Moral Rightness Comes in Degrees

ABSTRACT: *This article questions the traditional view that moral rightness and wrongness are discrete predicates with sharp boundaries. I contend that moral rightness and wrongness come in degrees: Some acts are somewhat right and somewhat wrong. My argument is based on the assumption that meaning tracks use. If an overwhelming majority of competent language users frequently say that some acts are a bit right and a bit wrong, this indicates that rightness and wrongness are gradable concepts. To support the empirical part of the argument I use the tools of experimental philosophy. Results from three surveys (*n = 715, 578, *and* 182*) indicate that respondents use* right *and* wrong *as gradable terms to approximately the same extent as color terms, meaning that rightness and wrongness come in degrees roughly as much as colors do. In the largest study, only 4 percent persistently used* right *and* wrong *as non-gradable terms.*

Introduction

There is an undisputed sense in which moral rightness and wrongness come in degrees. As Thomas Hurka explains, 'it is wrong to steal a car and wrong to murder, but murder is more seriously wrong than auto theft, which is more seriously wrong than breaking a promise to have lunch' (Hurka 2019: 41). It is also, in a similar sense, more importantly right to eradicate world poverty than to complete a book review on time.

Be that as it may. There is also another, theoretically more interesting sense in which moral rightness and wrongness may come in degrees. According to what I call the gradualist hypothesis, some acts are somewhat right *and* somewhat wrong. We can think of such acts as being located in a moral gray area. The gradualist hypothesis is compatible with, but does not entail, the idea that rightness and wrongness are vague concepts. An act can be somewhat right and somewhat wrong even if the boundary that separates the gray area from right and wrong is sharp. (Example: an act could be entirely right just in case none of the agent's moral obligations conflict, but somewhat right *and* somewhat wrong if they do.) By definition, vague concepts allow for Sorites series—that is, for sequences of similar instances in which there is no sharp boundary between correct and incorrect applications of a concept. Miriam Schoenfield suggests that some abortion cases may fit this description: 'Plausibly, we can create a Sorites series, admitting of borderline cases of permissibility, out of a series of abortions in which the fetus' age differ by a day (or a minute, or a second)' (Schoenfield 2016: 263).

The argument for the gradualist hypothesis that I offer in this article is based on the familiar idea that meaning tracks use. If an overwhelming majority of competent

CrossMark

language users frequently *say* that some acts are a bit right and a bit wrong, this indicates that rightness and wrongness *are* gradable concepts. The assumption that meaning tracks use can of course be accepted by people who reject the traditional idea that meaning *is* use in a strict semantic sense. It is uncontroversial that use is a reliable guide to meaning, but some of us reject Wittgenstein's claim that 'the meaning of a word *is* its use in the language' (Wittgenstein [1953] (2009): 25e, at §43, my emphasis).

If we wish to turn the meaning-tracks-use thesis into an argument for gradualism in ethics, it is not sufficient to show that a few individuals occasionally use right and wrong in a gradable sense. Not every usage of a word is thoughtful and sincere. However, if people *persistently* and *sincerely* talk about moral rightness and wrongness as if they allow for degrees, then this would be a strong reason for accepting the gradualist hypothesis. On a strict interpretation of the meaning-tracks-use thesis, an overwhelming majority of competent language users cannot be wrong about their persistent and sincere usage of moral terms.

The meaning-tracks-use thesis is widely endorsed by modern linguists. For instance, computational linguist Katrin Erk explains that so-called vector space models of meaning are based on the observation that, 'we can often guess what a word means from the contexts in which it is used. Thus, we can represent meaning as distribution, as observed contexts' (Erk 2012: 635). Vector space models have proven to be empirically successful. By observing how words are used in large corpora, by counting their occurrence in different sentences, computational linguists have developed software that predicts whether two words have similar meaning. (See, for example, Erk 2012: 637; Turney and Pantel 2010: 160). However, the computational approach requires large quantities of empirical data. Erk points out that, 'many phrases do not occur with sufficient frequency in a corpus to be represented through their distributional contexts' (Erk 2012: 637). This includes phrases relevant for assessing the gradualist hypothesis, such as 'this act is somewhat right and somewhat wrong' and 'this act is a bit right and a bit wrong'. To overcome this problem I offer support for the gradualist hypothesis based on data from three surveys ($n = 715$, 578, and 182) designed to test whether ordinary people use right and wrong as gradable terms. In the largest survey, no more than 4 percent of ordinary language users persistently used 'right' and 'wrong' as non-gradable terms. The statistical analysis also indicates that 'right' and 'wrong' are used as gradable terms to approximately the same extent as color terms, meaning that rightness and wrongness come in degrees just as much as colors do. Furthermore, by using multidimensional scaling techniques, it can be shown that rightness and wrongness are used in a gradable sense even if no gradable terms (such as 'right to some degree' or 'somewhat right and wrong') appear in the questions or answer options presented to respondents.

1. The Gradualist Hypothesis

The gradualist hypothesis is of recent origin. John Stuart Mill writes that, 'The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that acts are right *in proportion* as they tend to promote

happiness, wrong as they tend to produce the reverse of happiness' (Mill 1863: 9, my emphasis). If we read this literally, it follows that acts that produce half as much happiness as the optimal alternative(s) are half-right. However, Mill's gradualism seems to have been a slip of the tongue. In the rest of his writings, he never discusses this implicit notion of degrees of rightness, and he never uses the idea that rightness comes in degrees for supporting other claims or for rebutting objections to his ethical theory.

Many modern consequentialists believe that an act is right just in case no alternative brings about better consequences, and that every act that is not right is wrong. This criterion of rightness is obviously incompatible with the gradualist hypothesis. No matter what the consequences are, every act will be either right or wrong in the binary sense. This holds true even if some consequences are incomparable or on a par, because an act is considered to be right as long as its consequences are *not worse* than those of any alternative act. However, some consequentialist theories do allow for gradable notions of right and wrong. According to the multidimensional account of consequentialism that I discuss in *The Dimensions of Consequentialism* (Peterson 2013), an act's rightness or wrongness depends on several irreducible aspects, such as the total wellbeing produced by the act *and* the degree of equality with which wellbeing is distributed in society. According to this theory, moral rightness comes in degrees whenever no act is optimal with respect to all moral aspects.

