

*First International Conference of the International Association  
for Computing and Philosophy*

# Is Ethics Computable

Or, what other than *can* does *ought* imply?



*Presidential Address – Aarhus, Denmark – July 4<sup>th</sup>, 2011*



# Is Ethics Computable

Or, what other than *can* does *ought* imply?



*Everyone will readily agree that it is of the highest importance to know whether we are not duped by morality.*

*Emmanuel Levinas, 1961/1969, 21*



# Is Ethics Computable

## Or, what other than *can* does *ought* imply?



*As Daniel Dennett (2006) recently stated, AI 'makes philosophy honest.' Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas.*

*Anderson and Anderson, 2007, 16*

**Supposition: Computational ethics makes ethics honest.**

# Is Ethics Computable? – Talk Overview



- A Prefatory Story to Set the Stage
- Turing's Prediction and Its Moral Corollary
- Statement of the Problem
- On the Various Meanings of *Ought*
- To Be Perfectly Honest . . . With Ourselves



# A Reading from the *Genesis* of Twin Earth

For the narrative, see <http://faculty.evansville.edu/tb2/PDFs/Robot%20Genesis.pdf>



Wall-E and Eve™ Disney



UNIVERSITY  
OF  
EVANSVILLE  
Civic Mission... Sacred Trust



## Turing's Prediction and Its Moral Corollary



Alan Turing

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

*Turing, 1950, 442*



## Turing's Prediction and Its Moral Corollary



Alan Turing

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

*Turing, 1950, 442*

Will it be the case that by mid-century the use of words and general educated opinion will have altered so much that one will be able to speak of machines being moral without expecting to be contradicted?

## Turing's Prediction and Its Moral Corollary



Alan Turing

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

Will it be the case that by mid-century the use of words and general educated opinion will have altered so much that one will be able to speak of machines being moral without expecting to be contradicted?

**What is at stake?**



## Turing's Prediction and Its Moral Corollary

Things People Sometimes Meaningfully Say	Things People Do Not Usually Say
My computer thinks	My computer is thoughtful My computer is insightful My computer is wise
My computer hates me	My computer loves me My computer is depressed today My computer missed me last week
My computer is good [at computing]	My computer is [morally] good My computer is conscientious My computer is courageous



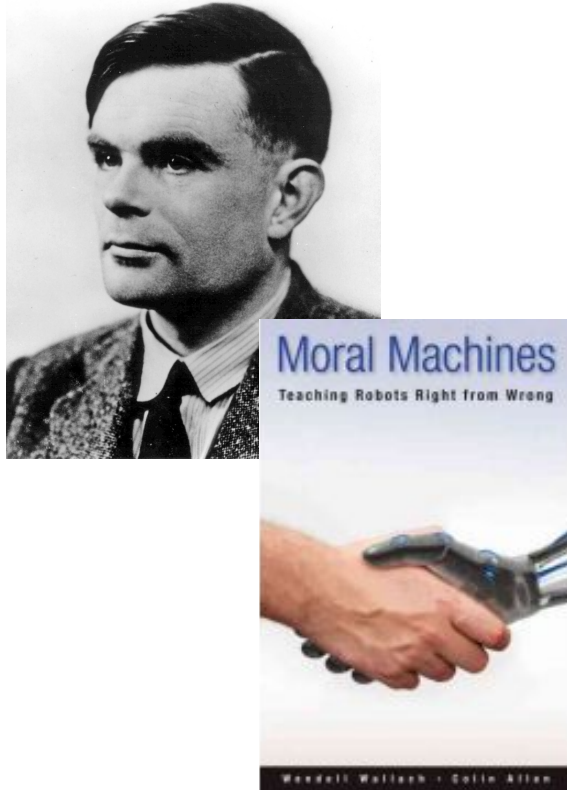
# Turing's Prediction and Its Moral Corollary

Things People May One Day Meaningfully Say	More Things People May One Day Meaningfully Say
My robot thinks	My robot is thoughtful My robot is insightful My robot is wise
My robot hates me	My robot loves me My robot is depressed today My robot missed me last week
My robot is good [at computing]	My robot is [morally] good My robot is conscientious My robot is courageous



