürichbergstrasse 43
8044 Zurich, Switzerland

Markus Kneer (markus.kneer@uzh.ch)
Department of Philosophy, Zürichbergstrasse 43
8044 Zürich, Switzerland

Abstract

The reasonable person standard is key to both Criminal Law and Torts. What does and does not count as reasonable behavior and decision-making is frequently determined by lay jurors. Hence, laypeople's understanding of the term must be considered, especially whether they use it predominately in an evaluative fashion. In this corpus study, we investigate whether laypeople use 'reasonable' mainly as descriptive, evaluative, or merely value-associated term, based on supervised machine learning models. We find that 'reasonable' is predicted to be an evaluative term in a majority of cases. This supports prescriptive accounts, and poses potential problems for descriptive and hybrid accounts of the term. Other terms often used interchangeably in jury instructions (e.g., 'careful,' 'ordinary,' 'prudent,' etc), however, are predicted to be descriptive. This indicates a discrepancy between the intended use of the term and the understanding lay jurors might bring into the court room.

Keywords: reasonable person standard; reasonableness; negligence; evaluative language; thick concepts; corpus linguistics; experimental jurisprudence

Introduction

The concept of reasonableness is key to practical rationality broadly conceived. As such it is of fundamental importance in our daily lives, economic decision making, and public governance. It also takes centre stage in the law, particularly in Common Law jurisdictions such as the UK and the US. Negligence in Torts is defined as a failure to exercise reasonable care (3rd Restatement of Torts, §3, see also Keating, 2022), criminal negligence is characterized as risk-taking that "involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation" (Modal Penal Code 2.02 (d)). The concept of reasonableness also plays a prominent role in constitutional law, contract law, administrative law and beyond (Gardner, 2015; Unikel, 1992; Zipursky, 2015).

But given its exceptional significance in decision making and the law, what exactly is reasonableness? "We can turn to legal, moral, or political theorists for clarity about reasonableness," Lawlor observes in a recent paper, "but often we find these theorists referring us back to our ordinary understanding." (2022, 1). The same is the case for the law, where decisions as to what does and does not constitute reasonable behavior and decision-making are frequently left to lay jurors. The law itself is rather tightlipped about the meaning of the expression 'reasonable,' and judges routinely refuse to elucidate it.

A central, and perhaps the most fundamental, debate concerning the expression 'reasonable' and the concept it denotes regards its type: whether it is a descriptive notion, capturing what is common, average or statistically likely, or whether it is an evaluative notion, referring to what is good, appropriate, or what an upright citizen would do. This question, which has "bedevilled and divided courts and scholars for centuries" (Miller and Perry, 2012, 323), is still hotly debated. Some have defended the descriptive account (Dressler, 1994; Zalesne, 1996). A greater number advocate the prescriptive account, tying the normative force of reasonableness to welfare maximisation (Posner, 2014), community values (Tilley, 2016), context-dependent normative justification (Gardner, 2015), Kantian freedom (Miller and Perry, 2012) or virtuous character traits (Feldman, 1998). Stern (2023) traces the history in the change in interpretation from descriptive to prescriptive. But the descriptive/prescriptive dichotomy does not exhaust the space of possibilities. Zipursky has recently proposed a "hybrid" view, according to which reasonableness "involves a kind of judgment that is both normative and descriptive." (2015, 2150) In an interesting series of vignette-based studies, Tobia (2018) has reported some evidence in favour of such an account. Grossmann et al. (2020) have shown that the folk concept of reasonableness is related to social norms, with which a series of studies by Jaeger (2020) is broadly consistent. Kneer (2022) has demonstrated that folk judgments regarding the reasonableness of decisions and actions are strongly sensitive to outcome valence, and that this is likely not a bias. Although this small number of studies effectively exhausts the empirical literature concerning the folk concept of reasonableness, it is evident that there is a preliminary convergence on a view that characterizes the expression 'reasonable,' and the phenomenon it denotes, as at least partially evaluative or prescriptive.