In the example in table 1 (see Peterson 2013: 13) the first act, A, is optimal with respect to total wellbeing but scores poorly with respect to equality. The opposite is true of C. Act B is not optimal with respect to any of these aspects, but it scores pretty well with respect to both of them. Multidimensional consequentialists believe that all three alternatives are somewhat right and somewhat wrong, but B is right to a *higher degree* than A and C all things considered. Multidimensional consequentialists may, for instance, argue that C is almost entirely right (right to degree 0.99 on a scale from 0 to 1) while A and C are half right and half wrong (right to degree 0.5).

Table 1 Multidimensional consequentialists believe that B is more right than A and C, but not entirely right.

Act	Alice's well-being	Bob's well-being
A	99	1
B	50	49
C	0	0

Some deontologists also accept the idea that rightness and wrongness come in degrees. In *Moral Uncertainty and Its Consequences*, Ted Lockhart claims that 'actions come in varying degrees of moral rightness between "right" and "wrong"' (Lockhart 2000: 75). He argues that this view is particularly attractive for those who accept W. D. Ross's theory of prima facie duties: 'Even for some deontological theories, we have no great difficulty entertaining a many-valued concept of moral rightness. A prima facie duties theory, for example, may readily

regard actions that, ceteris paribus, accord with more prima facie duties than others as having greater moral rightness. Moreover, for some prima facie duties, it seems that we can conform to or violate them to greater or lesser degrees. Examples of such duties in W. D. Ross's theory include beneficence, nonmaleficence, and justice' (Lockhart 2000: 80).

Lockhart's proposal is a revisionary extension of Ross's theory, which Ross himself would not accept. Ross is committed to what Rob Lawlor calls a 'threshold account' of rightness, just as mainstream consequentialists: acts that have the right kind of properties make it across the threshold, and those acts are right (or permissible) in the binary sense (Lawlor 2009: 21). All other acts are wrong. For mainstream consequentialists, the right-making property is that of having optimal consequences, and for Ross the threshold a right act has to pass is to be fitting to perform in light of all applicable prima facie duties.

Somewhat surprisingly, Lockhart's own proposal is a hybrid view that collapses into a threshold account when only two alternatives are available: 'x has greater degree of moral rightness than y in situation S for agent A just in case, if x and y were the only alternatives open to A in S, then x would be morally right for A in S and y would be morally wrong for A in S' (Lockhart 2000: 81).

Lockhart's thesis is compatible both with Hurka's commonsensical version of gradualism mentioned above, and with Mill's gradualist account of utilitarianism. However, as pointed out by Campbell Brown, Lockhart's reductive analysis is unable to compare degrees of rightness across different situations. Brown notes that, 'a person who uses her mobile phone while watching a movie in a public cinema may be acting wrongly, but surely not as wrongly as one who does the same while driving a heavy lorry on a motorway thereby causing a major accident' (Brown 2016: 27). The problem for Lockhart is that the acts described by Brown are performed in different situations and can therefore not be compared according to his criterion. To address this problem, Brown develops a more general reductive analysis of degrees of rightness that tallies well with Mill's gradualist theory mentioned above (see Brown 2016).

2. Three Experimental Studies

Should we accept the idea that right and wrong are gradable concepts? Almost everyone agrees that meaning tracks use in the sense endorsed by modern linguists. By studying how people use words and phrases we can, under normal circumstances, make reliable inferences about their meaning. As noted above, this is an epistemic thesis. Whether meaning *is* use in a strict semantic sense is another issue. The truth of the gradualist hypothesis can thus be assessed by testing the following empirical conjecture:

> Competent language users persistently and sincerely use gradualist notions of moral rightness and wrongness.

To test this conjecture, I distributed a series of online questionnaires to three sets of respondents: students at Texas A&M University taking classes in engineering ethics in spring 2019 ($n = 715$) and spring 2020 ($n = 578$), and a group of US

citizens who voted in the 2016 election ($n = 182$). Students took the surveys for credit (about 0.5 percent of the total course grade) while respondents in the third study received between \$2.65 and \$3.00 in compensation. In all three studies, respondents were invited to answer up to twelve questions. Both the order of the questions and the answer options were randomized. At the request of the Institutional Review Board of Texas A&M University, which approved the study, no demographic information was collected.

Respondents were presented with up to five different types of tasks: abstract, semi-abstract, concrete, comparative, and open-ended tasks (detailed below). By using several different kinds of measures that test a single hypothesis the validity of the measurement instrument can be assessed. If the validity is high, we should expect the results to be roughly the same for all types of tasks.

3. Abstract Tasks

Subjects were invited to evaluate abstract statements about ethics, mathematics, colors, scientific evidence, and scientific facts on a seven-point Likert scale ranging from 'strongly disagree' (1) to 'strongly agree' (7). The following example pertains directly to the gradualist hypothesis:

A1. Moral rightness and wrongness come in degrees. The boundary that separates morally right acts from wrong ones is not always sharp. Some acts are somewhat right and somewhat wrong.

Study 1: average degree of agreement 5.7 ($n = 237$, std. dev. 1.4)
Study 2: average degree of agreement 5.4 ($n = 114$, std. dev. 1.3)
Study 3: average degree of agreement 5.4 ($n = 181$, std. dev. 1.5)

The following abstract statements were used as reference points for comparative purposes:

A2. In mathematics, truth comes in degrees. The boundary that separates true mathematical statements from false ones is not always sharp. Some mathematical statements are somewhat true and somewhat false.

Study 1: average degree of agreement 2.7 ($n = 219$, std. dev. 1.7)
Study 2: average degree of agreement 2.6 ($n = 79$, std. dev. 1.6)
Study 3: average degree of agreement 2.2 ($n = 181$, std. dev. 1.5)

A3. Colors come in degrees. The boundary that separates one color from another is not always sharp. Some color hues are somewhat red and somewhat blue.

Study 1: average degree of agreement 5.9 ($n = 233$, std. dev. 1.1)
Study 2: average degree of agreement 5.7 ($n = 80$, std. dev. 1.2)
Study 3: average degree of agreement 6.2 ($n = 182$, std. dev. 1.2)

All items were followed by a simple comprehension check. Responses from subjects who did not answer it correctly have been excluded from the analysis.