# Turing's Prediction and Its Moral Corollary

## Why the Difference?

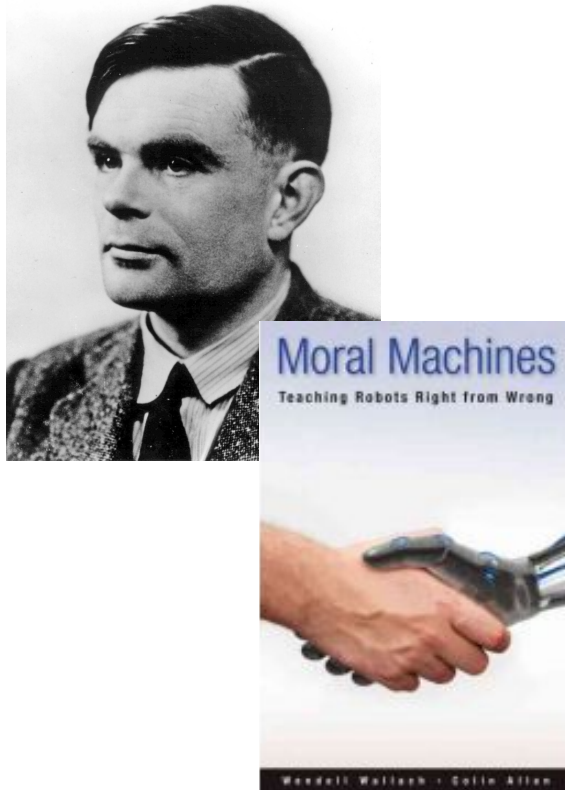


- Because robots, unlike computers, move around in the world?
- Because they exhibit externally-visible behavior and, hence, our judgments need not be based on inference, but observation?
- Because how we really judge another *person's* character is generally based on externals?



# Turing's Prediction and Its Moral Corollary

## Why the Difference?



- Because robots, unlike computers, move around in the world?
- Because they exhibit externally-visible behavior and, hence, our judgments need not be based on inference, but observation?
- Because how we really judge another *person's* character is generally based on externals?

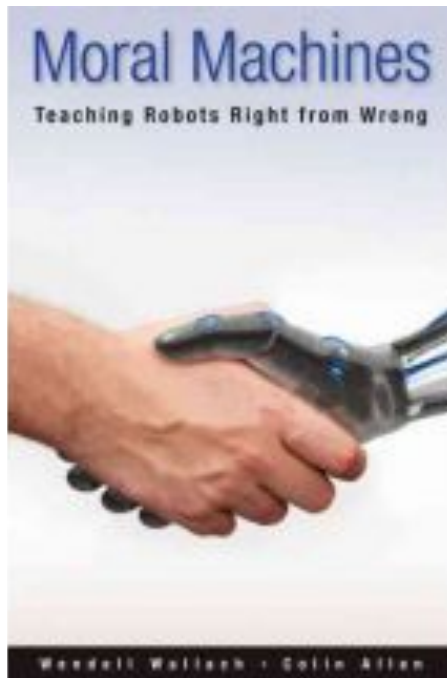
## A Quick Reminder

- Turing's famous test for intelligence (1950) uses a behavioral criteria based on indistinguishability from human behavior (or performance).



# Turing's Prediction and Its Moral Corollary

## The Moral Turing Test (MTT)



Oxford, 2009

- First proposed by Allen, Varner and Zinser in 2000
- Like Turing's test except it is not based on language or reasons but on actions

“Suppose the human judge in the MTT is provided with descriptions of actual, morally significant actions of a human and an AMA [automated moral agent], purged of all references that would identify the agents. If the judge correctly identifies the machine at a level above chance, then the machine has failed the test.” (Wallach and Allen 2009, 206)



## Statement of the Problem

“ ... let us suppose that sometime in the near future, we read the (rather long) headline, “First Robot Awarded Congressional Medal of Honor for Incredible Acts of Courage on the Battlefield.” What must we assume in the background for such a headline to make sense without profaning a nation’s highest award of valor? Minimally, fortitude and discipline, intention to act while undergoing the experience of fear, some notion of sacrifice with regard to one’s own life, and so forth, for what is courage without these things? That a robot might simulate them is surely not enough to warrant the attribution of virtue, unless we change the meaning of some terms.” (Beavers, 2011)



## Statement of the Problem

“ ... let us suppose that sometime in the near future, we read the (rather long) headline, “First Robot Awarded Congressional Medal of Honor for Incredible Acts of Courage on the Battlefield.” What must we assume in the background for such a headline to make sense without profaning a nation’s highest award of valor? Minimally, fortitude and discipline, intention to act while undergoing the experience of fear, some notion of sacrifice with regard to one’s own life, and so forth, for what is courage without these things? That a robot might simulate them is surely not enough to warrant the attribution of virtue, unless we change the meaning of some terms.” (Beavers, 2011)

In other words, interiority counts in ethics, unless we adopt strictly behavioral criteria. Furthermore, it seems that if computational ethics is to “make ethics honest,” then we must adopt such criteria. *This move, I will suggest, does not bode well for ethics. (And if not, then so be it.)*



## Statement of the Problem

The Principle of Utility:

“Actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure and the absence of pain; by unhappiness, pain and the privation of pleasure.” (1861, 7)