Given that the legal expression 'reasonable' is strongly tied to its ordinary language usage, that the instructions provided to juries are minimal, and that judges tend to refuse to elaborate further on its meaning, it is key to develop a better understanding of what, exactly, the folk concept of reasonableness is (Kneer, 2022; Tobia, 2018). This is our goal. To do so, we will first introduce some helpful distinctions with regards to evaluative concept classes from philosophy of language and moral philosophy. We then discuss a design to predict the concept class of a term based on sentiment metrics. Finally,

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

we present a corpus study that sheds light on the potential evaluative dimension of 'reasonable' and expressions closely associated to it. For this, we use a classifier trained on the sentiment dispersion in adjective conjunctions to predict whether 'reasonable' expresses a descriptive, a value-associated, or an inherently evaluative concept. Our results suggest that laypeople use 'reasonable' predominantly as a prescriptive term.

What kind of term is 'reasonable'?

In philosophy, evaluative terms are generally divided into different classes, such as thin and thick concepts (Eklund, 2011; Kirchin, 2019; Roberts, 2013; Tappolet, 2004; Väyrynen, 2013; for empirical studies see e.g., Baumgartner et al., 2022; Willemsen et al., 2021; Willemsen and Reuter, 2021). Thin terms like 'good' or 'bad' are all about pure evaluation without getting into the nitty-gritty of what exactly is being evaluated. For example, saying "John is good" evaluates John in a positive way, but does not specify why, or in what way, we evaluate him as good. Thick terms, on the other hand, provide extra descriptive information along with the evaluation. For example, the utterance "John is brave" tells us that John shows mental or moral strength to face danger, difficulty, or fear and evaluates him positively. It is commonly assumed that thick terms evaluate *by virtue* of the descriptive properties they denote. Accordingly, "John is brave" evaluates John positively *because* he shows mental or moral strength. The characteristic feature of thick and thin terms we focus on in this study is that their default semantic content includes an evaluative component.

Unlike inherently evaluative terms such as thin and thick concepts, descriptive terms like 'permanent' or 'yellow' do not inherently communicate an evaluation. However, that does not mean that these words cannot be used in an evaluative way in certain situations.[1] For example, a fan of red Ferraris might view a 'yellow Ferrari' as a negative thing. But, these impromptu evaluations are not seen as an inherent part of the term's standing meaning.

Recently, Reuter et al. (2022) have identified another class of concepts, the so-called value-associated concepts, which are neither fully descriptive nor inherently evaluative. Expressions denoting such concepts are *prima facie* descriptive, but they are often evaluatively charged because they tend to have common positive or negative associations (for psycholinguistic research, see e.g., Clore et al., 1987; Rensbergen et al., 2016; Võ et al., 2009). Take the utterance "It's rainy today." Most people probably have a negative association with rainy weather, like feeling bored or sad. In the context of beach vacations, for example, 'rainy' is likely to carry such a negative association. However, 'rainy' does not necessarily carry a negative evaluation. For a farmer who is currently experiencing a drought, a rainy day is a blessing. In either case, the evaluation does not pertain to the fact that water falls from the sky, but rather to the impact of rain in those specific circumstances.

The difference between value-associated and inherently evaluative terms lies in the distinction between pragmatic and semantic meaning. The evaluation communicated by value-associated expressions seems to be context-sensitive and thus potentially cancellable, akin to a conversational implicature. For inherently evaluative terms, on the other hand, there is hardly anyone who claims that the evaluation is just a particularised conversational implicature. Most say, like Roberts (2013), that the evaluation is in principle inseparable from the term's descriptive features, or that they are theoretically conceivable as two separate things, but have a very strong link (e.g., semantic entailment, presupposition, conventional implicature, etc.).

In this paper we are particularly curious about whether 'reasonable' conveys evaluative content and, if so, whether it is inherently evaluative or merely value-associated. We thus only focus on differences regarding the evaluativity of terms, leaving aside the descriptive component. Hence, we will only be examining whether 'reasonable' is purely descriptive, value-associated, or inherently evaluative (regardless of whether it is a thick or thin term). What is not yet entirely clear is how the descriptive, prescriptive, and hybrid accounts relate to these concept classes.

