"Wrongful discrimination" - a tautological claim?
An empirical study of the evaluative dimension of discrimination vocabulary

Pascale Willemsen (pascale.willemsen@uzh.ch)
University of Zurich, Department of Philosophy, Zollikerstrasse 117, 8008 Zurich, Switzerland

Simone Sommer Degn
University of Aarhus, Department of Political Science, Bartholins Allé 7, 8000 Aarhus, Denmark

Jan Alejandro Garcia Olier (janalejandro.garciaolier@uzh.ch)
University of Zurich, Department of Philosophy, Zollikerstrasse 117, 8008 Zurich, Switzerland

Kevin Reuter (kevin.reuter@uzh.ch)
University of Zurich, Department of Philosophy, Zürichbergstrasse 43, 8044 Zurich, Switzerland

Abstract

Is it tautological to call an action "wrongful discrimination?" Some philosophers and political theorists answer this question in the affirmative and claim that the term "discrimination" is intrinsically evaluative. Others agree that "discrimination" usually conveys the action's moral wrongness but claim that the term can be used in a purely descriptive way. In this paper, we present two corpus studies and two experiments designed to test whether the folk concept of discrimination is evaluative. We demonstrate that the term has undergone a historical development and is nowadays no longer used purely descriptively. Further, we show that this evaluation cannot be cancelled without yielding a contradiction. We conclude that the descriptive use of "discriminatory" is a thing of the past.

Keywords: discrimination; evaluative language; thick concepts; cancellability test; corpus study

Introduction

The United Nations condemns and prohibits discrimination under the following definition: "Discrimination is any unfair treatment or arbitrary distinction based on a person's race, sex, religion, nationality, ethnic origin, sexual orientation, disability, age, language, social origin or other status." It seems clear enough that discrimination, so understood, is considered morally wrong.

Many philosophers and political theorists also conceive of discrimination in a negative, moralized sense. However, some scholars are using the term in a non-moralized, descriptive way. According to such a descriptive understanding, "discrimination" solely points to differential treatment, which might not inherently carry moral judgment—such treatment could be morally negative, neutral, or even positive.

It is often suggested that the philosophical concept of discrimination should align with ordinary language. It seems only plausible that philosophers and political theorists aim to engage with actual social and political debates and to address the same thing as, say, the UN. This raises an intriguing practical-conceptual issue: Does the moralized or non-moralized sense of discrimination mirror ordinary usage? Or, to put it differently, do speakers always and with necessity evaluate an action as morally problematic when they call it "discrimination" or "discriminatory", making formulations like "wrongful discrimination" tautological?

In this paper, we address this question in various ways: We present the results of two corpus studies, analyzing the adjectives that most frequently co-occur with "discriminatory" (Study 1) and "discrimination" (Study 2) over the last two centuries. A historical perspective on the conceptual changes that DISCRIMINATION has undergone indicates that while the term was once used descriptively, it has recently been replaced by a clearly evaluative use (Study 2). Furthermore, we present the results of the cancellability test for evaluative language (Studies 3 & 4). Our results clearly show that participants find it contradictory to call an action discriminatory but to deny its wrongfulness. Taken together, the empirical evidence presented here supports a moralized understanding of discrimination that does not allow for a descriptive use.

Background

Some scholars use the term "discrimination" in such a way that it is synonymous with "wrongful", "unfair" or "unjust" discrimination. It follows from this that by calling an act discriminatory, we also call it wrongful. According to such a moralized understanding, "[t]o claim that someone discriminates is … to challenge her for justification; to call discrimination 'wrongful' is merely to add emphasis to a morally-laden term" (Wasserman, 1998, p. 805; for a similar view, see Halldenius, 2005). The *Moralized View*, as we will call it, often reflects the use of the term in media reporting and political debates.

The *Non-Moralized* or *Descriptive View* uses the term "discrimination" to signify that it is a specific type of differential treatment based on certain social group traits. Proponents of the descriptive view usually argue that discrimination can be used both in a descriptive and moral sense (Waldron, 1995, p. 83; Hellman, 2008; Gardner, 2017). Advocates of this view do not believe that calling an action discriminatory necessarily communicates a negative moral judgment: there are instances in which discrimination is

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

morally acceptable or even required. The Descriptive View is compatible with believing that discrimination is *often* or even prototypically morally wrong; it is just not always (some descriptivists defend an even weaker version, according to which it is not *necessarily)* the case (Lippert-Rasmussen, 2013; Eidelson, 2015).

