Forthcoming in Tobia, K. P. (Ed.). The Cambridge Handbook of Experimental Jurisprudence. Cambridge University Press.

The meaning of 'reasonable': Evidence from a corpus-linguistic study

Lucien Baumgartner[1] **and Markus Kneer**[2]

[1,2]Department of Philosophy, University of Zurich

August 25, 2023

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 based on supervised machine learning models, we investigate whether laypeople use the expression 'reasonable' mainly as a descriptive, an evaluative, or merely a value-associated term. We find that 'reasonable' is predicted to be an evaluative term in the majority of cases. This supports prescriptive accounts, and challenges descriptive and hybrid accounts of the term—at least given the way we operationalize the latter. Interestingly, other expressions 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 'reasonable' 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

1 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 center 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 "bedeviled and divided courts and scholars for centuries" (Miller and Perry, 2012, 323), is still hotly debated. Some have defended a descriptive account (Dressler, 1994; Zalesne,

1996). A greater number advocate a prescriptive account, tying the normative force of reasonableness to welfare maximization (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 favor 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. In recent corpus study, Nyarko and Sanga (2022) discovered that judges employ the term 'reasonable' in a hybrid manner, encompassing a broad spectrum of activities that range from 'ideal' to potentially less-than-ideal or 'typical' conduct, similar to terms like 'rational,' 'justifiable,' and 'realistic.' In contrast, laypeople tend to adopt a more prescriptive usage of the term, closer to 'valid,' 'prudent,' and 'sensible.' 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). To do so, in Section 2, we first motivate the need for empirical inquiry in somewhat more depth and respond to recent critical voices concerning experimental jurisprudence in passing. In Section 3, we then introduce some helpful distinctions with regards to evaluative concept classes from philosophy of language and moral philosophy. Moving forward, Section 4 presents our study design, which enables us to explore how sentiment metrics can be used to predict the class of a term. Finally, we provide the data (Section 5), methods (Section 6), and results (Section 7) of our corpus study, which sheds light on the potential evaluative dimension of 'reasonable' and expressions closely associated with 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 findings indicate that laypeople use 'reasonable' predominantly as a prescriptive term. Lastly, in Section 8, we discuss the implications for the law, address the limitations of our study, and suggest avenues for future research.

2 The Need for Empirical Inquiry

We hold that the tools of experimental jurisprudence are ideally suited to help clarify the central question at hand: whether the expression 'reasonable' denotes a descriptive, evaluative or hybrid concept, and what this means for legal practice. Experimental jurisprudence is a young discipline which elucidates key legal concepts and assumptions by aid of empirical means.

Over the last few years, experimental jurisprudence has covered a broad range of topics (for reviews, see Knobe and Shapiro, 2021; Prochownik, 2021; Tobia, 2022). One particularly active area of research regards *mens rea* (the subjective element of a crime, i.e. the "guilty mind") broadly conceived, under which the reasonable person standard falls as well. Recent work has focused on intentionality and epistemic states (Frisch et al., 2021; Kneer and Bourgeois-Gironde, 2017, 2018; Kneer et al., 2023; Kobick and Knobe, 2009; Macleod, 2015; Mott and Heiphetz, 2023; Nadelhoffer, 2006; Tobia, 2020), recklessness and negligence (Kneer and Machery, 2019; Kneer and Skoczeń, 2023; Margoni and Brown, 2023; Murray et al., 2023; Nobes and Martin, 2022), willful ignorance (Kirfel and Hannikainen, 2023; Kirfel and Phillips, 2023), the interaction between the attribution of mental states such as foresight and causation (Güver and Kneer, 2023; Knobe and Shapiro, 2021;

Lagnado and Channon, 2008; Sytsma, 2019) and—as mentioned in the introduction—the notion of reasonableness, which is of key relevance for the less inculpating types of *mens rea*.

Jiménez (2023, this volume; see also Jiménez, 2021) has recently raised doubts about experimental jurisprudence. With special jurisprudential concepts such as *intention*, *cause*, *consent* and the like, Jiménez observes, there "might be three notions [...] at play: that of the theorist, that of the lawyer, and that of the ordinary person" (2023, 5). Legal practitioners have "a *know-how* that allows them to engage in the practice of legal reasoning," which "entitles them to make claims about the meaning of those concepts independently of both theoretical and everyday understandings" (2023, 6). When it comes to the law, then, "those in the driving seat of legal concepts are legal officials and participants, not laypeople" (2023, 6). Consequently, Jiménez argues, "[e]xperimental jurisprudence can help us determine the content of legal concepts when it studies legal experts," though "findings about lay cognition do not support direct inferences or conclusions about legal concepts" (2023, 7).