How do the concept classes just sketched map onto the differing views of 'reasonable'—descriptive, prescriptive, and hybrid – discussed in the introduction? For the descriptive account, the story is very straightforward—as the name suggests, it conceives 'reasonable' as a descriptive term devoid of evaluative features. According to the prescriptive account, 'reasonable' refers to an ideal and thus, arguably, evaluates by default. Intuitively, it seems quite plausible that 'reasonable' is a thick term rather than a thin term, as it has more descriptive content than thin terms like 'good' or 'great.' The tricky part, though, is figuring out the difference between the prescriptive and hybrid accounts. Tobia illustrates the latter as follows:

> The [criminal law's affirmative] defense [of duress] applies to an allegation of criminal conduct where the person "was coerced to [act] by the use of, or a threat to use, unlawful force... that a person of *reasonable firmness* in his situation would have been unable to resist." (MPC §2.09(1)) In applying this standard, it seems clear that both statistical and prescriptive considerations are crucial. We care about both the firmness most people *would* have in the relevant situation and what firmness someone *should* have in that situation. (Tobia, 2018, 308)

The question is whether the relation between *would* and *should* is indicative of a strong semantic relation or just an association. A strong version of the hybrid account would conceive of 'reasonable' as a thick term. A weak account, on the other hand, might consider it as a value-associated term. It seems plausible that advocates of hybrid theories take that

[1] For a discussion of what philosophers have called "evaluative variability," see, e.g., Blackburn (1992); Dancy (1995); Väyrynen (2011, 2013, 2021).

the descriptive and prescriptive dimensions of 'reasonable' to be independent of each other. This implies that in certain contexts, one of the dimensions can be cancelled out, which explains the flexible use of the term. For example, an action (such as paying taxes) may be deemed prescriptively required but may not be performed by many individuals. On the other hand, the statistical likelihood of an action does not necessarily imply that it adheres to a prescriptive ideal. Hence, the weak account is more adept at capturing the hybrid view. To summarize, we think that the descriptive view predicts that 'reasonable' is a descriptive term, the prescriptive view thinks it is an inherently evaluative term, and the hybrid view conceives of it as a value-associated term—or at least this would be the most plausible construal of the view.

Design

What is the best way to operationalize the aforementioned concept classes in the context of quantitative corpus studies? Perhaps the key challenge resides in operationalizing the distinction between merely value-associated terms and inherently evaluative terms. However, it seems plausible that evaluative terms co-occur frequently with terms of similar valence, whereas value-associated terms co-occur less frequently together with other evaluative terms because they do not evaluate by default. Hence, thick and thin terms should have consistently higher and/or stable co-occurring sentiment scores than value-associated and descriptive terms. We thus suggest conceiving of the discussed classes as clusters on an underlying continuum (viz. co-occurring sentiment), ranging from purely descriptive to inherently evaluative terms, with value-associated terms in between. In the following, we will present how this design has been used previous research.

Previous empirical studies on evaluative concepts has focused on the sentiment distribution of adjectives in coordinating conjunctions, e.g. "*cruel* and *manipulative* interrogation" (Baumgartner, 2022; Willemsen et al., 2021). Adjectives in these 'and'-conjunctions typically have a similar sentiment polarity and intensity (Elhadad and McKeown, 1990; Hatzivassiloglou and McKeown, 1997): positively evaluating adjectives are commonly used in conjunction with other positive adjectives, descriptive ones are paired with neutral ones, and negative with negative ones. Thus, Willemsen et al. (2021) argue, both adjectives in coordinating conjunctions are mutually informative with regards to evaluativeness. In other words, the sentiment distribution of conjoined adjectives for any term X is considered a good indicator of X's own evaluativeness. E.g., if 'reasonable' is used in conjunction with 'good,' 'laudable,' and 'intelligent,' this indicates that 'reasonable' carries a positive evaluation. We will use this design for our classification task.[2]

In a recent study, Baumgartner (2022) used this design to classify inherently evaluative terms (thick and thin terms) versus terms that do not evaluate by default (descriptive and value-associated terms). However, the classifiers for this task did not exceed $\approx 63\%$, which means that the classifier is only correctly predicting the label for about 63% of the sample. The author suspects that the main reason for this is that descriptive and value-associated terms operate in fundamentally different ways. In fact, the data suggests that value-associated concepts are much more similar to thick and thin concepts than to descriptive concepts. We thus expect improving the classification by treating descriptive, value-associated, and inherently evaluative terms as distinct classes. Accordingly, we have a classification task with three classes and will be using multi-class models.

[2]Note that this means that we do not look at legal phrases like "beyond reasonable doubt."