The data sets are not normally distributed. It is therefore appropriate to perform a Mann-Whitney U-test. This is a nonparametric test for the null hypothesis that the distributions of two data sets are identical and therefore have the same median value. Table 2 summarizes the results for Study 2, which is less likely than the others to yield significant results due to its smaller sample size. The Mann-Whitney U-test indicates that moral rightness and wrongness is judged to come in degrees to a significantly higher extent (p <0.01, one-tailed) than truth in mathematics is judged to come in degrees, but there is no statistically significant difference between A1 (moral rightness and wrongness come in degrees) and A3 (colors come in degrees), not even at p <0.05. This indicates that rightness and wrongness is judged to come in degrees to approximately the same extent as colors are judged to come in degrees. The results of Study 1 offer additional support to this conclusion: all comparisons in Study 1 yield significant results ($p < 0.01$, one-tailed) except that between colors and moral rightness. However, in Study 3 the comparison between colors and moral rightness also yields a significant difference (p <0.01, one-tailed), indicating that the extent to which colors and moral rightness is reported to come in degrees is not *precisely* the same.

Table 2 Mann-Whitney U-tests for data in Study 2.

	A2. In mathematics, truth comes in degrees . . . ($n = 79$)	A3. Colors come in degrees . . .($n = 80$)
A1. Moral rightness and wrongness come in degrees . . . ($n = 114$)	*Mann-Whitney U = 1111 p-value <0.00001. Significant at p < 0.01; one-tailed.*	*Mann-Whitney U = 4014 p-value = 0.0778. Not significant at p < 0.05; one-tailed.*

The abstract tasks also included the following general statements about scientific evidence and scientific facts:

A4. Scientific evidence comes in degrees. The boundary that separates theories corroborated by evidence from those that are not is not always sharp. Some scientific theories are somewhat supported by evidence and somewhat unsupported.

Study 1: average degree of agreement 4.2 ($n = 239$, std. dev. 1.8)
Study 2: average degree of agreement 4.8 ($n = 97$, std. dev. 1.4)
Study 3: average degree of agreement 4.3 ($n = 182$, std. dev. 1.8)

A5. In science, facts come in degrees. The boundary that separates correct scientific claims from incorrect ones is not always sharp. Some scientific claims are somewhat correct and somewhat incorrect.

Study 1: average degree of agreement 3.4 ($n = 236$, std. dev. 1.8)
Study 2: average degree of agreement 4.1 ($n = 68$, std. dev. 1.7)
Study 3: average degree of agreement 3.9 ($n = 185$, std. dev. 1.8)

In all three studies, the Mann-Whitney U-test for A4 versus A5 indicates that scientific evidence is judged to come in degrees to a somewhat higher extent than scientific facts in all three studies (in Study 2, $U = 2615.5$, p-value < 0.00001; significant at $p < 0.01$; one-tailed). Moreover, scientific facts are also judged to come in degrees to a significantly higher extent than mathematical truths in all three studies (in Study 2, $U = 1454.5$, p-value < 0.00001; significant at $p < 0.01$; one-tailed), and moral rightness and wrongness are reported to come in degrees to a significantly higher extent than scientific evidence in all three studies (in Study 2, $U = 3792.5$, p-value = 0.00004; significant at $p < 0.01$; one-tailed). Finally, colors are reported to come in degrees to a significantly higher extent than scientific evidence in all three studies. (In Study 2, $U = 2317$, p-value < 0.00001; significant at $p < 0.01$; one-tailed).

These findings can be represented as a *hierarchical order* with four ordinal levels. Table 3 summarizes the results for Study 1 and 2. (As noted, in Study 3 colors come in degrees to a slightly higher extent than moral rightness and should therefore be represented on a separate sublevel.) The difference in agreement in table 3 between all pairs of levels is statistically significant at $p < 0.01$, except that between scientific evidence and scientific facts, which is significant at $p < 0.05$.

Table 3 Four ordinal levels of agreement in Study 1 and 2. Level 1 is the highest level, and all pair-wise differences between items at different levels are statistically significant at $p < 0.01$; one-tailed.

Level 1	Colors come in degrees . . . Moral rightness and wrongness come in degrees . . .
Level 2	Scientific evidence comes in degrees . . .
Level 3	In science, facts come in degrees. . .
Level 4	In mathematics, truth comes in degrees . . .

An alternative explanation of the results in table 3 could be that respondents tend merely to report their belief in how much *disagreement* there is in a certain domain, not that the phenomena themselves come in degrees. To control for this, the following items were included in Study (table 4):

Table 4 Moral disagreement does not explain the findings.

	N	Avg.	Std. dev.
It is common that people disagree on moral issues.	114	6.1	1.0
Moral rightness and wrongness sometimes come in degrees even when there is no disagreement.	114	5.2	1.3
It is common that people disagree on mathematical issues.	79	3.2	1.6
Truth in mathematics sometimes come in degrees even when there is no disagreement.	79	3.2	1.6

These numbers indicate that respondents do believe that there is more disagreement on moral issues than on mathematical issues ($U = 768$, p-value < 0.00001; significant at $p < 0.01$; one-tailed). This is not surprising. However,

respondents *also* believe that moral rightness and wrongness come in degrees even when there is no disagreement, and they do so to a much higher extent than for mathematical issues. ($U = 1587.5$, p-value < 0.00001; significant at $p < 0.01$; one-tailed). This casts doubt on the alternative explanation, but fits well with the gradualist hypothesis. There is little reason to think that respondents merely reported their belief in how much disagreement there is in a certain domain.

4. Semi-abstract Tasks

All three studies included two semi-abstract tasks in which subjects were invited to complete moral statements by selecting one of a set of pre-defined alternatives. The first of these tasks ($n = 242$, 287, and 90) was formulated as shown in table 5:

Table 5 Responses to S1. Lying in a situation in which doing so would bring about the best consequences is . . .

	Study: 1	2	3
. . . always morally right.	2.8%	12.6%	2.2%
. . . always morally wrong.	6.6%	0.7%	10.0%
. . . sometimes right to some degree, but also wrong to some degree.	89.3%	75.0%	58.9%
. . . always either right or wrong but sometimes more importantly right or more seriously wrong.	NA	5.8%	NA%
. . . either right or wrong, there is no middle position, but the truth of a moral judgment is relative to one's moral standard or cultural norms.	NA	NA	17.8%
. . . sometimes right and sometimes wrong, but never a bit right and a bit wrong.	NA	NA	6.7%
. . . not possible to assess.	1.2%	5.8%	NA
. . . I don't know if this is right or wrong.	NA	NA	4.4%

The gradualist analysis was the most frequently selected answer option in all three studies. The differences in gradualist responses in Study 1 (88.2 percent), Study 2 (75.0 percent), and Study 3 (58.9 percent) can be explained by the fact that respondents were presented with four answer options in Study 1, five in Study 2, and six in Study 3. In Study 3 about 17.8 percent chose the relativist answer option, which was not available in the other studies. The more options respondents have to choose from, the less likely is it that everyone selects the same option. If right and wrong had been binary concepts many subjects could arguably have been expected to favor the Kantian answer option ('always morally wrong') or the utilitarian answer option ('always morally right').