John Stuart Mill

## Statement of the Problem

“It is the business of ethics to tell us what are our duties, or by what test we may know them; but no system of ethics requires that the sole motive of all we do shall be a feeling of duty; on the contrary, ninety-nine hundredths of all our actions are done from other motives, and rightly so done if the rule of duty does not condemn them.” (1861, 17)

The morality of an action is thus determined by its conformity to a rule (in this case the principle of utility, but this is irrelevant to the current discussion.) **Interiority does not make a moral difference.**

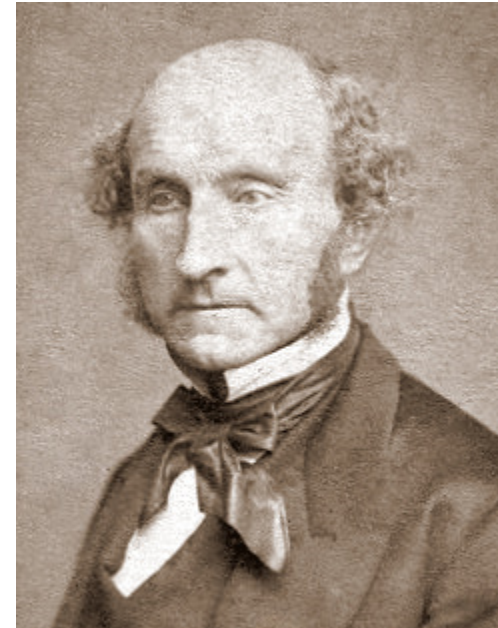


John Stuart Mill

## Statement of the Problem

The feeling of conscience may wane, but obligation may not. Thus, conscience cannot be a necessary condition for moral action. (1861, 29, *paraphrased*. See Section III for an extended discussion.)

**“We do not call anything wrong unless we mean to imply that a person ought to be punished in some way or other for doing it—** if not by law, by the opinion of his fellow creatures; if not by opinion, by the reproaches of his own conscience. This seems the real turning point of the distinction between morality and simple expediency.” (1861, 47)



John Stuart Mill

## Statement of the Problem

But:

The word “punishment” implies moral culpability and moral (i.e., non-causal) responsibility.

These imply, in turn, that it is in an agent’s power to commit or refrain from committing an action and that a choice has been made accordingly.

This implies, in turn, a reason (motive, desire, impulse, etc.) for acting that *ex hypothesi* is irrelevant to Mill, since moral action must be evaluated on a consequentialist (i.e., externalist) criterion.



## Statement of the Problem

But:

The word “punishment” implies moral culpability and moral (i.e., non-causal) responsibility.

These imply, in turn, that it is in an agent’s power to commit or refrain from committing an action and that a choice has been made accordingly.

This implies, in turn, a reason (motive, desire, impulse, etc.) for acting that *ex hypothesi* is irrelevant to Mill, since moral action must be evaluated on a consequentialist (i.e., externalist) criterion.

**Is this morally coherent?**



## Statement of the Problem

Is Mill a precursor to moral methodological behaviorism, even if his liberalism would make him want to resist this characterization?

*(In the process of “making ethics honest” will research in computational ethics push us in the same direction?)*



John Stuart Mill

## Statement of the Problem

It seems that to make Mill coherent, we have to define the psychology of the agent so that internals (conscience, moral responsibility and accountability) **are sufficient but not necessary conditions** for moral action, *apparently while preserving culpability*.

And, in turn, we get a series of collapsing moral and social distinctions:

- Acting for the sake of duty vs. acting in accordance with duty
- *Being* good vs. merely acting so
- Civic law and social custom vs. moral law
- Regarding moral agency, persons vs. things



## Statement of the Problem

Leaving us with no way to distinguish between:

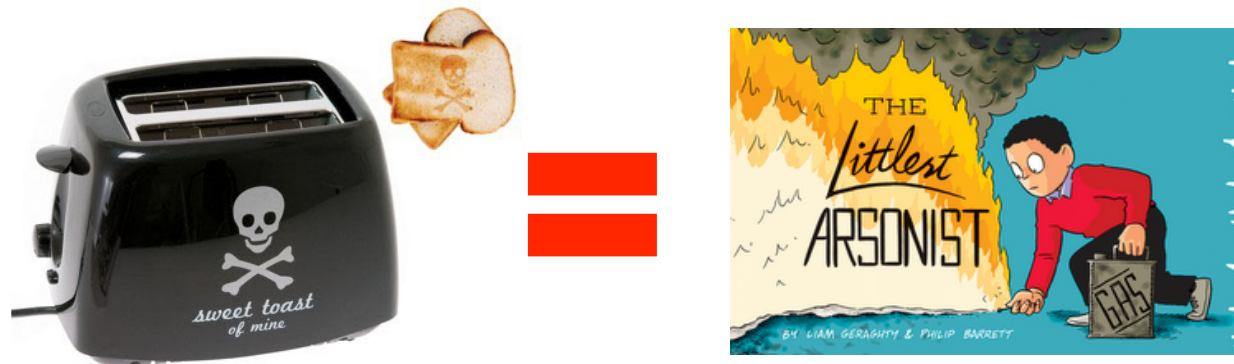
- My toaster was responsible for my house burning down
- Joe inadvertently knocked over a candle and was thus responsible for my house burning down
- Bob, the arsonist, purposely set fire to my house and was thus responsible for my house burning down