The argument is fine as far as it goes, though it mischaracterizes experimental jurisprudence, which is not quite the blunt tool Jiménez makes it out to be. First, from the get-go of experimental jurisprudence, studies with legal experts were a staple of the discipline. For instance, Spamann and Klöhn (2016) explore character bias and recourse to precedent in US federal judges. Scholars who are conducting work in and relevant to Civil Law jurisdictions, where trials are decided by professional judges, tend to sample *precisely* the audience Jiménez argues of relevance: legal professionals. Kneer and Bourgeois-Gironde (2017) investigate French judges' concept of *intention* and report it as being sensitive to both the Knobe effect and the severity bias. The results for French lawyers are very similar (Kneer and Bourgeois-Gironde, 2018), and Tobia (2023) also reports close-to-large (r=.49) correlations between perceived blame and intentionality attributions among US judges. For German legal experts, by contrast, Prochownik et al. (2020, 2023, this volume) fail to find a severity effect on intentionality attributions. Moa Lidén and her collaborators work almost exclusively with legal experts (principally judges from Sweden) and have found them to be as susceptible to a host of biases as laypeople (see, e.g., Lidén, 2023; Lidén et al., 2018, 2019). Noel Struchiner, Guilherme Almeida, Ivar Hannikainen and colleagues, too, have conducted a plethora of studies with legal experts from Brazil on several topics (see, e.g., Almeida et al., 2023; Struchiner et al., 2020a, 2020b). Piotr Bystranowski, Bartosz Janik, Maciej Próchnicki, Izabela Skoczeń and colleagues, have published interesting data with Polish legal experts (Bystranowski et al., 2022; Skoczeń and Smywiński-Pohl, 2022). The research of the cross-cultural experimental jurisprudence initiative (Hannikainen et al., 2018) routinely includes not just one, but four or five different expert samples in their exploration of central legal concepts and intuitions (see, e.g., Hannikainen et al., 2022; Kneer et al., 2023). In short, where data from legal professionals *matters*, it has frequently been collected.

Second, whereas in Civil law jurisdictions such as France, Sweden, Brazil, Poland or Germany studies with judges and lawyers are key because trials are decided by legal experts, in Common Law jurisdictions such as the UK and the US, lay concepts of special jurisprudence are of more importance than Jiménez is perhaps prepared to acknowledge. *Mens rea*, the subjective element in a criminal trial, for instance, is standardly adjudicated by lay jurors. Take the example of *intention*: It is not only unlikely that lay jurors leave their ordinary concept of *intention* at the door when entering the courtroom, but they are frequently instructed by the judge to recur to precisely that ordinary concept. The English courts, for instance, have made this very explicit, stating that "the legal meaning of the word 'intention' is the ordinary meaning of the word" (Herring, 2012, 135). In *R v Moloney* [1985], Lord Bridge argued:

> The golden rule should be that, when directing a jury on the mental element necessary in a crime of specific intent, *the judge should avoid any elaboration or paraphrase of what is meant by intent, and leave it to the jury's good sense to decide whether the accused acted with the necessary intent*, unless the judge is convinced that, on the facts and having regard to the way the case has been presented to the jury in evidence and argument, some further explanation or elaboration is strictly necessary to avoid misunderstanding. (*R v Moloney* [1985] AC 905, 926, italics added).

Courts routinely refuse to provide juries with clarifications of *intention* and other *mens rea* concepts and are only prepared to do so in exceptional cases (for discussion, see, e.g., Güver and Kneer, 2023; Herring, 2012).[1] One explanation for this is that the law, which, after all, aspires to be *public*, frequently assumes *correspondence* between legal and lay expressions and the concepts they denote (Kneer, 2022; Tobia, 2021b, 2018). Where such a *Correspondence Assumption* is in place and, in particular in Common Law countries, where, e.g., criminal trials are decided by lay juries, it stands to reason to investigate lay concepts which are simply supposed to be the concepts of relevance. In a nutshell, then, a general critique of experimental jurisprudence as put forth by Jiménez risks missing the mark. It does not adequately characterize the discipline and its many studies conducted with legal practitioners on the one hand, and it underappreciates the fact that lay concepts frequently—though by no means universally—play *the* central role in legal practice.