Study 2 included an answer option designed to capture Hurka's notion of degrees, according to which some acts are more importantly right or more seriously wrong. This answer was selected by no more than 5.8 percent of respondents.

In Study 1, a separate group of respondents was presented with the answer option 'sometimes a bit right and a bit wrong' instead of 'sometimes right to some degree, but also wrong to some degree'. About 83.6 percent ($n = 116$) selected 'sometimes a bit right and a bit wrong', compared to 89.3 percent ($n = 244$) for the group presented with the option 'sometimes right to some degree, but also wrong to some degree'. The difference between 89.3 percent and 83.6 percent is not significant ($X^2 = 8.396$, the p-value is < 0.039; not significant at $p < 0.01$). This is an indication of robustness. The results reported here do not depend on minor alterations of the wording of the gradualist hypothesis.

Unsurprisingly, the vast majority also reported gradualist responses for the second semi-abstract item S2 ($n = 119$, 288, and 90), as shown in table 6:

Table 6 Responses to S2. Exceeding the speed limit in an emergency is . . .

	Study: 1	2	3
. . . always morally right.	5.9%	4.5%	15.6%
. . . always morally wrong.	1.7%	7.9%	4.4%
. . . sometimes right to some degree, but also wrong to some degree.	88.2%	75.3%	48.9%
. . . always either right or wrong but sometimes more importantly right or more seriously wrong.	NA	6.9%	NA%
. . . either right or wrong, there is no middle position, but the truth of a moral judgment is relative to one's moral standard or cultural norms.	NA	NA	6.7%
. . . sometimes right and sometimes wrong, but never a bit right and a bit wrong.	NA	NA	23.3%
. . . not possible to assess.	4.2%	5.5%	NA
. . . I don't know if this is right or wrong.	NA	NA	1.1%

A possible explanation of why so many respondents in Study 3 selected the binary response 'sometimes right and sometimes wrong, but never a bit right and a bit wrong' (23.3 percent) or 'always morally right' (15.6 percent) could be that speeding, but not lying, is viewed as less morally problematic by experienced drivers. US citizens who voted in the 2016 election are on average older than college students and are thus more likely to drive.

5. Concrete Tasks

Study 1 included six concrete tasks in which respondents were invited to assess brief descriptions of particular acts. As these tasks were not designed to study respondents' views on moral relativism or Hurka's hypothesis, the number of answer options was limited to four. In Study 1 the following task was evaluated by all respondents ($n = 715$):

C1.   John lies to a future employer about his qualifications. Due to inadequate background checks his lies go undetected. He is offered, and accepts, a job he is not qualified for.

What John did was morally right.	1.1%
What John did was morally wrong.	91.3%
What John did was right to some degree, but also wrong to some degree.	6.0%
What John did cannot be assessed from a moral point of view.	1.6%

Although relatively little information is provided in the vignette, over ninety percent reported that John's act was wrong. The next item (C2, $n = 363$) serves as a reference point for what seems to be a case of someone doing something right:

C2.   Jared's colleague Bob struggles to understand a new task for work. Jared has no plans for the evening and volunteers to help Bob to get up to speed. Between 5 pm and 8 pm Jared helps Bob to figure out how to solve the new task.

What Jared did was morally right.	74.9%
What Jared did was morally wrong.	1.9%
What Jared did was right to some degree, but also wrong to some degree.	2.5%
What Jared did cannot be assessed from a moral point of view.	20.6%

The gradualist hypothesis states that *some* acts are somewhat right and somewhat wrong, not that *all* are. Therefore, the findings for C1 and C2 neither refute nor confirm the gradualist hypothesis. However, data for the following tasks, C3–C6, support it. For these items, the gradualist answer option 'What [Denise/the captain/Anna/Miriam] did was right to some degree, but also wrong to some degree' was the most frequently selected answer in all three studies (see .

C3. Denise is in severe pain. She asks her spouse Adam to drive her to the hospital for treatment. Although he knows that her condition is not life-threatening, Adam drives 15 miles above the speed limit to bring Denise to the hospital as fast as he can.

C4. An experienced airline captain flies through a volcanic ash cloud that cause the engines to malfunction. To prevent panic among the passengers, the captain decides to lie to the passengers: 'The airport at our destination is closed due to bad weather. We will land at a nearby airport. There is no danger'.

C5. Anna is proud of a new webpage she has designed for a client, and the client is also pleased with it. Anna asks her junior colleague Josh to share his opinion. Josh does not like the webpage, but because Anna and the client seem to like it Josh decides to lie: 'I think your new webpage looks great, congratulations!'

C6. Miriam is on her way to the airport. Due to severe congestion on the highway, she realizes that there is a risk she will miss her international flight. When the road finally clears she drives 15 miles above the speed limit to make up for the time lost because of the congestion.

Table 7 Relative frequencies for items C3 to C6. The four answers options were the same as in C1 and C2.

	C3 ($n = 329$)	C4 ($n = 368$)	C5 ($n = 353$)	C6 ($n = 386$)
Right	10.9%	14.4%	5.7%	3.6%
Wrong	13.6%	15.5%	23.5%	37.0%
Degree	69.3%	71.5%	62.9%	43.0%
Not assessable	6.1%	1.4%	8.0%	12.9%

Table 8 summarizes pair-wise chi-square tests for all combinations of C1–C6. The degrees of freedom for all comparisons is df = 3, so for $p < 0.01$ the critical chi-square value is 11.34, and for $p < 0.001$ it is 16.27. Note that *all* pair-wise comparisons are statistically significant. However, the chi-square values for C1 and C2 stand out: they are ten to one hundred times higher than the values for all other items. (See the dashed box in table 8.) From a statistical point of view, the explanation is that C3–C6 are items in which the gradualist answer option is the most popular one; therefore, C3—C6 have more in common with each other than with C1 and C2.

Table 8 Pairwise chi-square tests for all combinations of C1 – C6. For $p < 0.01$ the critical chi-square value is 11.34, and for $p < 0.001$ it is 16.27 (df = 3). Note that *all* comparisons are statistically significant at $p < 0.001$.