The disagreement between Moralized and Descriptive Views revolves around the question of whether "discrimination" semantically entails wrongfulness, that is: whether it is possible to call an action "discriminatory" but to *not* call it morally wrong by doing so. In this paper, we aim to answer the *conceptual* questions of whether the concepts of DISCRIMINATION and DISCRIMINATORY, as well as their corresponding terms, semantically entail moral wrongness and, more generally, a negative evaluation. [1] Thus, we investigate whether ordinary uses of the term are more in line with the Moralized or the Descriptive View.

Thick Evaluative Language

While discrimination is a central topic within political philosophy and political theory, very few empirical studies have explored the concept of discrimination (see Lippert-Rasmussen et al. 2024; Harnois, 2023 for exceptions). For our investigation, it makes sense to adopt a slightly different perspective, namely to discuss discrimination through the lens of metaethics and philosophy of language. As it stands, there are two options: DISCRIMINATION is either a descriptive concept or an evaluative, moralized concept with a descriptive dimension. In metaethics and philosophy of language, concepts that combine evaluative and descriptive content are also called "thick concepts", and given the moral nature of the evaluation in question, we suspect that DISCRIMINATION is a thick *ethical* concept.

Within the thick concept literature, it is an essential point of contention how exactly thick concepts hold their evaluative and descriptive contents together and whether it is possible to use that concept non-evaluatively. Willemsen and colleagues (Willemsen & Reuter, 2021; Willemsen et al., 2024) developed a useful empirical tool for investigating whether a value-laden term communicates its evaluation through lexical or pragmatic means, namely the cancellability test for evaluative language. Reuter, Baumgartner, and Willemsen (2023) further demonstrate how we can detect evaluative terms using corpus-linguistic means.

We believe the thick-concept debate mirrors the underlying disagreement about DISCRIMINATION, allowing the above-mentioned tools to investigate how the term "discrimination" conveys moral wrongness. Philosophers arguing for a moralized understanding have not been very explicit concerning the linguistic means by which "discrimination" conveys moral wrongness. However, given that they consider the negative evaluation *necessarily* conveyed, it is plausible to consider it part of the concept's lexical meaning. Accordingly, being wrongful would be part of the necessary, defining semantic features of discrimination.

Study 1: The Evaluative Nature of "Discriminatory"

In our first corpus study, we aim to identify the evaluative character and intensity of the term "discriminatory". Previous studies (e.g., Elhadad & McKeown 1990, Willemsen et al. 2023) have suggested investigating the evaluative aspects through the connective "and", which is frequently utilized to link adjectives of the same polarity. Consequently, if the term "discriminatory" is mainly deployed descriptively, we anticipate that this term will be frequently conjoined with other descriptive or even with positive terms, e.g., "discriminatory and specific". On the other hand, if the term is chiefly used as a negatively evaluative term, the conjoined adjectives should also predominantly be negative, like "unjust" and "offensive".

Methods

The NOW corpus (Davies, 2016-), a comprehensive collection of news articles from 2010 to the present, is accessible at https://www.english-corpora.org/now/. To obtain the desired outcomes, one merely inputs the query <ADJ and discriminatory> into the search field, yielding the results in Table 1. As a control term, we undertook an analysis of adjectives that are commonly paired with the term "selective". The term "selective" is conceptually akin to "discriminatory" in that it has the same or at least a very similar descriptive content without being evaluative. Some dictionaries even list the two terms as synonymous (e.g., Merriam-Webster).

In order to determine the evaluative polarity and intensity of the adjectives frequently conjoined with "discriminatory" and "selective", we used *sentiment values*. Sentiment dictionaries such as SentiWords (Baccianella et al., 2010; Gatti et al., 2016) encode both the polarity (positive vs. negative) and the intensity for an enormous number of adjectives. This comprehensive resource assigns sentiment values that span from '-1', denoting a strongly negative connotation, to '+1', indicative of a highly positive sentiment.

