The Vagueness of Robot Emotions

Phil Serchuk[1,2] **Ehud Sharlin**[2] **Martin Lukac**[3] **Marek Perkowski**[3]
[1]Department of Philosophy, University of Calgary, pserchuk@ucalgary.ca
[2]Department of Computer Science, University of Calgary, ehud@cpsc.ucalgary.ca
[3]Department of Electrical and Computer Engineering, Portland State University, mperkows@ee.pdx.edu, lukacm@cecs.pdx.edu

ABSTRACT
As the field of robotics matures, robots will need some method of displaying and modeling emotions. One way of doing this is to use a human-like face on which the robot can make facial expressions that correspond to its internal emotional state. Yet the connection between a robot's emotional state and its physical facial expression is not an obvious one; in particular, there is no principled method of mapping changes in emotional states to changes in facial expressions. We give a philosophical analysis of the problem and show that it is rooted in the vagueness of robot emotions. We then outline several methods that have been used in the philosophical literature to model vagueness and propose an experiment that uses our humanoid robot head to determine which philosophical theory is best suited to the task.

INTRODUCTION
It has been argued that the ability to display emotions on a human-like face is both an important and necessary step in making robots and computer agents more accessible to the general public [1, 7]. The emotional model for a robotic face will have two key components: a mapping of emotional states to facial positions and a method of transitioning between different pairs of emotional states and facial positions. For example, we map 'surprise' on our robot to a raising of its eyebrows and an opening of its mouth. We say the robot is 'not surprised' when its eyebrows fall and its mouth closes. Some method is now needed of transitioning between 'surprised' and 'not surprised' that both displays the transition on the robot's face and captures the emotional state of the robot throughout the transition.

We begin by describing the philosophical problem of vagueness and show how it is relevant to the domain of robot emotions; in particular, we argue, it is the vagueness of robot emotions that prevents us from finding a principled method of selecting a particular facial position or gesture to use as the boundary between two different emotional states. We then outline an experiment designed to test the validity of our approach, and give a detailed description of our interface—a humanoid robot head.

A THEORETICAL ACCOUNT OF VAGUENESS IN ROBOT EMOTIONS
Suppose we want to build an advanced robotic office assistant that, among other things, delivers mail to office employees. This hypothetical robot will have a face capable of displaying emotions to its users. For the remainder of this paper, we denote emotional states by a boolean variable S_α, whose value is returned by the function t. We use S_h to denote the emotional state 'happy' and $\neg S_h$ to denote the emotional state 'not happy'. Thus, the robot is 'happy' if and only if $t(S_h) = 1$; the robot is 'not happy' if and only if $t(S_h) = 0$. We write $P(\hat{v})$ to denote the robot's facial gesture, or position, as given by a vector $\hat{v}$; that is, the robot is said to be in position $P(\hat{v})$ when its servos are set to the values prescribed by the value of $\hat{v}$. To denote the robot's emotional state α, given a facial gesture or position expressed as a vector $\hat{v}$, we use $S_{\alpha_{\hat{v}}}$. We write $|\hat{v}| = i$ to express $\hat{v}$ in terms of i, which is some complex value generated by the positions of the robot's many servos.

We define four facial positions: $P_{\hat{v1}}$ and $P_{\hat{v2}}$ show the robot as being clearly 'happy', where $|\hat{v1}| = 100$ and $|\hat{v2}| = 85$; $P_{\hat{v4}}$ and $P_{\hat{v5}}$ show the robot as being clearly 'not happy', where $|\hat{v4}| = 15$ and $|\hat{v5}| = 0$. We also define $\hat{v3}$, such that $|\hat{v3}| = 50$. For example, the robot might be smiling widely in $P(\hat{v1})$ and slightly less so in $P(\hat{v2})$; however, when in $P(\hat{v4})$, the robot might frown slightly.

Consider a vector $\hat{v6}$, where $|\hat{v6}| = n$. When $n = 100$ $\hat{v3} = \hat{v6}$, and $t(S_{h_{\hat{v6}}}) = 1$. Each time the robot successfully completes a task we increase n to make the robot look happier, and each time the robot fails to complete a task we decrease n to make the robot look sadder. For example, if the robot spills coffee on a human we might set $n = n - 70$, but if the robot delivers an envelope to the wrong employee we might only set $n = n - 1$.