That said, though, it is evident that the law, across the Civil/Common Law divide, *does* employ a plethora of technical concepts, for which correspondence with lay concepts (if they even exist) is *not* assumed. For our purposes, the question thus arises whether the legal expression 'reasonable' is supposed to be understood in its ordinary language sense, or whether it denotes a *technical* legal concept distinct from the lay concept. Interestingly, here the response is even clearer than with most other legal concepts. Take Gardner (2015), for instance, who distinguishes "questions of law" from "questions of fact." The latter include, e.g., "medical questions, mathematical questions, *questions of ordinary linguistic usage*, etc." (2015, 7; italics added) and, he highlights, "the question of reasonableness"—and thus naturally also the concept of *reasonableness*—"belongs, for the most part, to the same list of non-legal questions" (2015, 7). What is more, the connection between 'ordinariness' and 'reasonableness' that can be frequently found in the jury instructions, Gardner suggests, should be interpreted as a matter of reassuring the jury "that the standard of reasonableness is to be set, *not by lawyers, but by 'ordinary' folk like themselves*" (2015, 7; italics added). But if this is so, and given how little the law and jury instructions have to say about the meaning of the expression 'reasonable,' it seems relatively obvious that the folk, in determining *what* is reasonable, are expected to employ *their* concept of *reasonableness*, not a technical variation thereof, whose meaning is not even properly explained to them. The concept, and corresponding standard of reasonableness, is, as Giestfeld (2020, 338) observes, "essentially defined by the lay understanding of jurors" (for further discussion, see also Jaeger, 2020, 2023, this volume, and Nyarko and Sanga, 2022). Tobia (2021b, 1) writes that "[l]aypeople's commonsense understandings, or 'ordinary concepts,' are at the root of many important legal concepts"—including, e.g., *intent*, *knowledge*, and *reasonableness*. We would go beyond this cautious formulation and argue—following *inter alia* Gardner—that the reasonable person standard "exists to allow the law to pass the buck, to help itself *pro tempore* to standards of justification that are not themselves set by the law" (2015, 36). The legal concept of *reasonableness* is hence by and large *coextensive* with the lay concept—not just the 'root' of it.

In sum, we have suggested that there are grounds for (considerable) correspondence between the lay concept of *reasonableness* and its legal equivalent. To shed further light on central questions regarding the legal concept—e.g. whether it is a normative or descriptive concept, or whether it is outcome-sensitive etc.—we must explore what, exactly, the lay concept of *reasonableness* consists in. And as discussed in the opening sections, although it is helpful to do so, e.g., by vignette studies in the vein of Grossmann et al., Jaeger, Kneer, or Tobia, there is a vast array of evidence at our disposal, which has received little attention to date: namely, linguistic usage, which can reveal what kind of concept the expression 'reasonable' denotes. For example, Nyarko and Sanga (2022) presented a novel statistical test for examining disparities in word usage across various corpora. In one of their application studies, the authors successfully identified notable disparities in the utilization of the term 'reasonable' between legal practitioners and laypersons, thereby challenging the empirical reliability of the correspondence assumption. Corpus studies of this sort have recently begun to receive some attention in legal theory (Lee and Mouritsen, 2018, 2021; Mouritsen, 2010; for a roadmap through the debate, see Tobia, 2021a). Our corpus study wants to add to this literature: By aid of corpus linguistic methods, we aim to clarify a notion central to

[1] Regarding *intention* in particular, see R v Allen [2005] Crim LR 698; R v Phillips [2007] EWCA Crim 1042.

the law—'reasonableness'. In doing so, we want to demonstrate the use-value of experimental jurisprudence broadly conceived and hope to convince even critics such as Jiménez who eye it skeptically.

3 What kind of term is 'reasonable'?

In philosophy, evaluative terms are generally divided into different classes, such as thin and thick terms (Eklund, 2011; Kirchin, 2019; Roberts, 2013; Tappolet, 2004; Väyrynen, 2013; for empirical studies see, e.g., Baumgartner et al., 2022; Willemsen and Reuter, 2021; for thick terms in law, see, e.g., Enoch and Toh, 2013; Willemsen et al., 2023). 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, an utterance of "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 like thin and thick terms, 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.[2] For example, a fan of red Ferraris might view the characterization "yellow Ferrari" as a negative thing. But impromptu evaluations of this sort are not seen as an inherent part of the term's standing meaning.

Recently, Reuter et al. (2022) have identified another class of terms, so-called value-associated terms, 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.[3] For inherently evaluative terms, on the other hand, only very few scholars argue that the evaluation conveyed by such terms is a conversational implicature (see, e.g., Väyrynen, 2012). Most scholars, like Roberts (2013), hold that if it is possible to conceive of the descriptive and evaluative features separately, they are at least strongly intertwined (e.g. through semantic entailment, presupposition, or conventional implicature).

In this paper we explore whether 'reasonable' is a purely descriptive, value-associated, or inherently evaluative expression (regardless of whether it is a thick or thin term). To begin with, let us examine in what ways the different accounts of 'reasonable'—descriptive, prescriptive, and hybrid—discussed in the introduction map onto the concept classes just sketched. For the descriptive account, the story is very straightforward: As the name suggests, it conceives of 'reasonable' as a descriptive term devoid of evaluative features. According to the prescriptive account, by contrast, 'reasonable' refers to an ideal and thus,

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

[3] Cancellability refers to the ability of an implicature or meaning conveyed in a statement to be revoked or negated in a specific context, while conversational implicatures are inferences derived from the context and not explicitly stated in the utterance itself.

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*' Tobia points to 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 a value-associated term. It seems plausible that advocates of hybrid theories take 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, it seems that the hybrid view of 'reasonable' might be better captured by the weak account, which casts it as a value-associated term.