Results

Table 1 lists the most frequently conjoined terms. We calculated the weighted average SentiValue for the 20 most frequent adjectives in each condition. These terms comprise 50.7% of the usage for "discriminatory" and 48.4% for "selective", offering a comprehensive insight into how these terms are typically employed.[2]

[1] In this paper, we assume that terms and concepts stand in a relationship which is close enough to make inferences about the concepts from investigating the use of terms.

[2] We excluded "potent" from the analysis as "potent and selective" is a standard phrase in chemistry and does not reflect the ordinary usage of the term.

Table 1: List of the 10 most frequently conjoined terms with discriminatory (left) and selective (right).

discriminatory		selective	
Term	Number	Term	Number
unfair	556	potent	422
arbitrary	443	careful	139
racist	373	cautious	85
divisive	173	narrow	84
hateful	172	sensitive	55
unconstitutional	153	efficient	46
unjust	145	active	40
offensive	119	biased	39
sexist	91	partial	37

The analysis revealed a weighted average SentiValue of -0.512 for terms associated with "discriminatory", indicating a generally negative sentiment. Of the 100 most frequently connected adjectives with "discriminatory", only "selective" and "extreme" have a positive sentiment value according to the SentiWords dictionary.[3] In contrast, terms related to "selective" had a weighted average SentiValue of 0.109, suggesting a slightly positive or neutral sentiment. The independent t-test (t = -8,72, p < 0.001) demonstrates a statistically significant difference in the SentiValues between "discriminatory" and "selective".

Discussion

The findings suggest that the terms conjoined with "discriminatory" and "selective" possess significantly different average sentiment values. Specifically, terms associated with "discriminatory" tend to carry a very negative sentiment compared to those linked with "selective." These results provide substantial support for the view that "discriminatory" is an evaluative term.

Study 2: Time-Course Analysis of "Discrimination"

The meanings of words often evolve over time, as seen in examples like "gay", "broadcast", and "conspiracy theory" (Reuter & Baumgartner, forthcoming). Corpus Studies, using corpora such as Corpus of Historical American English (https://www.english-corpora.org/coha/) as well as Google NGRAM viewer are instrumental in tracking possible shifts in meaning. In Study 2, we show that an examination of the term "discriminatory" within the COHA and NGRAM viewer datasets reveal that while initially employed as a purely descriptive term, its usage has undergone a significant shift, transforming it into an evaluative term in contemporary usage.

When searching for <ADJ discrimination> in COHA, we find that the most frequent adjectives preceding "discrimination" in the first half of the 19th century were positive, such as "nice", "just", "careful", "clear", and "great".[4] An example from the North American Review in 1846 illustrates this: "Dr. Palfrey sums up very briefly, but with nice discrimination, the qualities on which his popularity as a preacher depended." From the 1860s, phrases like "unjust discrimination" and "unfair discrimination" gained prominence, emerging as the two predominant terms on COHA. However, and importantly, by the 1930s, these phrases started to strongly decrease again, and instead, people started to speak almost exclusively about specific forms of discrimination like "racial discrimination", with "religious discrimination" following as a distant second. Google's Ngram Viewer corroborates these findings (see Figure 1). Entering terms like "nice discrimination, unjust discrimination, racial discrimination" into its search field highlights the rise of negative adjectives in the 1860s and the subsequent dominance of "racial" in the context of discrimination.

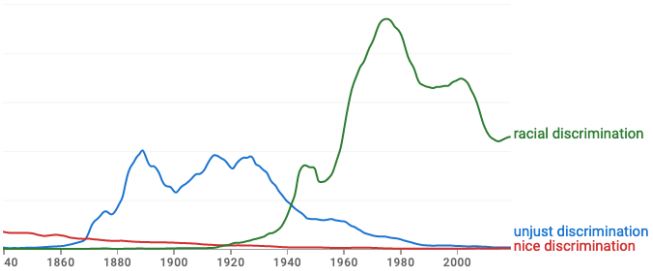

Figure 1. Ngram results for "nice discrimination", "unjust discrimination" and "racial discrimination" from 1840 - 2019.