The Boundary Between S_h and $\neg S_h$

The fundamental problem with this Boolean model is that we cannot determine the n that marks the location where the robot changes from state S_h to state $\neg S_h$. Suppose the face is in position $P_{\hat{v3}}$. We are just as justified in holding that position $P_{\hat{v3}}$ corresponds to state $\neg S_h$ as we are in holding that it corresponds to state S_h. We do not want to arbitrarily assign $P_{\hat{v3}}$ to S_h or $\neg S_h$, as to hold that position $P_{\hat{v3}}$ is the precise cut-off between emotional states S_h and $\neg S_h$ is to hold that while some large number of negative events may occur to move the robot's face from position $P_{\hat{v1}}$ to $P_{\hat{v3}}$, the robot's emotional state remains constant at S_h; yet if one small, seemingly insignificant event causes the robot's face to move past position $P_{\hat{v3}}$ and toward $P_{\hat{v4}}$, the robot's emotional state will suddenly change from S_h to $\neg S_h$. This is clearly counterintuitive as, among healthy humans, small, insignificant actions do not cause sudden changes in emotional states.

Suppose we use $P_{\hat{v3}}$ as the boundary between S_h and $\neg S_h$. Recall that the face of the office robot is in position $P_{\hat{v6}}$ and suppose that, at a given point during the day, $n = 85$: since, in this case, $\hat{v6} = \hat{v2}$, we know the robot is 'happy'. If the robot spills coffee on an employee then $n = 85 - 70 = 15$ and $\hat{v6} = \hat{v4}$; thus, the robot becomes 'not happy'. This result conforms to our intuition that if a person is happy and she makes a big mistake, such as spilling coffee on a co-worker, she will become unhappy.

Now suppose $n = 85$ and the robot brings a piece of mail to the wrong employee: $n = 85 - 1 = 84$. This result also conforms to our intuitions: if a person is happy and makes a small mistake, such as bringing a piece of mail to the wrong person, they don't stop being happy. Now suppose this occurs 36 times: first $n = 84 - 1 = 83$, then $n = 83 - 1 = 82$, and so on, until $n = 50 - 1 = 49$. When the robot incorrectly delivers the mail for the 35^{th} time it is still 'happy'; but because $n = 50$, its smile will have almost completely disappeared. However, once the robot delivers mail to the wrong person for the 36^{th} time, its mood will immediately change from 'happy' to 'not happy', even though its physical appearance — the diminished smile that is visible when $n = 50$ — has changed by an imperceptibly small amount, namely by one unit, such that $n = 50-1 = 49$. The problem is that while the physical transition between the appearance of 'happy' and the appearance of 'not happy' is gradual, as it is with humans, the transition between S_h and $\neg S_h$ is not: despite its physical appearance, the robot will believe it is as happy when $n = 51$ as it is when $n = 85$. When n is decremented by one unit the change in facial position will be hardly noticeable, yet its mood instantaneously changes from 'happy' to 'sad': from S_h to $\neg S_h$.

The Logic

The problem in the above example lies in our belief that one insignificant event does not make the difference between emotional states. The linguistic vagueness that philosophers typically study works similarly: words like 'bald' are said to be vague because one hair does not seem to make the difference between baldness and non-baldness. Let $S_{h_{\hat{v6}_n}}$ indicate that the robot is in state S_h and in position $P_{\hat{v6}}$, given the value of n in $\hat{v6}$. Thus, $t(S_{h_{\hat{v6}_{100}}}) = 1$, since when $n = 100$, $\hat{v6} = \hat{v1}$. In boolean logic, we symbolize our belief that one insignificant event does not make the difference between S_h and $\neg S_h$ as follows: $\forall m(S_{h_{\hat{v6}_m}} \supset S_{h_{\hat{v6}_{m-1}}})$. Yet given this, we can prove that there is no n that marks the cut-off between S_h and $\neg S_h$ (see Figure 1).