Naturally, it is up to advocates of hybrid accounts to further clarify what, exactly, they mean by 'hybrid.' We take it that no party to the debate considers *reasonable* a thin evaluative concept: Such concepts are very rare, and all accounts of reasonableness tie it to some particular feature (welfare maximization, community values, normative justification, virtuous character traits etc., see section 1) that goes beyond the pure, simple, and unqualified good. Hence, those who cast *reasonable* as an evaluative concept presumably consider it a thick concept (i.e. an evaluative concept with some descriptive dimension). But if 'hybrid' accounts are not to simply collapse into thick concept accounts, the only option available for such views is to understand 'reasonable' as a value-associated expression. To summarize, we think that the descriptive view predicts that 'reasonable' is a descriptive term, the prescriptive view considers it a thick 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.

4 Design

What is the best way to operationalize the aforementioned concept classes in the context of quantitative corpus studies? It seems plausible that evaluative terms co-occur frequently with terms of similar valence, whereas value-associated terms co-occur less frequently with other evaluative terms because they do not evaluate by default. Hence, thick and thin terms (e.g. 'cruel,' 'good') should have consistently higher and/or stable co-occurring sentiment scores than value-associated (e.g. 'sunny') and descriptive terms (e.g. 'wooden'). 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 in extant research.

Previous empirical studies on evaluative concepts have focused on the sentiment distribution of adjectives in coordinating conjunctions (e.g. Baumgartner, 2022; Willemsen et al., 2023). Coordinating conjunctions include expressions like "What a *cruel* and *manipulative* interrogation!" or "He is a *reasonable* and *careful* driver." 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. (2023) 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 t is considered a good indicator of t's own evaluativeness. For instance, if 'reasonable' were frequently used in conjunction with 'good,' 'laudable,' and 'intelligent,' this would indicate that 'reasonable' carries a positive evaluation. We will use this design for our classification task.[4]

In this study, we adopt a design similar to that employed by Baumgartner (2022). The author utilized sentiment dispersion in coordinating conjunction to classify two types of terms: inherently evaluative terms (thick and thin terms) and terms that do not inherently convey an evaluation (descriptive and value-associated terms). Logistic regression models were utilized to achieve this binary classification task. However, the performance of the classifiers for this task was limited, achieving an accuracy of only $\approx 63\%$. This indicates that the classifier correctly predicts the label for about 63% of the sample. Baumgartner suspects that the main reason for this is that descriptive and value-associated terms operate in fundamentally different ways. Interestingly, the data suggests that value-associated concepts are much more similar to thick and thin concepts than to descriptive concepts. We thus expect that the classification can be improved by treating descriptive, value-associated, and inherently evaluative (thick and thin) terms as distinct classes. Accordingly, the classification task expands to three classes, necessitating the use of multi-class models.

5 Data

5.1 Training/validation set

The classifiers were trained and validated on corpus data compiled by Baumgartner (2022), comprising a total of 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 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 pre-selection 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:[5]

- **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

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

[5] The categorization of adjectives into descriptive, evaluative, and value-associated terms in Reuter et al. (2022) was based on claims derived from the literature on thick concepts, the authors' intuitions, and controlled for by the terms' sentiment value in the SentiWords dictionary. This explains why, e.g., 'loud' (sentiment: -0.11) is a descriptive term, whereas 'quiet' (sentiment: 0.41) is value-associated.

to form the evaluative class. In other words, we do not distinguish between these subclasses but rather treat them indiscriminately as evaluative terms. To ensure a balanced sample, the training/validation set was reduced to a random subsample of 2,000 observations per class (total $n = 6,000$).

5.2 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), or by aid of a lack of "ordinary care" (Texas, New York, Illinois). 'Ordinary' is a prima facie descriptive expression (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; note that Bear and Knobe, 2017 have argued that even 'normal' carries a partly evaluative meaning though). 'Careful' seems more on the normative side of the fence. We have also included 'rational,' to which utilitarians in the Law & Economics tradition such as Posner (2014) want to reduce the reasonable, 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,' as it is commonly encountered in court proceedings, and we anticipate that the antonym will exhibit similar behavior.

For the prediction set, we collected 37,174 Reddit comments containing conjoined adjectives 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.[6] 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 excluded from the analysis. These modifiers and intensifiers affect the expressed sentiment in ways that are beyond the scope of our study.

6 Methods

6.1 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 in ordinary English.

6.2 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).[7] For the rSVM, the data is additionally pre-processed during training (scaled and centered). 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).

[6] 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).

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

6.3 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.

7 Results

7.1 Sentiment distribution

Let us have a brief look at the overall sentiment dispersion, before delving into the results of the classification task. Figure 1 depicts the relation between the sentiment value of a target adjective and the average sentiment of its conjoined adjectives.[8] As can be seen, 'reasonable' and its affiliated target adjectives (green) have a tendency to cluster with value-associated terms (blue) in between evaluative (yellow) and descriptive terms (grey). In general, the behavior of our target expressions is less homogeneous than anticipated. Notably, the terms 'average' and 'ordinary' are much closer to the neutral midpoint than the rest of the target terms. Surprisingly, 'normal' is quite far removed from 'ordinary' and 'average,' indicating a more value-laden use. This is an important finding in its own right, confirming the results of Bear and Knobe (2017), who argue that *normal* is a dual-character concept with both descriptive and prescriptive features.