Why did terms such as "unjust" and "unfair" initially emerge as the predominant adjectives paired with "discrimination", only to later fade from common usage? The explanation appears to be rooted in the evolving nature of the term "discrimination" itself. In the 19th century, when "discrimination" was primarily used in a descriptive sense, there was a perceived need to explicitly express disapproval by coupling it with adjectives like "unjust" and "unfair". However, over the following 80 years, the term "discrimination" underwent a semantic shift, acquiring an inherently evaluative dimension that conveyed strong disapproval. By the 1930s and 40s, the redundant use of these adjectives was no longer considered necessary. Thus, "discrimination" began to be used independently, its evaluative content understood in its standalone usage.

While it might be conjectured that the observed rise and fall in the frequency of "unjust discrimination" could be attributed to fluctuations in the use of "unjust" or "discrimination" individually, our analysis suggests otherwise. To explore this possibility, we examined potential correlations between these terms. Intriguingly, our findings indicate that the usage of "unjust discrimination" did not show a significant correlation with either "unjust" or "discrimination" independently. For this purpose, we sourced

[3] We checked those uses individually and could not find any evidence that even those uses of "discriminatory" are descriptive.

[4] While a time-course analysis of "discriminatory" would have been desirable, there are too few hits on COHA to do a proper statistical analysis.

relative occurrence frequencies from the Google NGRAM viewer and conducted a correlation analysis. Particularly from 1860 to 1960, the correlation coefficient was approximately -0.02, revealing a very weak and almost negligible negative correlation. This implies that, in this timeframe, the parallel trends in the use of "unjust" and "discrimination" had little to no bearing on the usage of the compound phrase "unjust discrimination".

Study 3: Discrimination and Its Descriptive Siblings

The corpus study provides clear evidence that the ordinary concept of discrimination is, in fact, evaluative. This may come as a surprise to advocates of the Descriptive View. Should it be possible to employ "discriminatory" in a purely descriptive manner, then, within a vast corpus like NOW, we would expect to find instances where "discriminatory" is used devoid of any evaluative component. However, Descriptivists may argue that our findings do not definitively prove that the term cannot be used in a descriptive sense.

In this study, we apply an experimental design, namely the cancellability test for evaluative language by Willemsen and Reuter (2021). In a nutshell, the cancellability test examines whether a piece of information derivable from a target statement can be explicitly denied without creating a contradictory statement. Our aim is thus to investigate whether it is possible for a speaker to explicitly deny a discriminatory act's wrongness.[5] If the Descriptivist view is right, cancelling the negative evaluation of "discriminatory" should sound felicitous, similar to descriptive terms like "differential" and "selective".[6] The Moralized View, on the other hand, may predict that cancelling the negative evaluation of "discriminatory" does yield a contradiction.

We recruited 1042 participants via Prolific who completed an online survey in Qualtrics. Participants were at least 18 years old, English native speakers, and had a minimal approval rate of previous studies of 90%. Before engaging with the actual experiment, participants were presented with a training round where they were asked whether two phrases were a "contradiction". Participants who failed both training round questions were excluded from the analyses. The final sample after exclusions consisted of 1019 participants (gender-balanced; M_{age} = 41.91).

Methods

For this experiment, we implemented a 2 × 4 × 2 full-factorial design with Phrase (policy, behavior), Concept (discriminatory, differential, preferential, selective), and Cancellability Clause (wrongness, evaluation) as between-subject factors and contradiction ratings as our dependent measure. Participants were randomly assigned to one of 16 between-subject conditions, consisting of one of the following two test phrases:

1. Policy (CP): The corporation's policy is X…
2. Behavior (B): This behaviour is X…

"X" stands for one of the four concepts. Finally, we added one of two Cancellability Clauses to our test phrases, namely:

1. Wrongness: … but it is not morally wrong.
2. Evaluation**:** … but by that I am not saying something negative about it. I mean this in a fully neutral way.

Here is an example of the phrases we used for our experiment: "The corporation's policy is discriminatory but it is not morally wrong."

Participants then had to rate the extent to which the phrase was a contradiction, using a scale from 1 = "definitely not" to 9 = "definitely yes."

In line with the theoretical discussion, we pre-registered several hypotheses which are described in greater detail in the pre-registration. Most central to the debate between the Moralized and Descriptive Views is, however, the prediction that Concept (but not Phrase) has a significant effect on contradiction ratings. More specifically, we predicted that, regardless of the Cancellability Clause, contradiction ratings for "discriminatory" would be significantly above the neutral midpoint and also above its descriptive counterparts—differential, preferential, and selective.