1	$S_{h_{\hat{v6}_{100}}}$	
2	$\forall m(S_{h_{\hat{v6}_m}} \supset S_{h_{\hat{v6}_{m-1}}})$	
3	$S_{h_{\hat{v6}_{100}}} \supset S_{h_{\hat{v6}_{99}}}$	$\forall$E, 2
4	$S_{h_{\hat{v6}_{99}}}$	$\Rightarrow$E, 1, 3
$\vdots$	$\vdots$	
$2,000$	$S_{h_{\hat{v6}_1}}$	$\Rightarrow$E, $\vdots$
$2,001$	$S_{h_{\hat{v6}_1}} \supset S_{h_{\hat{v6}_0}}$	$\forall$E, 2
$2,002$	$S_{h_{\hat{v6}_0}}$	$\Rightarrow$E, $(2,000)$, $(2,001)$

Figure 1. Proof: if the robot is 'happy' when in $P_{\hat{v6}_{100}}$, then it is 'happy' when in $P_{\hat{v6}_0}$

Thus, having started with an obviously true premise, namely $S_{h_{\hat{v6}_{100}}}$, we can conclude that when $m = 0$ the robot is still 'happy'; $t(S_{h_{\hat{v6}_0}}) = 1$. We could construct a similar argument, starting with $\neg S_{h_{\hat{v6}_0}}$ and assuming $\forall m(\neg S_{h_{\hat{v6}_m}} \supset \neg S_{h_{\hat{v6}_{m+1}}})$, to show that when $m = 100$ the robot is still 'not happy': $t(\neg S_{h_{\hat{v6}_{100}}}) = 1$. We have thus been able to prove two seemingly contradictory facts. We proved both that the robot is 'happy' when $m = 0$ and its facial expression is 'not happy', and that the robot is 'not happy' when $m = 100$ and its facial expression is 'happy'. While boolean logic requires that we find some vector, and thus some value of m, to use as the boundary between S_h and $\neg S_h$, philosophers have developed several theories of vagueness that we use to develop a framework for the proper modeling of robot emotions.

PHILOSOPHICAL THEORIES

Typically, philosophical theories of vagueness give both a metaphysical account of the phenomenon of vagueness and an account of our linguistic use of vague predicates. While our work is focused on modeling emotions in practical applications, the metaphysical content of these theories should not be disregarded: as the field of human-robot interaction matures it is likely that scholarship will focus not only on the results achieved by a given method but on the correctness of using that method in the first place. In this section we present an overview of several theories of vagueness that seem particularly well-suited for the task of modeling robot emotions and controlling facial expressions; it is not our intent to give a detailed account of the philosophical literature. References for further reading are provided.

3-Valued Logics

For any variable P in boolean logic, either P is true or P is false. By introducing additional truth values, many-valued logics allow P to take other values. Many-valued logics are extensions of classical logic and always have 'true' and 'false' as truth values which behave, in relation to one another, as they would in classical logic [3, p. 5]. We use a Łukasiewicz 3-valued logic, which has the following truth values: $[0, \frac{1}{2}, 1]$. A truth value of $\frac{1}{2}$ represents an indeterminable truth value that is assigned to that which is possible and exists between 'the true' and 'the false'; that is, $\frac{1}{2}$ is truer than what is false but falser than what is true [8]. Given two variables P and Q, negation, conjunction, disjunction, and implication are defined thusly:

$t(\neg P) = 1 - t(P)$
$t(P \wedge Q) = min(t(P), t(Q))$
$t(P \vee Q) = max(t(P), t(Q))$
$t(P \rightarrow Q) = min(1, 1 - t(P) + t(Q))$

To model emotions in a 3-valued logic, we first determine which positions *clearly* correspond to S_h and which *clearly* correspond to $\neg S_h$. There are two ways that a 3-valued logic represents states that are neither *clearly* S_h nor *clearly* $\neg S_h$: on the *truth gap* theory, these borderline cases are neither S_h nor $\neg S_h$ while on the *truth glut* theory they are both S_h and $\neg S_h$. On both *truth gap* and *truth glut* theories we have the following truth assignments: $t(S_{h_{\hat{v6}_{100}}}) = 1$, $t(S_{h_{\hat{v6}_{50}}}) = \frac{1}{2}$, and $t(S_{h_{\hat{v6}_{0}}}) = 0$. Recall that in boolean logic, if the robot is not in state S_h, such that $t(S_h) = 1$, then the robot must be in state $\neg S_h$, such that $t(S_h) = 0$. On the 3-valued approach, however, a robot that is not in state S_h need not be in state $\neg S_h$: when $t(S_h) = \frac{1}{2}$ the robot is not in state S_h nor is it in state $\neg S_h$. For more information on 3-valued logics and vagueness, see [8],[9], and [10].