Data

Training/validation set

The classifiers were trained and validated on the corpus compiled by Baumgartner (2022), comprising 18,301 Reddit comments. Each comment contains a coordinating conjunction of two adjectives, e.g., "What a *cruel* and *sad* world!" One of the two adjectives is considered the *target adjective*—the adjective of primary interest; the other one is the *conjoined adjective*. The latter is secondary in the sense that it is only used to convey additional information about the former. The set of these target adjectives consists of a preselection of evaluative and non-evaluative terms featured in Reuter et al. (2022). The authors have annotated the concept class of each target adjective, distinguishing between five classes:

- **descriptive terms**: dry, large, loud, narrow, permanent, short, wooden, yellow.
- **value-associated terms**: quiet, rich, shiny, sunny, tall, broken, bloody, closed, empty, rainy.
- **thick moral terms**: compassionate, courageous, friendly, generous, honest, cruel, rude, selfish, reckless, vicious.
- **thick non-moral terms**: beautiful, delicious, funny, justified, wise, boring, disgusting, insane, stupid, ugly.
- **thin terms**: good, great, terrific, bad, terrible, awful.

The other adjective in the conjunction, the conjoined adjective, is freely variable. The corpus contains the adjectives' sentiment based on the sentiWords dictionary (Baccianella et al., 2010; Esuli and Sebastiani, 2006; Gatti et al., 2016). The dictionary codes a sentiment on a continuous scale from $-1 \leqslant x \leqslant 1$ (-1 = highly negative, 0 = neutral, 1 = highly positive). The data also includes the animacy state (animate vs inanimate) and the entity type (e.g., abstract, person, object, etc) of the subject/object of the predication, based on the xrenner-algorithm by Zeldes and Zhang (2016).

In this paper, we are only interested in the difference between descriptive, evaluative, and value-associated terms. Hence, we pool thin, thick moral and non-moral terms together to form the evaluative class. To ensure a balanced

sample, the training/validation set was reduced to a random subsample of 2,000 observations per class (total $n = 6,000$).

Prediction set

The prediction set is based on expressions which figure prominently in US jury instructions on the one hand and legal theory and philosophy on the other. The official jury instructions of the most populous US states for civil negligence cases, all of which define the latter in terms of reasonableness, standardly invoke a failure to behave like the "reasonably careful person" (California, Illinois, Florida, Pennsylvania) or "a person of ordinary prudence" (New York), as well as the lack of "ordinary care" (Texas, New York, Illinois). 'Ordinary' is rather descriptive (even though it can also be used in a demeaning tone), and frequently elucidated in terms of what is 'average' or 'normal' in the literature (see e.g., Tobia, 2018; Zipursky, 2015). 'Careful' seems more on the normative side of the fence. We have also included 'rational,' to which utilitarians in the Law & Economics tradition want to reduce the reasonable (e.g., Posner, 2014), as well as 'sensible' and 'responsible' which are frequently used as synonyms for 'reasonable' by philosophers (e.g., Lawlor, 2022). Lastly, we include the antonym of 'reasonable,' viz. 'unreasonable.'

For the prediction set, we collected 37,174 Reddit comments containing adjectives conjunction including the following terms:

- **prediction terms**: reasonable, careful, rational, sensible, responsible, ordinary, normal, prudent, average, unreasonable

The data was collected and annotated to match the training/validation set.[3] Only 'and'-conjunctions were considered. Previous research has shown that conjunctions with 'but,' 'or,' or 'yet' work quite differently (Elhadad and McKeown, 1990; Hatzivassiloglou and McKeown, 1997). Comments which include a negation of the adjectives (e.g. 'not,' 'hardly,' or 'barely') or any other adverbial modifier (e.g. 'very,' 'rather,' or 'mostly') were discarded.

Methods

Task

In a first step, we train and validate a classifier distinguishing between three classes: descriptive, value-associated, and evaluative terms. Thereafter, we use the best model to generate predictions for the term 'reasonable' and the terms often associated with it in jury instructions. Based on these predictions, we can determine whether 'reasonable' is a descriptive, value-associated, or evaluative term.

[3]The Reddit data was collected using the Pushshift API (Baumgartner et al., 2020) in R (v4.1.0). The dependency parsing was conducted using the `stanza` toolkit (v1.3.0) provided by the Stanford NLP Group (Qi et al., 2020) in Python (v3.7.11). Both the coreference resolution and the animacy detection are conducted with `xrenner` (v2.2.0.0) by Zeldes and Zhang (2016), based on the pretrained Electra model for GUM7, using Python (v3.7.11). For the sentiment annotation we used the `quanteda`-package (v3.0.0) in R (v4.1.0).