Results

We conducted a 2 × 4 × 2 global ANOVA with Phrase (policy v. behavior), Concept (discriminatory, differential, preferential, and selective), and Cancellability Clause (wrongness v. evaluation) as between-subjects factors. The analysis revealed a small, but significant effect of Phrase ($F(1, 1003) = 13.150$, $p < .001$; $\eta_p^2 = .013$) and a large effect of Concept Class ($F(3, 1003) = 87.279$; $p < .001$, $\eta_p^2 = .207$).[7] The results are depicted and detailed in Figure 2.

We took a closer look at the data and conducted planned comparisons, keeping Phrase constant (either policy or behavior) across conditions. Per our pre-registration, we also kept the Cancellability Clause conditions fixed in either "wrongness" or "evaluation". The analyses revealed that, when the test Phrase was Policy, mean contradiction ratings for "discriminatory" in both the wrongness (M = 6.02; SD = 3.015) and evaluation (M = 6.19; SD = 2.760) cancellability clause conditions were significantly above the neutral midpoint (all $ps < .005$; *all* $ds > .33$). When the Phrase was Behavior, only in the evaluation cancellability condition were the mean contradiction ratings for "discriminatory" (M = 5.69; SD = 2.850) significantly higher than the neutral midpoint ($p = .029$; $d = .24$). In the wrongness condition, mean contradiction ratings for "discriminatory" (M = 5.05;

[5] For other applications of this test, see, e.g. Almeida, Struchiner, & Hannikainen (2021), Baumgartner et al. (2022), Coninx et al. (2023), and Sytsma et al. (2023).

[6] All supplementary materials for Studies 3 and 4 can be found here.

[7] The three-way ANOVA also revealed a significant interaction between the concept class and cancellability clause variables ($F(3, 1003) = 4.446$, $p = .004$; $\eta_p^2 = .013$) and a significant main effect of the cancellability clause on contradiction ratings ($F(1, 1003) = 30.365$, $p < .001$; $\eta_p^2 = .029$). There was no significant three-way interaction between Phrase, Concept class, and Cancellability Clause on contradiction ratings ($F(3, 1003) = 1.361$, $p = .253$; $\eta_p^2 = .004$).

SD = 3.300) were not significantly above the neutral midpoint (*p = .455; d = .014*).

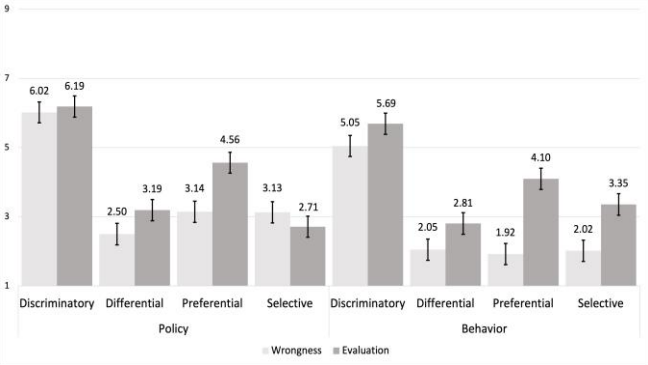

Figure 2: Mean contradiction ratings for all conditions. The error bars indicate the standard error around the means.

These results provide some evidence for the prediction that contradiction ratings for "discriminatory" would be above the neutral midpoint in both Cancellability Clause conditions. Next, we observed that regardless of which Phrase or Cancellability Clause conditions we kept constant, mean contradiction ratings for "differential", "preferential", and "selective" were significantly below the neutral midpoint (all *Ms* < *4.11*) all *ps* < *.009*; all *ds* > *.30*), except for "preferential" when Phrase test was Policy and the Cancellability Clause was Evaluation (*M = 4.56; p* = .093; *d* = .15). Also, regardless of the Phrase or Cancellability Clause condition, mean contradiction ratings for "discriminatory" (all *Ms* > 5.05) were significantly higher than for "differential", "preferential", and "selective" (all *Ms* < *4*.57; all *ps* < *.001*, all *ds* > *.59*).