Epistemicism

Epistemicism [11] holds that words like 'tall' and 'bald' have sharp boundaries that are necessarily unknowable to us. If we extend this view to emotions like 'happy', we must hold that when the robot is in $P_{\hat{v}3}$ it is really either in state S_h or $\neg S_h$ — but we can never know which state it truly belongs in. Of course, we can program the robot to treat $P_{\hat{v}3}$ as though it corresponds to S_h, but this will yield the same sudden change in emotional state that we are trying to avoid. The strengths of this theory are primarily metaphysical, and the only practical advantage it offers is that it allows us to model vagueness, and emotions, in boolean logic.

To model emotions according to epistemic principles, we need to hold that there are precise facial expressions — some of which are unknown to us — that correspond to the robot appearing happy; all other facial expressions correspond to the robot looking unhappy. Thus, $P_{\hat{v}1}$ corresponds to the robot looking happy and $P_{\hat{v}5}$ corresponds to the robot looking unhappy. Yet according to the epistemicist, we have no way of knowing what the actual position of the cut-off is: we only know that it exists. Because epistemicism was developed for linguistic vagueness, many complications arise when attempting to model robot emotions using this theory. One possibility is to have the robot inform the user that it does not know whether it is happy or sad when in $P_{\hat{v}3}$; another is to have the robot stop expressing emotion altogether when it is in indeterminate positions like $P_{\hat{v}3}$ and continue expressing emotions when it returns to a position where its emotional state is clear, such as $P_{\hat{v}2}$.

Fuzzy Logic

Fuzzy logic has been proposed by philosophers and used by computer scientists to model linguistic vagueness; computer scientists have already developed several emotional models based on fuzzy logic [4, 5]. Fuzzy logic is an infinitely-valued logic, with truth values represented on the interval of real numbers $[0, 1]$. Variables, and in this case emotional states, are represented by fuzzy sets and objects in the domain are members of each set to varying degrees; the degree to which a particular variable belongs in the set **TRUE** is a variable's degree of truth. Negation, conjunction, and disjunction are defined as they are in 3-valued logic, and implication typically is as well, although other definitions are sometimes used. Thus, we initially know that $t(S_{h_{\hat{v6}_{100}}}) = 1$ and $t(S_{h_{\hat{v6}_{0}}}) = 0$. We map the other values of i to truth values using a membership function. Suppose the robot is *clearly* happy when $i > 80$ and *clearly* unhappy when $i < 30$. One possible membership function is:

if $i < 30$ then $t(S_{h_{\hat{v6}_{i}}}) = 0$
if $30 \leq i \leq 80$ then $t(S_{h_{\hat{v6}_{i}}}) = \frac{i-30}{50}$
if $i > 80$ then $t(S_{h_{\hat{v6}_{i}}}) = 1$

Using this function, the robot's emotional state changes along with its smile. When $i = 70$ the robot is 'happy' to degree 0.8 and 'not happy' to degree 0.2, $t(S_{h_{\hat{v6}_{70}}}) = 0.8$; when $i = 60$ the robot is 'happy' to degree 0.4 and 'not happy' to degree 0.6, $t(S_{h_{\hat{v6}_{60}}}) = 0.4$. More information on fuzzy logic and fuzzy set theory can be found in [2] and [6].

THE INTERFACE

Our interface is a robot head (see: Fig. 2) developed by the authors from Portland State University. Although we illustrated the vagueness of robot emotions in the previous section with 'happy', we found that the head worked much better when expressing surprise; thus, we chose to implement 'surprise' for the purposes of our experiment. This decision does not affect our theoretical grounding. Instead of an office assistant, suppose we want to build a robot that will spend time with the elderly. Among other things, this robot will watch movies along with a human companion. There are many different emotions that such a robot would need to be capable of expressing: it would need to get happy and sad at the right times, it would need to dislike 'bad' characters and empathize with 'good' ones, and it would need to get 'surprised' when something frightening or startling occurred.