Models

The training data is split randomly into a train and validation set, based on an 80–20% ratio. We train and compare the following models: penalized multinomial regression (MNL), support vector machines with radial basis function kernel (rSVM), and random forest (RF).[4] For the rSVM, the data is additionally pre-processed during training (scaled and centred). The training includes 10-fold repeated cross-validation for all three models. The optimal tuning parameters are automatically chosen to maximize accuracy (tune length = 20).

Variables

The set of selected variables includes the sentiment of the conjoined adjective as well as its square product, the difference between the sentiment of the two adjectives and its square product, the polarity of the target adjective, the animacy state of the object of predication, a dummy coding whether the target adjective is mentioned first or second, and the timestamp of the comment.

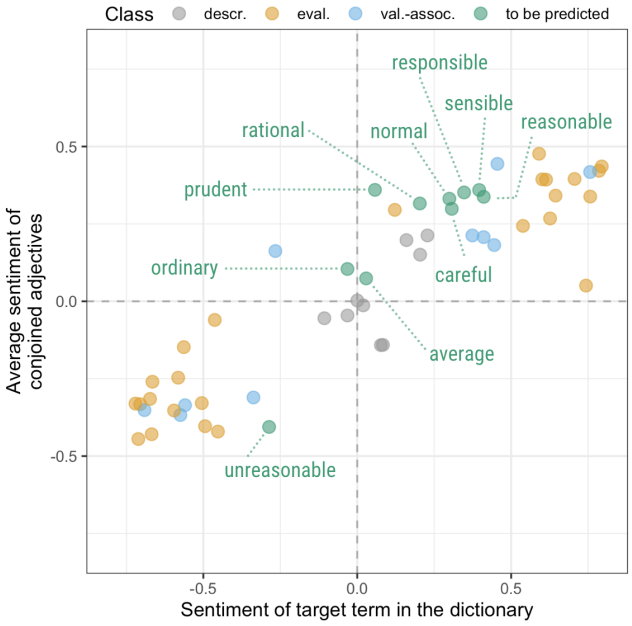

Figure 1: Sentiment dispersion in adjective conjunctions.

Figure 1 depicts the relation between the sentiment value of a target adjective and the average sentiment of its conjoined adjectives. As can be seen, 'reasonable' and its affiliated target adjectives cluster with value-associated terms in between evaluative and descriptive terms.

Results

Training and validation

Each model was first trained using 10-fold cross-validation to select the best tuning parameters (based on accuracy). Then, we generated predictions in the validation set to compare

[4]The models were built with `caret` (v6.0-90) in R (v4.1.0).

the models. Table 1 shows the model performances. The best model is the RF (mtry= 2) with an accuracy of 91.17% ($\kappa = 86.76$). The rSVM model (C= 760.228, $\sigma = 1.517$) has 82.08% accuracy ($\kappa = 73.12$). The optimal MNL (decay= 0.0215) has an accuracy of 46.08% ($\kappa = 19.34$). All models significantly exceed the no-information rate (34.58%), on 0.05-alpha level. The RF models performs significantly better than the other models, based on 95% confidence intervals.[5]

Table 1: Performance evaluation metrics [%].

	Accuracy (95% CI)	Kappa
Random forest	91.17 (89.42, 92.71)	86.76
rSVM	82.08 (79.79, 84.21)	73.12
Multinominal	46.08 (43.23, 48.95)	19.34

For the final RF model, we dropped the animacy state and the order dummy as predictors, based on the increase in the prediction error (MSE).

Table 2 contains the confusion matrix for the RF model. The misclassification rate is generally lower than 4.2% per cell. That said, it is highest for evaluative and value-associated concepts: 4.17% of value-associated terms are misclassified as evaluative concepts, and 1.58% of evaluative terms are mistaken as value-associated terms.

Table 2: Confusion matrix for the RF model [%].

		True Class		
		descr.	eval.	val.-assoc.
Pred.	descriptive	31.58	0.67	0.42
	eval.	1.58	29.58	4.17
	val.-assoc.	0.42	1.58	30.00

Given the high accuracy of the RF model (91.17%), we are confident that it accurately reflects the classes we are interested in.[6] The data in the prediction set was collected from the same platform (Reddit). Hence, we expect that this model can adequately classify the observations in the prediction set.