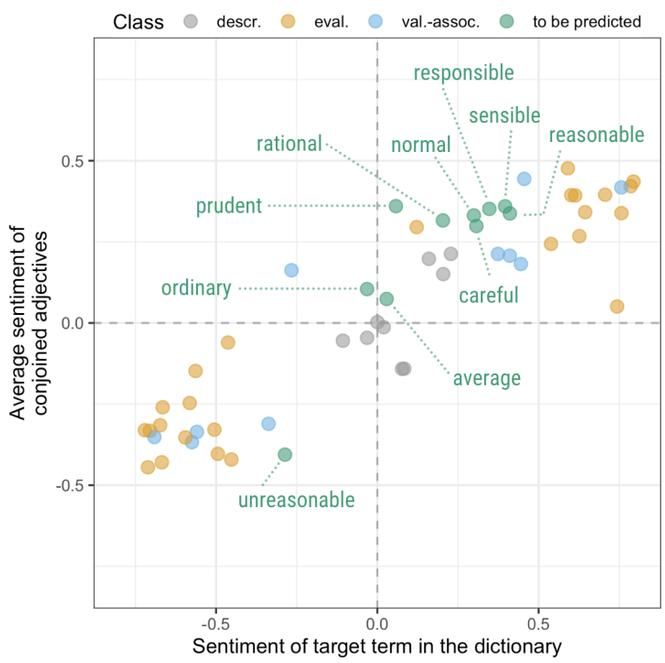

Figure 1: Average conjoined sentiment in relation to the sentiment of the target term.

Figure 2 shows the conjoined sentiment dispersion for the target terms in the prediction set, which delivers a few additional insights:

1. The term 'average' is interesting because it is used in both positive and negative contexts. Depending on the situation, it can convey a sense of mediocrity or being typical, but it can also imply fairness or adequacy.

2. 'Ordinary' stands out because it shows a significant peak around the middle of the sentiment spectrum. This suggests that it is often used to describe things that are considered neither particularly positive nor negative.

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

3. On the other hand, terms like ‘normal,’ ‘prudent,’ and ‘unreasonable’ exhibit a bimodal distribution, which means that they are used in two distinct ways (exhibiting moderate and strong sentiment). The bimodal distribution for ‘normal’ and ‘prudent’ is limited to the positive side of the sentiment spectrum, whereas for ‘unreasonable’ it occurs on the negative side.

4. In contrast, terms like ‘careful,’ ‘rational,’ ‘reasonable,’ and ‘sensible’ exhibit a unimodal right skew. This means that they are primarily used in a positive sense, and—given the location of the peaks—tend towards the neutral midpoint rather than the extremely positive endpoint of the scale.

Based on these results, we can expect significant differences between the target terms in our classification task.

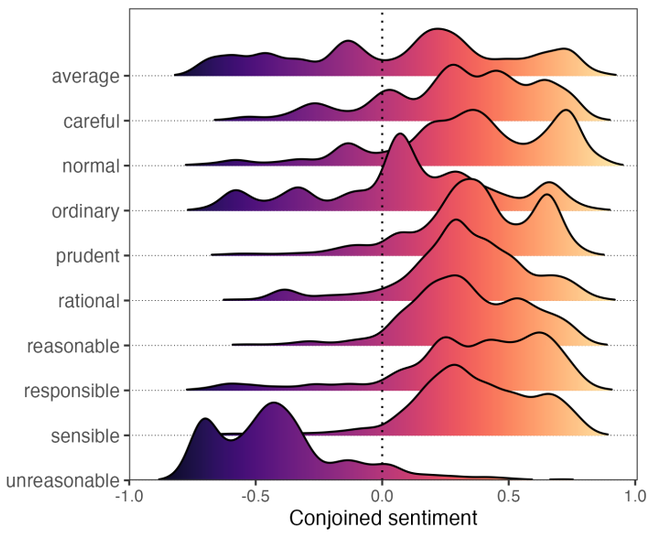

Figure 2: Conjoined sentiment dispersion for target terms in the prediction set.

7.2 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 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 model performs significantly better than the other models, based on 95% confidence intervals.

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		
		descriptive	evaluative	value-assoc.
Prediction	descriptive	31.58	0.67	0.42
	evaluative	1.58	29.58	4.17
	value-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.[9] Given that the data in the prediction set was collected from the same platform (Reddit), we anticipate that our model will effectively classify the observations in this set.

7.3 Predictions

The RF classifier predicts a roughly equal spread of 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, 'careful' appears to behave very differently, even though it is the go-to expression to explicate 'reasonable' in jury instructions. In contrast to 'reasonable,' 'careful' is primarily value-associated (46.81%). The antonym of 'reasonable,' i.e. 'unreasonable,' is even more clearly 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 [%].

	evaluative	value-assoc.	descriptive
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 3 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 multidimensional scaling (MDS). 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). MDS 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.