Figure 2. Our Humanoid Interface

Only the head is robotic: the torso and hands are props used to enhance the robot's realism in our experiments. The robot head we are using consists of many parts, each of which is controlled by a servo. The robot consists of twelve servos, including:

- four which control the eyebrows, denoted as $Eb_1 \dots Eb_4$ (see: Fig. 3.1).
- four which control the face, denoted as $Fa_1 \dots Fa_4$ (see: Fig. 3.2).
- one which controls both eyes, denoted as Ey (see: Fig 3.3).
- one which controls the tongue, denoted as To (see: Fig 3.4).
- one which controls the mouth, denoted as Mo (see: Fig 3.5).
- one which controls its upper lip, denoted as Ul (see: Fig 3.6).

Each servo has a different amplitude of real movement which is scaled to a value between 0 and 100. We denote the scaled position of a servo α by $P(\alpha)$. For example, $P(Fa_2) = 50$ means that the second facial servo is in position 50. The following figure shows each of the robot's servos; when necessary, a dot has been placed to indicate a servo's position.

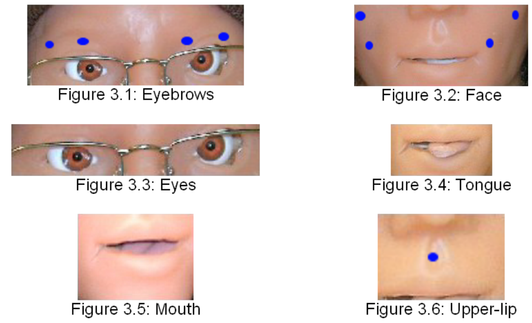

Figure 3. Servo Positions

I use scripts to control the robot's movement. In our experiment, scripts cause the robot to look 'surprised' while the user is watching the robot. Once the script finishes, the robot will hold its 'surprised' position for a predetermined period of time. Once this time has passed, the robot's face returns to its normal state. We denote the scaled position of a servo α prior to the running of a script as $P(\alpha_{(\text{old})})$; we use $P(\alpha_{(\text{new})})$ to denote the servo's position attained after a given script has completed but before the face has returned to normal. If no subscript is used, then $P(\alpha)$ denotes a servo's 'new' or current position.

To denote the level of 'surprise' expressed by a given script σ we use a vector $\hat{v}$. We let $\hat{v} = \{\alpha, \beta, \chi, \delta, \epsilon, \zeta, \eta\}$ and define $|\hat{v}| = \alpha * w_1 + \beta * w_2 + \chi * w_3 + \delta * w_4 + \epsilon * w_5 + \zeta * w_6 + \eta * w_7$. Note that while $w_1 \dots w_7$ are arbitrary weights, $\hat{v}$ nonetheless reflects the level of 'surprise' shown by our robot face.

- α reflects upward movement of the eyebrows, where $\alpha = \Sigma(P(Eb_{i(\text{new})}) - P(Eb_{i(\text{old})}))$, when $1 \leq i \leq 4$ and $P(Eb_{i(\text{new})}) > P(Eb_{i(\text{old})})$.
- β denotes the final position of the eyebrows, where $\beta = \Sigma P(Eb_i)$ where $1 \leq i \leq 4$.
- χ returns a fixed value if the eyes move a sufficiently large distance, if $|P(Ey_{(\text{new})}) - P(Ey_{(\text{old})})| > 5$.
- δ reflects the movement of the upper lip, where $\delta = P(Ul)$ if $P(Ul_{(\text{new})}) \neq P(Ul_{(\text{old})})$.
- ϵ reflects downward movement of the mouth, where $\epsilon = P(Mo_{(\text{new})}) - P(Mo_{(\text{old})})$, if $P(Mo_{(\text{new})}) > P(Mo_{(\text{old})})$.
- ζ denotes the final position of the mouth, where $\zeta = P(Mo)$.
- η reflects the amount of time that a script σ's final positions are held.

SELECTING A THEORY

Because we are not currently interested in the metaphysical claims of the theories described in the previous section, we must use some other criterion to determine which theory is best suited for mapping facial expressions to a robot's emotional state. Throughout this section we use a Boolean variable S_s, whose value is returned by the function t, to model the emotional state 'surprise'; we use $\neg S_s$ to denote the state 'not surprised'. Thus, the robot is 'surprised' if and only if $t(S_s) = 1$ and the robot is 'not surprised' if and only if $t(S_s) = 0$.