Predictions

The RF classifier predicts a roughly equal spread of concept classes: terms in the prediction set are 37.00% descriptive, 31.28% evaluative, and 31.72% value-associated. However, the group of selected adjectives is not as homogeneous as expected. Table 3 shows the proportions of the predicted classes for each adjective. 'Reasonable' is evaluative in 52.63% of cases, 30.19% value-associated, and 17.18% descriptive, which is very similar to what we find for 'sensible.' Interestingly, its antonym, 'unreasonable,' belongs to a different class, as it is mostly value-associated (78.95%). Lastly, 'average', 'ordinary', and 'prudent' are predominantly descriptive. Hence, it seems that 'reasonable' is used very differently from the other terms often associated with it in jury instructions.

Table 3: Shares of predicted class for each target adjective in the prediction set [%].

	eval.	val.-assoc.	descr.
sensible	53.48	29.67	16.85
reasonable	52.63	30.19	17.18
rational	38.99	02.50	58.51
responsible	37.07	50.05	12.88
careful	30.19	46.81	23.00
normal	26.35	46.61	27.04
unreasonable	18.70	78.95	02.35
prudent	18.39	07.21	74.40
ordinary	14.09	01.61	84.30
average	07.61	08.70	83.69

Figure 2 shows how close our terms of interest (circles) are to each other as well as to terms from the validation set (triangles). It depicts the centroids of the RF proximity measures after multi-dimensional scaling (MSD). The terms are color coded with the respective predicted class based on the mode. Note that the Figure only shows a random subsample of the prediction set ($n = 800$ per term). MSD allows to illustrate the similarity of high-dimensional data in a 2D space, in such a way that the relative distances in the higher-dimensional space are preserved in the lower-dimensional space.

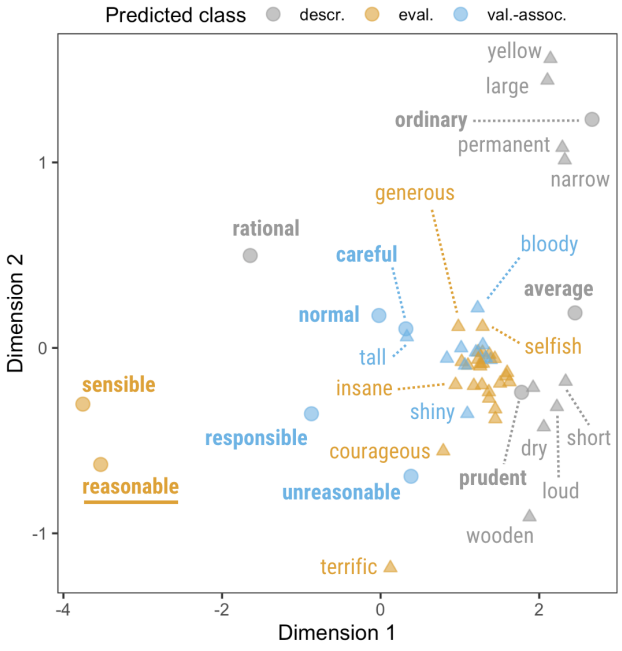

Figure 2: Multi-dimensional scaling (MDS) plot of the RF proximity matrix. Each data point represents the centroid of the respective term, color coded with its predicted class (mode). Triangles are terms from the validation set, whereas circles are from the prediction set.

[5]The data and scripts for this analysis are publicly available on the Open Science Framework repository at https://osf.io/tfasc/.

[6]While the classifier is not perfect as it does not have a $\approx 98\%$ prediction accuracy, it is still considered a very good performance for many applications, especially for complex problems.

As we can see, 'reasonable' is very similar to 'sensible,' but very dissimilar to 'average,' 'prudent,' 'ordinary.' What is interesting is that 'ordinary' is part of a cluster of descriptive terms including 'large,' 'narrow,' 'permanent,' and 'yellow.' 'Prudent,' too, is part of a descriptive cluster together with 'dry,' 'loud,' and 'short.'

In sum, we again find 'reasonable' to be used quite differently from the terms by means of which it tends to be explained in jury instructions, namely 'average', 'prudent', and 'ordinary.' The latter are mostly used as descriptive adjectives, whereas 'reasonable' is used much more evaluatively by laypeople.