	C1	C2	C3	C4	C5
C2	18840.98				
C3	3645.44	1804.09			
C4	3168.97	2235.24	56.29		
C5	1830.94	3503.68	33.32	82.6	
C6	658.66	5462.24	218.55	330.83	105.72

It is also worth pointing out that the chi-square values for C6 are four to ten times higher than the corresponding values for C3, C4, and C5. What could explain this? Note that in items C3, C4, and C5 some widely accepted norm is violated for a good reason: By violating the speed limit, or by lying, the agent brings about good consequences for others. A large majority reported that such norm violations are somewhat right and somewhat wrong. However, in C6 the agent violates a norm for what appears to be a selfish reason. Fewer subjects considered this to be

somewhat right and somewhat wrong, and about twice as many considered it to be wrong.

Combining the findings for the abstract, semi-abstract, and concrete tasks I found that only 4 percent in the largest study (Study 1, n = 715) persistently used right and wrong as non-gradable terms.

6. Comparative Tasks

Study 1 included a fourth type of task designed to test the gradualist hypothesis *without* including gradualist terms such as 'degree' and 'somewhat right' in the vignettes. If gradualist terms appear in questions or answer options, respondents might be more willing to apply such terms than they otherwise would (see Messick and Jackson 1961; see also the discussion below.)

Subjects were asked to make pairwise comparisons between items C1–C6 and a seventh item, C7:

> C7. After graduation Zofia decides to do unpaid volunteer work for Engineers without Borders for a couple of months before joining Petersen Consulting in Dallas, TX.

The comparative task was formulated as follows:

> Assess the acts performed by the agents. How similar are the moral properties of the two acts? (If one is right and the other is wrong they are not very similar, but if both are right, or both are wrong, they are very similar.)
> [C 1]
> [C 2]
> [Seven-point Likert scale, ranging from 'very similar' to 'very dissimilar'.]

Pairwise comparisons of seven items, C1–C7, requires 21 comparisons. Each respondent was invited to make four comparisons, which yielded 2830 comparative data points (*n* = 98 to 141). To verify that subjects understood the comparative task correctly, three identical comparisons were included in the questionnaire, C1–C1, C6–C6, and C7–C7. The average dissimilarity reported for these items were, 0.1, 0.2, and 0.3, which indicates a good understanding of the task. Table 9 summarizes the results:

Dissimilarities can be interpreted as distances in an *n*-dimensional geometric space. The more dissimilar two items are, the farther apart is their location in space. The twenty-one comparisons listed in table 9 can range over twenty dimensions, but by applying multidimensional scaling techniques, the dimensionality of a multidimensional dataset can be reduced (see Kruskal and Wish 1978).

Table 9 Average degree of dissimilarity, ranging from 0 (very similar) to 6 (very dissimilar).

	C1	C2	C3	C4	C5	C6	C7
C1	0.1						
C2	5.5	N/A					
C3	3.6	3.8	N/A				
C4	3.6	4.1	3.0	N/A			
C5	3.4	4.3	3.0	2.6	N/A		
C6	3.7	5.0	2.2	3.7	3.2	0.2	
C7	5.7	1.5	4.2	4.6	4.5	5.2	0.3

The aim of a classical multidimensional scaling is to represent the original data set by a new set of points in a smaller number of dimensions such that the Euclidean distance between each pair of points in the new set approximates the distance in the original multidimensional data set. Ideally, each pairwise distance (similarity) in the original data set (table 9) should be exactly the same as the corresponding Euclidean distance in the new representation. However, as we reduce the number of dimensions some minor errors will typically be introduced into the new representation. This is acceptable as long as the errors are small.

Figure 1 shows a classical multidimensional scaling of table 9. The maximum error is 0.36 units, which is a relatively large error (see below). When interpreting figure 1, it is important to keep in mind that item C1 is almost unanimously (91.3 percent) considered to be an example of wrongdoing, whereas C2 is widely considered (74.9 percent) to be an example of an agent doing something right. It is thus not surprising that C1 and C2 are located far apart along the x-axis. C7 can also be taken to be an example of someone doing something right, so it is equally unsurprising that C7 is located close to C2. The locations of C1, C2, and C7 can thus be explained *without* invoking the hypothesis that rightness and wrongness come in degrees.

However, the locations of C3–C6 cannot be easily explained without assuming that rightness and wrongness come in degrees. Although these items are located *somewhat* to the left in figure 1, they are not close to C1 along the *x*-axis. Moreover, recall that C3–C6 are items in which it is widely believed that the agent's act was somewhat right and somewhat wrong (69.3 percent for C3; 71.5 percent for C4; 62.9 percent for C5; and 43.0 percent for C6). This fits well with their locations in figure 1: items C3–C6 are, literally speaking, located 'between' the entirely wrong act in item C1 and the entirely right acts in C2 and C7.

If the binary hypothesis had been true, it would have been possible to represent all seven items along a single dimension, and all items would have been clustered in two distinct areas in the figure: right and wrong. However, the findings in table 9 cannot be represented in such a way and are thus incompatible with the binary hypothesis.

That said, it remains to explain why C6 and C3 in figure 1 are located above C4 and C5 on the *y*-axis. Note that C3 and C6 are cases in which the agent violates the speed limit for a good reason, whereas C4 and C5 are cases in which the agent lies for what seems to be a good reason. This suggests the following somewhat speculative interpretation: The more similar the agent's *reason* for some acts are, the closer
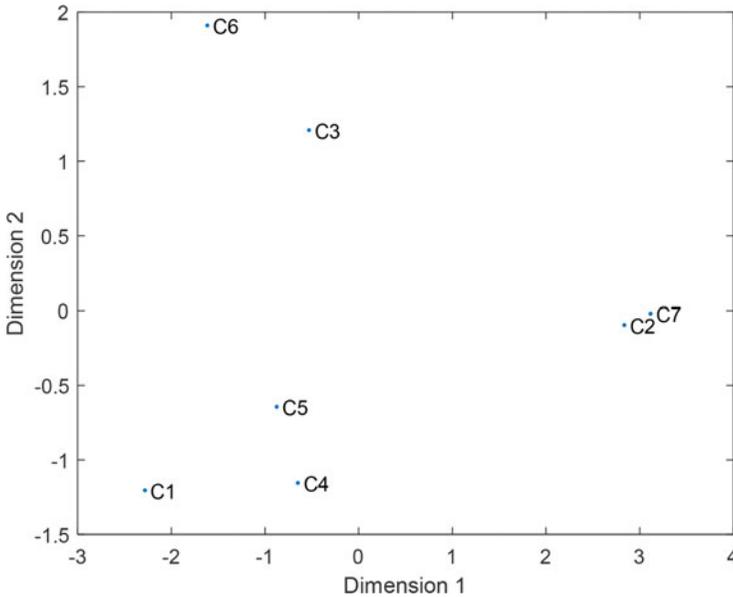
Figure 1. A classic multidimensional scaling of table 9. The maximum error is 0.36 units. Recall from above that a vast majority (91.3 percent) thought the agent acted wrongly in C1, whereas the majority (74.9 percent) thought the agent acted rightly in C2. For the options in between the most popular answer was that the agent acted somewhat right and somewhat wrong: 69.3 percent for C3; 71.5 percent for C4; 62.9 percent for C5; and 43.0 percent for C6. No data are available for C7.

are their locations on the y-axis. If so, the underlying similarities between C3 and C6, and C4 and C5, seem to be visible in the figure, which indicate that data in table 9 have a reasonable degree of validity.