Each theory of vagueness treats positions that are neither clearly S_s nor clearly $\neg S_s$ differently: on the 3-valued approach such a position corresponds either to both S_s and $\neg S_s$ or to neither S_s nor $\neg S_s$, while on the fuzzy logic approach it partially corresponds to S_s and partially corresponds to $\neg S_s$. On the epistemic view it corresponds to either S_s or $\neg S_s$, but it is impossible for us to know which. We propose conducting experiments with a robot face to test user's perceptions of emotions that are not clearly 'surprised' nor clearly 'not surprised' to see how they perceive this borderline area.

Our experiment is being conducted in parallel with a separate experiment on vagueness in natural language. Participants will choose to participate in an experiment on natural language; it is only when they arrive to participate in the experiment that they will see the robot. At this point, participants will be told that they are doing the experiment with a co-participant—a robot. Both the human and the robot will be completing two tasks: one the same, one different. The human participant will be told that both participants will complete a questionnaire on their use of vague predicates in natural language, with the goal of generating data about the differences between robot and human vocabulary. The human will be told that the robot, as part of its second task, will be shown images at random intervals. The human's second task will be to describe the emotional state of the robot after the robot appears 'surprised' from the images it sees. The human participant will complete questions on vague predicates while thinking that the robot is doing the same thing.

Every few questions the participant will be notified, via their computer screen, that the robot is about to see an image. At this point the robot will be programmed to act startled. Once the reaction is complete, a screen will appear asking the participant to describe the robot's emotional state. Once the participant answers this question, they will continue answering questions about vague predicates. The user will be able to choose one of the following options when describing the robot's emotional state:

- surprised
- not surprised
- surprised and not surprised
- neither surprised nor not surprised
- either surprised or not surprised, but unsure of which
- partially surprised and partially not surprised

The first two options conform to a traditional boolean emotional model, while the last four are used to represent the theories of vagueness that were described in the previous section. We hypothesize that users will choose one of the first two options when the robot's state is obvious; when the state is difficult to determine, we believe users will choose one of the last four options. If a significant percentage of users choose one of the last four options when the robot is in an intermediate state, then we will have evidence indicating which theory of vagueness can be best used to control robot emotions and facial expressions. If the data shows that users associate intermediate positions with a given theory of vagueness, then we will use literature on that theory to develop a more detailed model of robot emotions.

CONCLUSION

In this paper we reported our efforts to model synthetic emotions in a humanoid robot head. We described the philosophical problem of vagueness and showed how it poses a practical problem for the implementation of emotions in Boolean logic. We presented three philosophical theories of vagueness that could be used to accurately model robot emotions and map changes in emotional states to physical gestures. Finally, we described a user study—that we are currently conducting—that will allow us to evaluate the viability of our approach and the effectiveness of these theories. We hope that this work will allow researchers in human-robot interaction to explore the strengths of non-Boolean logics in the domain of robot emotions.

ACKNOWLEDGMENTS

We'd like to thank Dr. Svetlana Yanuskevich from the University of Calgary Biometrics Lab, Andrew Read from the University of Calgary Science Workshop, and Brian Scowcroft from the University of Calgary Department of Computer Science.

REFERENCES

1. C. L. Breazeal. *Designing Sociable Robots*. MIT Press, 2002.
2. J. J. Buckley and E. Eslami. *An Introduction to Fuzzy Logic and Fuzzy Sets*. Physica-Verlag, 2002.
3. S. Gottwald. *A Treatise on Many-Valued Logics*. Studies in Logic and Computation. Research Studies Press Ltd., 2001.
4. L. C. Jain, H.-N. Teodorescu, and A. Kandel, editors. *Soft Computing in Human-Related Sciences*. CRC Press, 1999.
5. R. Langari and J. Yen. *Fuzzy Logic: Intelligence, Control, and Information*. Prentice-Hall Inc., 1999.
6. H. T. Nguyen and E. A. Walker. *A First Course in Fuzzy Logic*. CRC Press, 1997.
7. D. A. Norman. *Emotional Design*. Basic Books, 2004.
8. N. Rescher. *Many-valued Logic*. McGraw-Hill, 1969.
9. M. Tye. Vague objects. *Mind*, 99(396):535–57, 1990.
10. M. Tye. Vagueness: Welcome to the quicksand. In T. Horgan, editor, *Spindel Conference: Vagueness*, pages 1–22. University of Memphis, 1995.
11. T. Williamson. *Vagueness*. Routledge, 1994.