Discussion

Our research reports an intriguing discovery: the expression 'reasonable' is *not* just a straightforward descriptive term. In fact, only 17.18% of uses in our sample fall into this category. Interestingly, other words that are commonly used alongside 'reasonable' in jury instructions, like 'average,' 'ordinary,' 'rational,' and 'prudent' are primarily descriptive. In terms of multidimensional proximity, 'reasonable' inhabits a very different part of the space. Considering that these terms are used somewhat interchangeably in jury instructions, our data suggests that jurors enter the courtroom with a different concept than the one intended or expected by legislators. However, it is important to keep in mind that we are not directly comparing the language of laypeople and experts and therefore cannot make direct inferences about possible differences between the two. And yet, judging from the fact that laypeople use 'reasonable' in a completely different way than other terms used to characterize 'reasonable' in the jury instructions, this suggests, at least indirectly, a certain discrepancy in language use. Furthermore, our results align with the findings by Willemsen et al. (2021) that laypeople tend to use certain terms in a more evaluative manner, compared to legal professionals. This disparity in understanding can have significant ramifications during trial, as jurors and legal professionals may not be on the same page. For a more comprehensive investigation of this discrepancy, further comparative studies are required.

Our results have multiple implications: First, our findings challenge the notion that 'reasonable' is purely descriptive, as advocated by Dressler (1994) and Zalesne (1996), on the one hand, and—more importantly—as one might infer from the jury instructions of US states on the other. Second, our results can help adjudicating between the prescriptive and hybrid accounts of the term. We cash out the difference between the two such that the prescriptive account predicts a primarily evaluative use, while the hybrid account predicts 'reasonable' to be value-associated. Based on this operationalization, our findings favor the prescriptive account, as 52.63% of the uses in our sample were evaluative, compared to only 30.19% being value-associated. Third, our data suggests a strong relationship between 'reasonable' and 'sensible,' which might be an indication for synonymy, as Lawlor (2022) has proposed. Fourth, another noteworthy finding is that 'rational' is much more descriptive than 'reasonable.' Perhaps this can be explained by different connotations. Grossmann et al. (2020) suggest that both 'rational' and 'reasonable' are ordinarily associated with being sensible, intelligent, and logical, but they nevertheless carry very distinct connotations. While 'rational' generally refers to self-interest and agency, 'reasonable' is applied in cases where people are socially-minded and caring. These differences in connotations could lead to different adjectival co-occurrences, which ultimately affects their sentiment dispersion. Lastly, we find that 'reasonable' and its antonym 'unreasonable' belong to different classes. This might be related to a general asymmetry in positive and negative adjectives (e.g., Baumgartner et al., 2022; Willemsen and Reuter, 2021), or connected to different connotations as well. Thus, research that takes connotations into account is needed.

One drawback of our study is that it is limited to the evaluative dimension of concepts, and ignores the descriptive dimension altogether. Further research is needed to understand the relationship between the two for terms like 'reasonable.' Including descriptive content would allow us to distinguish thin from thick concepts, which would make for a more fine-grained analysis. For this purpose, it might be advisable to make a switch to more complex embedding models, like word2vec (e.g., Jatnika et al., 2019; Lilleberg et al., 2015; Toshevska et al., 2020). Besides that, more theoretical work is necessary with regards to the two possible operationalizations of the hybrid account, i.e. whether 'reasonable' is thick or value-associated.

In conclusion, our study represents a first step in examining various accounts of legal terms such as 'reasonable' by mapping them onto different concept classes, and rigorously examining them empirically. Our study thus not only presents key insights regarding the semantics of 'reasonable,' but demonstrates how experimental philosophy of language can contribute to legal theory and practice.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J., and Io, P. (2020). The Pushshift Reddit Dataset. *ArXiv Preprint*, page 2001.08435v1.

Baumgartner, L. (2022). Why are reckless socks not (more of) a thing? Towards an empirical classification of evaluative concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Baumgartner, L., Willemsen, P., and Reuter, K. (2022). The polarity effect of evaluative language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44:729–735.

Blackburn, S. (1992). Through thick and thin. *Proceedings of the Aristotelian Society, supplementary*, 66:284–299.

Clore, G. L., Ortony, A., and Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53:751–766.

Dancy, J. (1995). In defense of thick concepts. *Midwest Studies in Philosophy*, 20:263–279.

Dressler, J. (1994). When heterosexual men kill homosexual men: Reflections on provocation law, sexual advances, and the reasonable man standard. *Journal of Criminal Law and Criminology*, 85:726.