As we can see, 'reasonable' is very similar to 'sensible'—a term philosophers consider synonymous with 'reasonable.' As we can also see, 'reasonable' is very dissimilar to 'average,' 'prudent,' and 'ordinary'—expressions descriptive accounts of reasonableness tend to emphasize, which also play a prominent role in jury instructions. What is interesting is that 'ordinary' (cf. the jury instructions of Texas, New York, and Illinois) is part of a cluster of descriptive terms including 'large,' 'narrow,' 'permanent,' and 'yellow.' 'Prudent' (cf.

[9] 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.

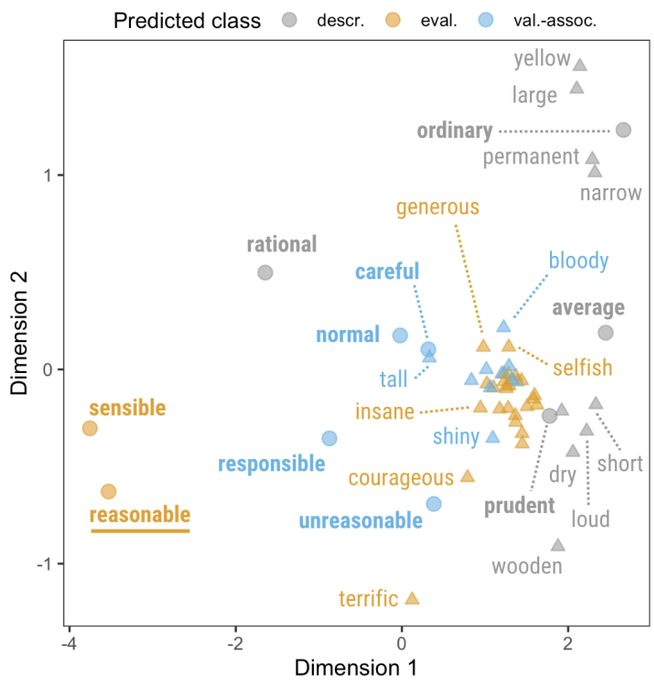

Figure 3: MDS plot of the RF proximity matrix. Each data point represents the centroid of the respective term, which is color-coded with its predicted class (mode). Triangles are terms from the validation set, whereas circles are from the prediction set.

the jury instructions of New York), too, is part of a descriptive cluster together with 'dry,' 'loud,' and 'short.'

In sum, we 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 on Reddit.

8 Discussion

Our research reports an intriguing discovery: the expression 'reasonable' is most often *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 to elucidate '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. (2023) that laypeople tend to use certain terms in a more evaluative manner, compared to legal professionals—at least if the jury instructions are used as a proxy thereof. 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 two implications: First, they challenge the notion that 'reasonable' is purely descriptive, as advocated e.g. 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 in 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 primarily 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. Interestingly, 'careful,' which is also frequently used in jury instructions to elucidate what 'reasonable' means, fits the hybrid account (operationalized as a value-associated account). However, it is problematic to explicate an evaluative expression in terms of a value-associate one.

The corpus study offers additional insights on expressions relevant to the debate on reasonableness. Our data suggests a strong relationship between 'reasonable' and 'sensible,' which might be an indication of synonymy, as Lawlor (2022) has proposed. 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 affect their sentiment dispersion. Our findings are also consistent with recent interesting results by Jaeger (2020), which suggest that the lay understanding of 'reasonable' differs considerably from economic rationality proposed by scholars of the Law & Economics tradition. 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 urgently needed.

One drawback of our study is that it is predominantly concerned with the evaluative dimension of concepts, without delving deeper into the descriptive dimension. Including descriptive features in the analysis of terms like 'reasonable' would give us a more comprehensive semantic representation. For instance, it is conceivable that the descriptive content of 'average' aligns with that of 'reasonable,' despite their differences in evaluation. It would also enable us to model the aforementioned distinct connotations of the expressions 'reasonable' and 'rational' suggested by Grossmann et al. (2020). For these purposes, it might thus be advisable to make a switch to more complex semantic representations, like word2vec (e.g. Jatnika et al., 2019; Lilleberg et al., 2015; Toshevska et al., 2020). Furthermore, a comparative analysis of the language used in the Reddit corpus with other corpora, such as newspaper articles or legal transcripts, could provide valuable insights into the variations in language use across different contexts. Nyarko and Sanga (2022) present a method to test for differences in word embeddings across different corpora. In one of their application studies, the authors find differences in how the term 'reasonable' is used in an ordinary corpus and a legal corpus. Laypeople's words that are closest to 'reasonable' ('valid,' 'prudent,' 'sensible') reflect an emphasis on 'ideal' conduct, while judges' closest words ('rational,' 'justifiable,' 'realistic') encompass both 'ideal' and potentially less-than-ideal or 'typical' conduct. These findings align with our own, highlighting the evaluative nature of the lay understanding of reasonableness, and raise further questions about whether the legal notion aligns with a more descriptive perspective. More comparative work is needed to map out the details of these contextual differences.