A drawback of the two-dimensional scaling is that the maximum error in figure 1 is, as noted, relatively large. It is therefore appropriate to consider a three-dimensional scaling. See figure 2. In this representation, Kruskal's stress value is 1.07×10^{-6}, which indicates a good fit. Figure 2 confirms the conclusion of figure 1: C1 (the entirely wrong act) is located far apart from C2 and C7 (the entirely right acts), but C3–C6 are located *between* the entirely right and entirely wrong items. In figure 3, it is thus also reasonable to interpret the x-axis as a visual representation of an act's degree of rightness. The interpretation of the y-axis is the same as in figure 1, but we can leave it open how the z-axis is to be interpreted as that is of no importance of the present discussion.

Figures 1 and 2 are based on the assumption that all similarities in table 9 can be represented in a metric space. Because it is hard to know if this assumption is true, nonmetric multidimensional scalings are also worth considering. In this type of representation distances are interpreted as ordinal orderings: the aim is to preserve the *ordering* among the original points in a lower number of dimensions. Figure 3 shows a two-dimensional nonmetric multidimensional scaling of data in table 9. Kruskal's stress value is 8.5×10^{-5}, which indicates a good fit. This figure confirms the previous conclusions: C3–C6 is located *between* C1 and C2 and C7, which
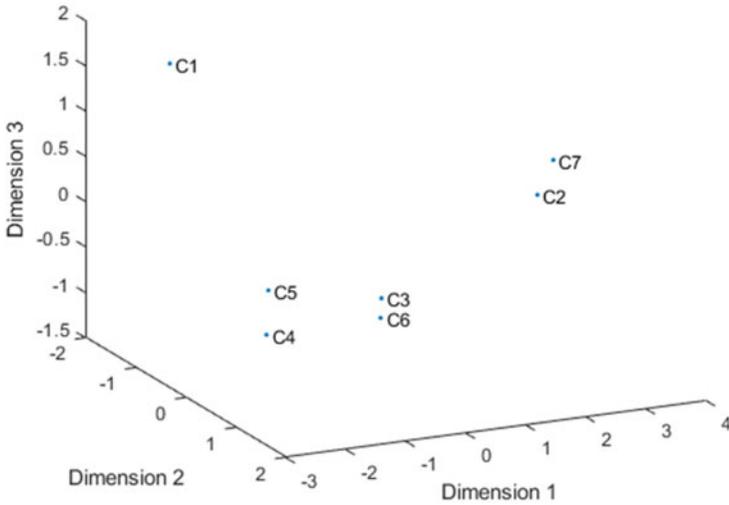
Figure 2. A three-dimensional metric multidimensional scaling of table 9 Kruskal's stress value is 1.07×10^{-6}. Note that C3–C6 are located 'between' C1 (the entirely wrong act) and C2 and C7 (the entirely right act).
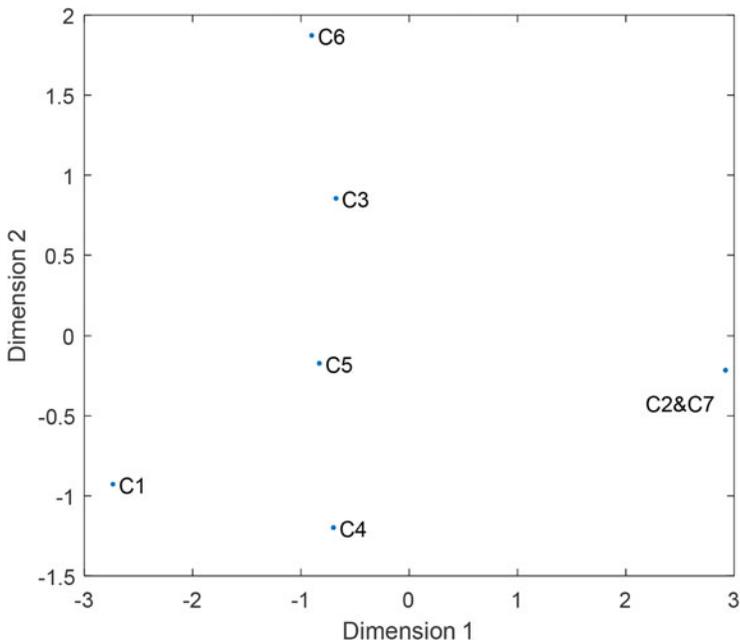
Figure 3. A two-dimensional nonmetric multidimensional scaling of table 9. Kruskal's stress value is 8.5×10^{-5}. Note that C3–C6 are located 'between' C1 (the entirely wrong act) and C2 and C7 (the entirely right act).

tallies well with the conclusion that the act C1 is entirely wrong, while the acts C3–C6 somewhat right and somewhat wrong, and C2 and C7 entirely right.

In summary, all three multidimensional representations fit well with the gradualist hypothesis, but they are incompatible with the binary one.

7. A Socratic Midwifery Effect?

Study 3 included an open-ended task in which half of the respondents were invited to write a couple of sentences about a case without relying on any predefined answer options. The other half was invited to select one of the following answer options: 'right', 'wrong', 'right to some degree, but also wrong to some degree', and 'I don't know, I would need more information'. The purpose was to study to what extent gradualist answers occur naturally. The study was conducted in May 2021, during the COVID-19 pandemic. Below is the case:

> Anouska's 80-year-old father is reluctant to take the COVID vaccine because he is concerned about possible side effects. She explains to him that it is primarily younger women who have been affected by severe side effects, so for him the benefit would definitely exceed the risk. In Anouska's opinion, it would be irrational of her father to not take the vaccine. He eventually gives in and schedules an appointment, but only after Anouska pressures him to do so. She also deceives him by exaggerating the benefits and minimizes some of less severe side effects. Evaluate Anouska's behavior from a moral point view. What is your moral conclusion all things considered?