Eklund, M. (2011). What are Thick Concepts? *Canadian Journal of Philosophy*, 41(1):25–49.

Elhadad, M. and McKeown, K. R. (1990). Generating Connectives. In *Proceedings of the 13th conference on Computational linguistics*, pages 97–101.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422.

Feldman, H. L. (1998). Prudence, benevolence, and negligence: Virtue ethics and tort law. *Chicago-Kent Law Review*, 74:1431.

Gardner, J. (2015). The many faces of the reasonable person. *Law Quarterly Review*, 131:563–584.

Gatti, L., Guerini, M., and Turchi, M. (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Grossmann, I., Eibach, R. P., Koyama, J., and Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality versus reasonableness. *Science Advances*, 6.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

Jaeger, C. B. (2020). The empirical reasonable person. *Alabama Law Review*, 72:887.

Jatnika, D., Bijaksana, M. A., and Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157:160–167.

Keating, G. C. (2022). *Reasonableness and Risk: Right and Responsibility in the Law of Torts*. Oxford University Press.

Kirchin, S. (2019). Thick and Thin Concepts. *International Encyclopedia of Ethics*, pages 1–10.

Kneer, M. (2022). Reasonableness on the clapham omnibus: Exploring the outcome-sensitive folk concept of reasonable. In *Judicial Decision-Making*, pages 25–48. Springer, Cham.

Lawlor, K. (2022). A genealogy of reasonableness. *Mind*, forthcoming.

Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015*, pages 136–140.

Miller, A. D. and Perry, R. (2012). The reasonable person. *New York University Law Review*, 87:323.

Posner, R. A. (2014). *Economic analysis of law*. Wolters Kluwer.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics System Demonstrations*, pages 101–108.

Rensbergen, B. V., Deyne, S. D., and Storms, G. (2016). Estimating affective word covariates using word association data. *Behavior Research Methods*, 48:1644–1652.

Reuter, K., Baumgartner, L., and Willemsen, P. (2022). Tracing Thick and Thin Concepts Through Corpora. PhilSci Archive preprint, http://philsci-archive.pitt.edu/id/eprint/20584.

Roberts, D. (2013). Thick Concepts. *Philosophy Compass*, 8(8):677–688.

Stern, S. (2023). From clapham to salina: Locating the reasonable man. *Law & Literature*, pages 1–27.

Tappolet, C. (2004). Through thick and thin: good and its determinates. *Dialectica*, 58(2):207–221.

Tilley, C. C. (2016). Tort law inside out. *Yale Law Journal*, 126:1320.

Tobia, K. P. (2018). How people judge what is reasonable. *Alabama Law Review*, 70:293.

Toshevska, M., Stojanovska, F., and Kalajdjieski, J. (2020). Comparative analysis of word embeddings for capturing word similarities. In *Proceedings of the 6th International Conference on Natural Language Processing (NATP 2020)*, pages 9–24. Academy and Industry Research Collaboration Center (AIRCC).

Unikel, R. (1992). Reasonable doubts: A critique of the reasonable woman standard in american jurisprudence. *Northwestern University Law Review*, 87:326.

Väyrynen, P. (2011). Thick concepts and variability. *Philosopher's Imprint*, 11:1–17.

Väyrynen, P. (2013). *The Lewd, the Rude and the Nasty: A Study of Thick Concepts in Ethics*. Oxford University Press.

Väyrynen, P. (2021). Thick ethical concepts. The Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/thick-ethical-concepts/.

Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., and Jacobs, A. M. (2009). The berlin affective word list reloaded (bawl-r). *Behavior Research Methods*, 41:534–538.

Willemsen, P., Baumgartner, L., Frohofer, S., and Reuter, K. (2021). Examining evaluativity in legal discourse : A comparative corpus-linguistic study of thick concepts. PsyArXiv Preprints, https://psyarxiv.com/yxsp9/.

Willemsen, P. and Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*, 10:135–146.

Zalesne, D. (1996). Intersection of socioeconomic class and gender in hostile housing environment claims under title viii: Who is the reasonable person, the. *Boston College Law Review*, 38:861.

Zeldes, A. and Zhang, S. (2016). When Annotation Schemes Change Rules Help: A Configurable Approach to Coreference Resolution beyond OntoNotes. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes*, pages 92–101.

Zipursky, B. C. (2015). Reasonableness in and out of negligence law. *University of Pennsylvania Law Review*, 163:2131–2170.