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 also demonstrates how experimental philosophy of language can contribute to legal theory and practice.

References

Almeida, G. d. F. C. F. d., Knobe, J., Struchiner, N., and Hannikainen, I. R. (2023). Purposes in law and in life: An experimental investigation of purpose attribution. *Canadian Journal of Law & Jurisprudence*, 36(1):1–36.

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.

Bear, A. and Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, 167:25–37.

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

Bystranowski, P., Janik, B., Próchnicki, M., Hannikainen, I. R., da Franca Couto Fernandes de Almeida, G., and Struchiner, N. (2022). Do Formalist Judges Abide By Their Abstract Principles? A Two-Country Study in Adjudication. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 35(5):1903–1935.

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.

Enoch, D. and Toh, K. (2013). Legal as a thick concept. In Waluchow, W. J. and Sciaraffa, S., editors, *Philosophical Foundations of the Nature of Law*, pages 257–278. Oxford University Press.

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.

Frisch, L. K., Kneer, M., Krueger, J. I., and Ullrich, J. (2021). The effect of outcome severity on moral judgement and interpersonal goals of perpetrators, victims, and bystanders. *European Journal of Social Psychology*, 51(7):1158–1171.

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.

Giestfeld, M. A. (2020). Folk tort law. In Hanoch, D. and Zipursky, B. C., editors, *Research Handbook on Private Law Theory*, pages 338–355. Edward Elgar Publishing.

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

Güver, L. and Kneer, M. (2023). Causation and the Silly Norm Effect. In Prochownik, K. and Magen, S., editors, *Advances in Experimental Philosophy of Law*. Bloomsbury Publishing. Forthcoming.

Hannikainen, I. R., Kneer, M., Tobia, K., Dranseika, V., Almeida, G. d. F. C. F. d., Poama, A., Strohmaier, N., Lopez, A. R., Bystranowski, P., and Struchiner, N. (2018). Experimental Jurisprudence Cross-Cultural Study Swap. https://doi.org/10.17605/OSF.IO/SK7R3.

Hannikainen, I. R., Tobia, K. P., De Almeida, G. D. F. C. F., Struchiner, N., Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., Janik, B., Lauraitytė, E., Laakasuo, M., Liefgreen, A., Neiders, I., Próchnicki, M., Rosas, A., Sundvall, J., and Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, 119(44):e2206531119.

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.

Herring, J. (2012). *Criminal Law: Text, Cases, and Materials*. Oxford University Press, 5th edition.

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

Jaeger, C. B. (2023). Reasonableness from an Experimental Jurisprudence Perspective. In Tobia, K., editor, *Cambridge Handbook of Experimental Jurisprudence*. Cambridge University Press.

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.

Jiménez, F. (2021). Some doubts about folk jurisprudence: The case of proximate cause. *University of Chicago Law Review Online*, 2021:1.

Jiménez, F. (2023). The Limits of Experimental Jurisprudence. In Tobia, K. P., editor, *The Cambridge Handbook of Experimental Jurisprudence*. Cambridge University Press. Forthcoming.

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.

Kirfel, L. and Hannikainen, I. R. (2023). Why blame the ostrich? Understanding culpability for willful ignorance. In Prochownik, K. and Magen, S., editors, *Advances in Experimental Philosophy of Law*. Bloomsbury Press. Forthcoming.

Kirfel, L. and Phillips, J. (2023). The pervasive impact of ignorance. *Cognition*, 231:105316.

Kneer, M. (2022). Reasonableness on the Clapham omnibus: Exploring the outcome-sensitive folk concept of reasonable. In Bystranowski, P., Janik, B., and Próchnicki, M., editors, *Judicial Decision-Making*, pages 25–48. Springer, Cham.

Kneer, M. and Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, 169:139–146.

Kneer, M. and Bourgeois-Gironde, S. (2018). Intention, cause, et responsabilité: Mens rea et effet Knobe. In Ferey, S. and G'Sell, F., editors, *Causalité, Responsabilité et Contribution à la Dette*, pages 117–144. Editions Brylant.

Kneer, M., Hannikainen, I., Zehnder, M.-A., Almeida, G., F., A., Bystranowski, P., Dranseika, V., Janik, B., Garcia Olier, J., Güver, L., Liefgreen, A., Tobia, K., Próchnicki, M., Rosas, A., Skoczeń, I., Strohmaier, N., and Struchiner, N. (2023). The Severity Effect on Intention and Knowledge: A cross-cultural study with laypeople and legal experts. In preparation. https://bit.ly/3aJBtO9.

Kneer, M. and Machery, E. (2019). No luck for moral luck. *Cognition*, 182:331–348.

Kneer, M. and Skoczeń, I. (2023). Outcome effects, moral luck and the hindsight bias. *Cognition*, 232:105258.