The group asked to answer the question by writing a couple of sentences without relying on any predefined answer options ($n = 87$) submitted 3,453 words. All responses were analyzed and categorized manually. Gradualist conclusions were expressed by 16.1 percent of respondents, compared to 46.5 percent ($n = 95$) in the group presented with predefined answer options. (In the first group 29.0 percent reported that it was right to pressure Anouska's father to take the vaccine, compared to 16.1 percent in the second group; 26.0 percent concluded that it was wrong to pressure Anouska's father to take the vaccine, compared to 33.3 percent in the second group; and 29.0 percent presented with open-ended tasks stated views that were ambiguous or could not be reasonably classified as all-things-considered verdicts, compared to 4 percent in the second group who selected 'I don't know, I would need more information'.) Below are some examples of spontaneously submitted gradualist answers from respondents in the first group:

- 'What she did was morally grey, but the end result is positive'.
- 'From a moral standpoint, it would probably be both right and wrong, but erring toward wrong'.
- 'It's a little right and a little wrong'.
- 'It is a bit right and wrong to do what she did'.

It is not surprising that respondents used gradualist terms spontaneously, but it is surprising that gradualist responses occur almost three times as often (46.5 percent versus 16.1 percent) if a gradualist conclusion is included among a set of predefined answer options. This difference is significant at $p > 0.01$, one-tailed. What explains the difference?

One explanation could be that respondents' willingness to express gradualist views might be stimulated by a bit of philosophical midwifery. Socrates taught us that philosophers can help people articulate ideas they are unable to express clearly themselves, so I call this the Socratic midwifery effect. According to this explanation, respondents have a tacit but somewhat underdeveloped understanding of the gradualist hypothesis, which is clarified by the presence of a clearly stated gradualist answer option and therefore selected to a higher extent.

However, the findings reported in this paper do not permit us to state any definitive conclusion about the possible significance of a Socratic midwifery effect. About half of the respondents in Study 3 ($n = 95$) were presented with tasks C1, C3, and C4 described above. The rest ($n = 87$) were presented with versions of these tasks in which the original answer options ('a bit right and a bit wrong', and so on) had been replaced with more complex 'midwifery' answer options: 'There are reasons for, but also against, doing what [the agent] did. On balance it was neither entirely right nor entirely wrong; it was a bit right but also a bit wrong / it was right / it was wrong / I don't know'. For task C1, the percentage of gradualist answers increased from 10.5 percent to 27.6 percent (which supports the Socratic midwifery effect), but for C3, the percentage of gradualist answers decreased from 49.5 percent to 12.6 percent, and for C4 it decreased from 55.8 percent to 19.5 percent. This speaks against the Socratic midwifery effect and there is no obvious explanation of this anomaly. The significance of the Socratic midwifery effect could be a topic for future research.

8. Discussion

The empirical findings favor the gradualist hypothesis over its binary rival. The abstract, semi-abstract, concrete, and comparative measures indicate that if the way in which people use words and phrases is a reliable guide to meaning, then rightness and wrongness come in degrees. However, no empirical study is immune to criticism.

First, one might worry that the empirical findings may not uniquely support the gradualist hypothesis. Other hypotheses might be equally well supported by data. Consider, for instance, the suggestion that although no act is a bit right *and* a bit wrong, some right acts are right to a greater degree than others, and some wrong acts are more wrong than others. According to this alternative hypothesis, right and wrong function much like hot and cold: some cold objects are colder than other cold objects, but no cold object is a bit cold *and* a bit hot. Data for the comparisons reported above do not discriminate between this alternative hypothesis and the gradualist hypothesis. All we can conclude from the observation that some items (C3–C6) are located 'between' an act judged to be entirely wrong (C1) and others judged to be entirely right (C2 and C7) in a multidimensional scaling is that

traditional binary analyses offer a poor fit. The hypothesis that right and wrong function like hot and cold is, however, compatible with this result.

In response to this, the experimentalist may point out that the alternative hypothesis is not compatible with some of the other findings. Note, for instance, that the gradualist answer option 'right to some degree, but also wrong to some degree' was the most common response for items C3–C6—that is, the items located 'between' acts judged to be entirely right and wrong discussed above. This is evidence against the hypothesis that right and wrong function like hot and cold. If that hypothesis had been true, respondents would not have selected 'right to some degree, but also wrong to some degree'.

That said, it is of course possible to formulate other hypotheses that might fit better with data. Suppose, for instance, that respondents believe that acts are complex wholes composed of multivalent parts or aspects. According to this hypothesis, an act can be right with respect to one of its parts or aspects (for example, respect for autonomy) but wrong with respect to some other part or aspect (for example, fairness), but still be right in the binary sense all things considered. If so, respondents who responded 'right to some degree, but also wrong to some degree' may mistakenly have selected a gradualist phrase for expressing the view that an act is wrong with respect to some but not all of its parts or aspects.

A drawback of this hypothesis is that it offers a poor fit with some of the other findings. Consider, for instance, the responses to item S1 (table 5). The vast majority (83.6 percent in Study 1) reported that lying in a situation in which doing so would bring about the best consequences is 'sometimes right to some degree, but also wrong to some degree'. If respondents had believed that acts are complex wholes composed of different parts or aspects, they would arguably have selected 'I don't know' or 'not possible to assess' to a higher extent than they did, as it would have been unclear if the statement referred to the entire act or some of its parts or aspects. However, those answers were selected by no more than 4.4 percent and 5.8 percent. This indicates that a sophisticated philosophical distinction between wholes and parts does not seem to offer a better explanation of data. Another finding that does not fit well with the whole-part hypothesis is the observation that respondents agreed with the statement that 'moral rightness comes in degrees' to roughly the same extent that 'colors come in degrees' (see above). This is also difficult to reconcile with the part-whole hypothesis if we believe that colors come in degrees in an outright sense.

It is also worth keeping in mind that *all* theories are underdetermined by data, as pointed out by Willard Van Orman Quine (1975) and others. No matter what evidence one gathers for the gradualist hypotheses, it will always be possible to imagine some alternative hypothesis that fits equally well with the experimental findings. We will never be able to prove with certainty that our preferred hypothesis is the uniquely best one. The modest conclusion of this article should therefore be that the findings reported here offer a better fit with the gradualist hypothesis than do any of the alternative hypotheses discussed so far.