Discussion

The results provide further evidence in support of the Moralized View and raise serious doubts about the notion that "discriminatory" and "differential" can be used interchangeably. However, we also recognized a potential confound in our design. Many participants (117 out of 250) reported issues with understanding the term "differential" or what the sentences featuring the term meant. We, therefore, decided to choose a different approach that can test the Moralized and Descriptivist Views' key differences without relying on the term "differential".

Study 4:

In this pre-registered experiment, we ran a second cancellability study to investigate whether a speaker can call either a corporation's policy or a person's behavior discriminatory, but then deny that the policy/behavior (a) is morally wrong, or (b) that the speaker was saying something negative. As control cases, we used a variety of purely descriptive terms (common, new, unexpected), negative thick concepts (offensive, disrespectful, unfair), specific discrimination terms (racist, sexist, homophobic), and the term "selective", which worked as a good comparison in Study 3, as well as Corpus Study 1.

For this follow-up experiment, we recruited 1305 participants via Prolific who completed an online survey in Qualtrics. Participants were at least 18 years old, English native speakers, and had a minimal approval rate in previous studies of 90%. Before engaging with the actual experiment, participants were presented with a training round in which they were asked whether two phrases were a "contradiction". Participants who failed both training round questions were excluded from the analyses. The final sample consisted of 1256 participants (gender-balanced; M_{age} = 38.44).

Methods

We implemented a 2 × 5 × 2 between-subjects design with Phrase, Concept Class, and Cancellability Clause as between-subject factors. Participants were randomly assigned to one of our 20 between-subjects conditions. The test Phrases and Cancellability Clauses remained the same as in Study 3. The five Concept Classes were the following:

1. Neutral: with "unexpected", "new", and "common"
2. Negative Control: "unfair", "disrespectful", and "offensive".
3. Negative Discrimination: "racist", "sexist", and "homophobic".
4. "Discriminatory".
5. "Selective".

We selected the Negative Control terms based on two criteria. First, these terms are regarded as primary examples of thick ethical concepts and communicate a negative evaluation. Second, we selected evaluative terms that seem fitting in the context of discrimination. Discriminatory acts are usually unfair, demonstrate a lack of respect, and may cause offense. For Negative Discrimination terms, we only chose concepts that express some specific form of discrimination (based on race, sex, or sexual orientation), are familiar enough (other than "ageist"), are expressed by one single term, and do not include the term "discrimination".

Participants assigned to the Neutral, Negative Control, or Negative Discrimination Concept Classes were presented with all three terms of the class. For our analyses, we calculated and used the mean contradiction rating for the three terms of each class. Participants assigned to Concept Classes 4 and 5 only received a single target term, namely either "discriminatory" or "selective".

For this experiment, we predicted that Concept Class and Cancellability Clause had significant main and interaction effects on contradiction ratings. We also predicted that, regardless of the Cancellability Clause, contradiction ratings for "discriminatory" would be significantly higher than the neutral midpoint, the neutral Concept Class, and the "selective" concept. The hypotheses for this experiment are described in greater detail in the pre-registration.

Results

We first examined the potential effect of Phrase, Concept Class, and Cancellability Clause, as well as a two-way

interaction between Cancellability Clause and Concept Class on contradiction ratings. To this end, we conducted a 2 × 5 × 2 ANOVA with Phrase, Concept Class, and Cancellability Clause as between-subjects factors. Consistent with Study 3, we observed an effect of Concept Class (*F(4, 1236) = 230.152; p < .001;* η_p^2 = .427) and Cancellability Clause on contradiction ratings (*F(1, 1236) = 14.773; p < .001;* × = .012). We also found a significant interaction of Concept Class and Cancellability Clause (*F(4, 1236) = 28.491; p < .001;* η_p^2 = .084). In contrast to Study 3, however, we found no effect of Phrase (*F(1, 1003) = 13.150, p < .001;* η_p^2 = .013). The results are depicted and detailed in Figure 3.