Knobe, J. and Shapiro, S. (2021). Proximate Cause Explained: An Essay in Experimental Jurisprudence. *The University of Chicago Law Review*, 88(1):165–236.

Kobick, J. and Knobe, J. (2009). Interpreting intent: How research on folk judgments of intentionality can inform statutory analysis. *Brooklyn Law Review*, 75:409.

Lagnado, D. A. and Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770.

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

Lee, T. R. and Mouritsen, S. C. (2018). Judging ordinary meaning. *The Yale Law Journal*, 127(4):788–879.

Lee, T. R. and Mouritsen, S. C. (2021). The corpus and the critics. *The University of Chicago Law Review*, 88(2):275–366.

Lidén, M. (2023). *Confirmation Bias in Criminal Cases*. Oxford University Press.

Lidén, M., Gräns, M., and Juslin, P. (2018). The presumption of guilt in suspect interrogations: Apprehension as a trigger of confirmation bias and debiasing techniques. *Law and Human Behavior*, 42:336–354.

Lidén, M., Gräns, M., and Juslin, P. (2019). 'Guilty, no doubt': detention provoking confirmation bias in judges' guilt assessments and debiasing techniques. *Psychology, Crime & Law*, 25(3):219–247.

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.

Macleod, J. A. (2015). Belief states in criminal law. *Oklahoma Law Review*, 68:497.

Margoni, F. and Brown, T. R. (2023). Jurors use mental state information to assess breach in negligence cases. *Cognition*, 236:105442.

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

Mott, C. and Heiphetz, L. (2023). Mens rea in criminal cases: How contrast affects attribution of culpable mental states. In preparation.

Mouritsen, S. C. (2010). The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review*, 2010:1915.

Murray, S., Krasich, K., Irving, Z., Nadelhoffer, T., and De Brigard, F. (2023). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General*, 152:120–138.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9(2):203–219.

Nobes, G. and Martin, J. W. (2022). They should have known better: The roles of negligence and outcome in moral judgements of accidental actions. *British Journal of Psychology*, 113(2):370–395.

Nyarko, J. and Sanga, S. (2022). A statistical test for legal interpretation: Theory and applications. *The Journal of Law, Economics, and Organization*, 38(2):539–569.

Posner, R. A. (2014). *Economic Analysis of Law*. Wolters Kluwer.

Prochownik, K., Feiertag, R., Horvath, J., and Wiegmann, A. (2023). How much harm does it take? An experimental study on legal expertise, the severity effect, and intentionality ascriptions. In Tobia, K., editor, *The Cambridge Handbook of Experimental Jurisprudence*. Cambridge University Press.

Prochownik, K., Krebs, M., Wiegmann, A., and Horvath, J. (2020). Not as bad as painted? legal expertise, intentionality ascription, and outcome effects revisited. In *42nd Annual Conference of the Cognitive Science Society*, pages 1930–1936.

Prochownik, K. M. (2021). The experimental philosophy of law: New ways, old questions, and how not to get lost. *Philosophy Compass*, 16(12):e12791.

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.

Skoczeń, I. and Smywiński-Pohl, A. (2022). The context of mistrust: Perjury ascriptions in the courtroom. In Horn, L. R., editor, *From Lying to Perjury*, pages 309–352. De Gruyter Mouton.

Spamann, H. and Klöhn, L. (2016). Justice Is Less Blind, and Less Legalistic, than We Thought: Evidence from an Experiment with Real Judges. *The Journal of Legal Studies*, 45(2):255–280.

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

Struchiner, N., Almeida, G. d. F. C. F. d., and Hannikainen, I. R. (2020a). Legal decision-making and the abstract/concrete paradox. *Cognition*, 205:104421.

Struchiner, N., Hannikainen, I. R., and Almeida, G. d. F. C. F. d. (2020b). An experimental guide to vehicles in the park. *Judgment and Decision Making*, 15(3):312–329.

Sytsma, J. (2019). The character of causation: Investigating the impact of character, knowledge, and desire on causal attributions.

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. (2021a). The Corpus and the Courts. *University of Chicago Law Review Online*, 2021:1.

Tobia, K. (2021b). Law and the Cognitive Science of Ordinary Concepts. *Law and Mind: A Survey of Law and the Cognitive Sciences*.

Tobia, K. (2022). Experimental jurisprudence. *The University of Chicago Law Review*, 89(3):735–802.

Tobia, K. (2023). Legal concepts and legal expertise. *Synthese*. forthcoming.

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

Tobia, K. P. (2020). Testing Ordinary Meaning. *Harvard Law Review*, 134:726.

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.

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/.

Väyrynen, P. (2012). Thick concepts: Where's evaluation? In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*, volume 7, pages 235–270. Oxford University Press.

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. (2023). Examining evaluativity in legal discourse : A comparative corpus-linguistic study of thick concepts. In Prochownik, K. and Magen, S., editors, *Advances in Experimental Philosophy of Law*. Bloomsbury Press. Forthcoming.

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? *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.