This brings me to what is perhaps the most important worry: Is the fact that ordinary people seem to use right and wrong in a sense that permit for degrees a good reason for revising traditional, binary moral theories? If traditional binary

moral theories are meant to capture the same concept we use in everyday moral discussions, and meaning tracks use, then the answer is yes. As noted, defenders of the meaning-tracks-use argument do not have to insist that meaning *is* use in a strict semantic sense. The key premise is just that use is a reliable indicator of meaning. The observation that people use right and wrong as gradable terms *shows that* these terms are gradable, regardless of whether meaning *is* use.

A possible response from binary theorists, such as classic act-utilitarians and some Kantians, could be that the notions of right and wrong described in those theories are technical concepts, just like nearly all central concepts used in scientific theories. It is beyond the scope of this article to discuss this objection at length, but note that the analogy with technical concepts in science might be problematic because scientists use technical concepts for what seems to be a good reason. For instance, the technical concept of heat used in physics enables scientists to express nuanced claims about physical processes that cannot be expressed by the everyday concept. But the binary concepts of right and wrong are *less* nuanced than their gradualist counterparts. What would the point be of introducing technical notions of right and wrong that are less nuanced than our ordinary concepts? The binary concepts of right and wrong have no additional explanatory power compared to the gradualist ones. The gradualist folk notions of right and wrong allow us to articulate more precise moral verdicts than the blunt, binary concepts employed by traditional moral theories. By asserting that an act is somewhat right and somewhat wrong we can express nuances that get lost if we adopt a binary theory that forces us to conclude that an act is either right or wrong in the binary sense.

Another reason for dismissing the empirical findings as irrelevant could be that it may make no practical difference if the gradualist hypothesis is true or false. Agents always have to choose what to do, and according to this argument choices are ultimately binary. From a deliberative point of view, it would therefore be irrelevant if the gradualist hypothesis is correct. This objection is, however, too quick. In an excellent discussion of moral vagueness, J. Robert Williams (2017) distinguishes between what he calls 'moral oughts' and 'decision oughts'. (A similar distinction is sometimes made in discussions of moral uncertainty. See, for instance, Lockhart 2000: ch. 1). The term 'decision ought' is a technical term that refers to what a rational and morally conscientious agent would have most instrumental reason to do if she is motivated entirely by a desire to act in accordance with what morality demands of her. Everyone agrees that if all alternatives open to an agent are entirely right or wrong, then the agent is rationally permitted to perform any of the right acts, but none of the wrong ones. But what about acts that are somewhat right and somewhat wrong? Are we ever rationally permitted to perform such acts? There is no consensus on this, but here are some possible answers: (1) It is rationally permissible to perform every act *a* that is not entirely wrong (see Williams 2014, 2017; Bales 2018). (2) It is rationally permissible to perform act *a* if and only if no alternative act is right to a greater degree. (3) It is rationally permissible to perform act *a* if and only if *a* is not more wrong than right. (4) If act *a* is neither entirely right nor entirely wrong, then there is no fact of the matter as to whether it is rationally permissible to

perform *a* (see Rinard 2015). (5) If act *a* is neither entirely right nor entirely wrong, then it is rationally permissible to perform act *a* if and only if *a* is chosen randomly with a probability that reflects *a*'s degree of rightness (see Peterson 2013; Williams 2017).

It is beyond the scope of my purpose here to take a stand on which theory of decision-making defenders of the gradualist hypothesis ought to adopt. However, by considering the five views outlined it becomes clear that it does make a practical difference whether we accept the gradualist hypothesis. An agent who, for instance, believes that some act *a* is more right than wrong, but not entirely right, may reasonably disagree with a binary theorist who insists that *a* is right in a binary sense about whether it would be rationally permissible to perform *a*. This shows that it *does* matter whether the empirical evidence for the gradualist hypothesis can convince utilitarians, Kantians, and others to abandon traditional binary moral theories.

MARTIN PETERSON
TEXAS A&M UNIVERSITY
martinpeterson@tamu.edu

References

Bales, Adam. (2018) 'Indeterminate Permissibility and Choiceworthy Options'. *Philosophical Studies*, 175, 1693–1702.

Brown, Campbell. (2016) 'The Rightest Theory of Degrees of Rightness'. *Ethical Theory and Moral Practice*, 19, 21–29.

Erk, Katrin. (2012). 'Vector Space Models of Word Meaning and Phrase Meaning: A Survey'. *Language and Linguistics Compass*, 6, 635–53.

Hurka, Thomas. (2019) 'More Seriously Wrong, More Importantly Right'. *Journal of the American Philosophical Association*, 5, 41–58.

Kruskal, Joseph B., and Myron Wish. (1978) *Multidimensional Scaling*. Newbury Park: Sage.

Lawlor, Rob. (2009). *Shades of Goodness. Gradability, Demandingness and the Structure of Moral Theories*. Basingstoke: Palgrave Macan.

Lockhart, Ted. (2000). *Moral Uncertainty and Its Consequences*. New York: Oxford University Press.

Messick, Samuel, and Douglas N. Jackson. (1961). 'Acquiescence and the Factorial Interpretation of the MMPI'. *Psychological Bulletin*, 58, 299–304.

Mill, J. S. (1863). *Utilitarianism*. London: Parker, Son and Bourn.

Peterson, Martin. (2013) *The Dimensions of Consequentialism*. Cambridge University Press.

Quine, Willard van Orman. (1975) 'On Empirically Equivalent Systems of the World', *Erkenntnis*, 9, 313–28.

Rinard, Susanna. (2015). 'A Decision Theory for Imprecise Credences'. *Philosopher's Imprint*, 15. http://hdl.handle.net/2027/spo.3521354.0015.007.

Schoenfield, Miriam. (2016) 'Moral Vagueness Is Ontic Vagueness'. *Ethics*, 126, 257–82.

Turney, Peter D., and Patrick Pantel. (2010) 'From Frequency to Meaning: Vector Space Models of Semantics'. *Journal of Artificial Intelligence Research*, 37, 141–88.

Williams, J. Robert G. (2014) 'Decision-Making under Indeterminacy', *Philosophers' Imprint* 14. http://hdl.handle.net/2027/spo.3521354.0014.004.

Williams, J. Robert G. (2017) 'Indeterminate Oughts', *Ethics* 127, 645–73.

Wittgenstein, Ludwig. [1953] (2009) *Philosophical Investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte. Edited by P. M. S. Hacker and Joachim Schulte. 4th rev. ed. Malden: Wiley-Blackwell.