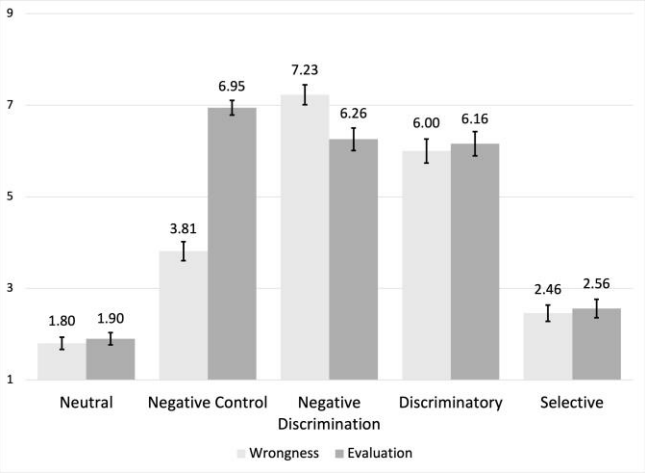

Figure 3. Mean contradiction ratings for all conditions. The error bars indicate the standard error of the means.

As in our Study 3, we conducted a series of planned comparisons to explore our specific hypotheses. Consistent with our prediction, we found that the mean contradiction rating for the evaluation Cancellability Clause (*M* = 4.77, *SD* = 3.021) was significantly higher than for wrongness (*M* = 4.24, *SD* = 3.077), (*p = .001; d = .17*) across Concept Class conditions. Also in line with our predictions, we observed that, in the wrongness Cancellability Clause condition, mean contradiction ratings for "discriminatory" (*M* = 6.00, *SD* = 2.883) were not only significantly above the neutral midpoint (*p < .001; d = .34*), but also higher than the Concept Class "neutral" (*M* = 1.80, *SD* = 1.550; *p < .001; d = 1.82*) and "selective" (*M* = 2.46, *SD* = 2.018; *p < .001; d = 1.42*). In the evaluation clause condition, mean contradiction ratings for "discriminatory" (*M* = 6.16, *SD* = 2.970) were significantly higher compared to the neutral midpoint (*p < .001; d = .39*), the "neutral" Concept Class (*M* = 1.90, *SD* = 1.499; *p < .001; d = 1.81*), and the concept "selective" (*M* = 2.56, *SD* = 2.259; *p < .001; d = 1.36*).[8]

Discussion

These results provide further evidence in favor of the evaluative view of discrimination and against the descriptive view. "Discriminatory" behaves quite differently from neutral terms and very similarly to other paradigmatic cases of negative thick concepts and other discrimination terms which are agreed by all sides to be negatively valenced.

General Discussion

We have examined the concept of DISCRIMINATION from different empirical perspectives. Study 1 and Study 2 suggest that while the historical use of "discrimination" was descriptive, it has since been replaced by a clearly evaluative term according to the corpus-linguistic analysis. The findings from the cancellability tests in Studies 3 and 4 provide further support for the evaluative view of discrimination. The empirical evidence of this article supports the Moralized View of discrimination. Whether ordinary usage of DISCRIMINATION is reflected among academics has important implications for how scholars communicate with each other and with the public. Researchers who subscribe to the Moralized View reflect ordinary usage, meaning that defenders of the Non-Moralized View should motivate why they diverge from ordinary usage when they do. In other words, our findings suggest that "discrimination is wrongful" is a tautological claim. Further studies may investigate whether specific examples of what the Descriptive View calls morally-neutrally "discrimination" mirrors laypeople's use, e.g., through a vignette study.

[8] Following our pre-registration, we conducted an ANOVA with both Cancellability Clauses (wrongness and evaluation) and the Concept Class conditions "discriminatory" and "negative control". Consistent with our prediction, the two-way interaction was small but significant (*F(1, 496) = 43.243; p < .001;* η_p^2 = .080). To take a closer look at the interaction, we conducted further pairwise comparisons keeping the Cancellability Clause fixed in either wrongness or evaluation. Also consistent with our prediction, in the wrongness clause condition contradiction ratings for "discriminatory" (*M* = 6.00, *SD* = 2.883) were significantly higher than for the negative control concept class (*M* = 3.81, *SD* = 2.309; *p < .001; d = .83).* In the evaluation condition, however, contradiction ratings for "discriminatory" (*M* = 6.16, *SD* = 2.970) were significantly lower than for the "negative control" concept class (*M* = 6.94, *SD* = 1.800; *p = .006; d = .32).*

A final set of planned comparisons revealed yet further predicted patterns. Consistent with previous results by, e.g., Willemsen & Reuter (2021) and Willemsen et al. (2024), in the evaluation condition, the mean contradiction rating for the "negative control" Concept Class (*M* = 6.94, *SD* = 1.800) was significantly above the neutral midpoint (*p < .001; d = 1.08).* Additionally, the mean contradiction ratings for "negative discrimination" concepts in both the wrongness (*M* = 7.23, *SD* = 2.451) and evaluation (*M* = 6.26, *SD* = 2.732) conditions were also significantly above the neutral midpoint (*ps < .001; ds > .46).*

Acknowledgments

We would like to thank three anonymous referees for their comments, as well as members of the Zurich XPhi Lab and the members of CEPDISC and the Political Theory section (Aarhus) for their feedback. Pascale Willemsen's research was funded by the Swiss National Science Foundation (SNF), project number PZ00P1_201737. Simone Sommer Degn was supported by the Danish National Research Foundation (DNRF144). Kevin Reuter was funded by an SNF Eccellenza Grant [PCEFP1_18108].

References

Almeida, G., Struchiner, N., & Hannikainen, I. (2021). Rule is a dual character concept. Available at *SSRN* 4018823.

Altman, A. (2016). Discrimination. *The Stanford Encyclopedia of Philosophy (*Winter 2016 Edition), Edward N. Zalta (Ed.). URL: https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, 10*, 2200–2204.

Baumgartner, L., Willemsen, P., & Reuter, K. (2022). The polarity effect of evaluative language. *Philosophical Psychology*, 1-18.

Coninx, S., Willemsen, P., & Reuter, K. (2024). Pain linguistics: A case for pluralism. *The Philosophical Quarterly, 74(1)*, 145-168.

Davies, M. (2016-). *Corpus of News on the Web (NOW).* Available online at https://www.english-corpora.org/now/

Elhadad, M., McKeown, K. (1990). Generating Connectives. Proceedings of the 13th conference on Computational linguistics, 3, 97–101.

Eidelson, B. (2015). *Discrimination and Disrespect*. Oxford University Press.

Gardner, J. (2018). Discrimination: The Good, the Bad, and the Wrongful. *Proceedings of the Aristotelian Society*, 118(1), 55–81.

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.

Halldenius, L. (2005). Dissecting "Discrimination". *Cambridge Quarterly of Healthcare Ethics*, 14(4), 455–463.

Harnois, C. E. (2023). The Multiple Meanings of Discrimination. *Social Psychology Quarterly*, 86(4), 413–431.

Hellman, D. (2008). *When is Discrimination Wrong?* Cambridge, MA: Harvard University Press.

Lippert-Rasmussen, K., Serritzlew, S., Laustsen, L., Sommer D., S., Albertsen, A. (2024). What is the folk concept of discrimination? Discriminators and comparators. *Philosophical Psychology*.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford University Press.

Reuter, K., Baumgartner, L. (forthcoming). Corpus analysis: A case study on the use of 'conspiracy theory', in Kornmesser, Bauer et al. (Eds.) *Experimental Philosophy for Beginners*. Springer Nature.

Reuter, K., Willemsen, P., Baumgartner, L. (2023). Tracing thick and thin concepts through corpora. *Language and Cognition*, 1–20.

Sytsma, J., Willemsen, P. & Reuter, K. (2023). Mutual entailment between causation and responsibility. *Philosophical Studies*, 180, 3593–3614.

Waldron, J. (1985). 'Indirect Discrimination'. In Guest Stephen, Milne Alan (Eds.), *Equality and Discrimination: Essays in Freedom and Justice*, pp. 93–100. Stuttgart: F. Steiner Verlag.

Wasserman D. (1998). Discrimination, concept of. In: Chadwick R. (Ed.), *Encyclopaedia of Applied Ethics*. Academic Press, San Diego, pp. 805–814.

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

Willemsen, P., Baumgartner, L., Cepollaro, B. and Reuter, K. (2024), Evaluative Deflation, Social Expectations, and the Zone of Moral Indifference. *Cognitive Science, 48*, e13406.