# Statement of the Problem

## The Problem

If interiority does not make a (necessary) moral difference, then all responsibility conflates to causal responsibility and there is no way to distinguish between moral wrong-doing, on the one hand, and illegal or socially-inappropriate behavior, on the other. To be moral is to behave according to a “code of conduct” and ethics is accordingly redefined.

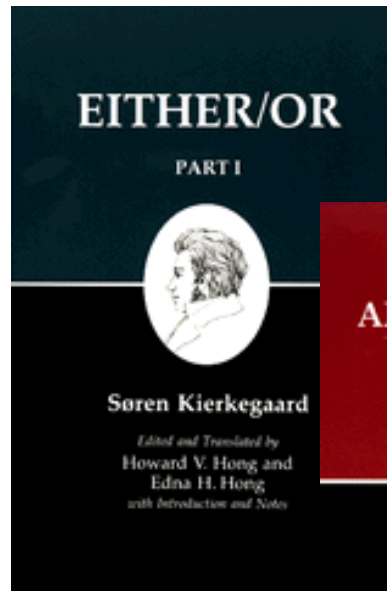


## On the Various Meanings of *Ought*

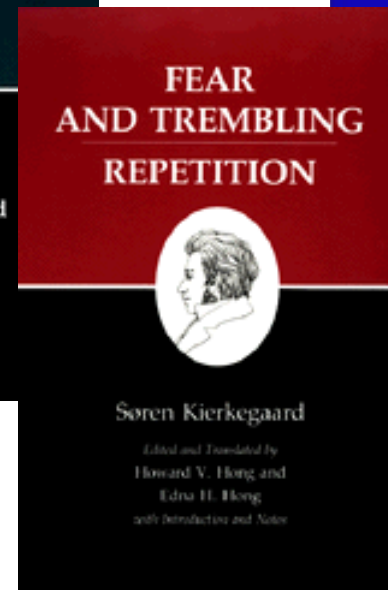


Søren Kierkegaard

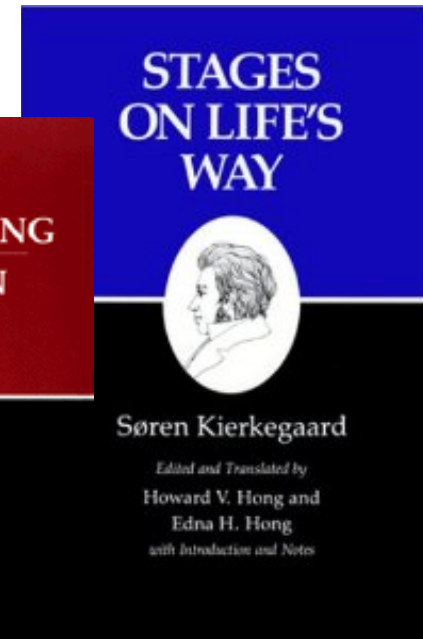
See Beavers, 1995, 2001 & Schrag, 1994.



1843



1843

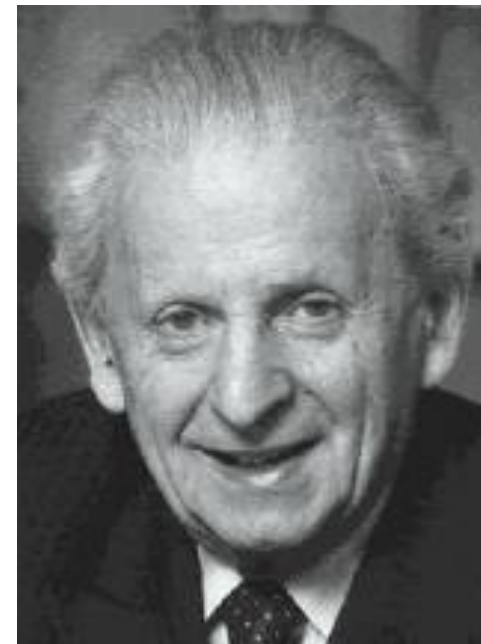


1845

## On the Various Meanings of *Ought*

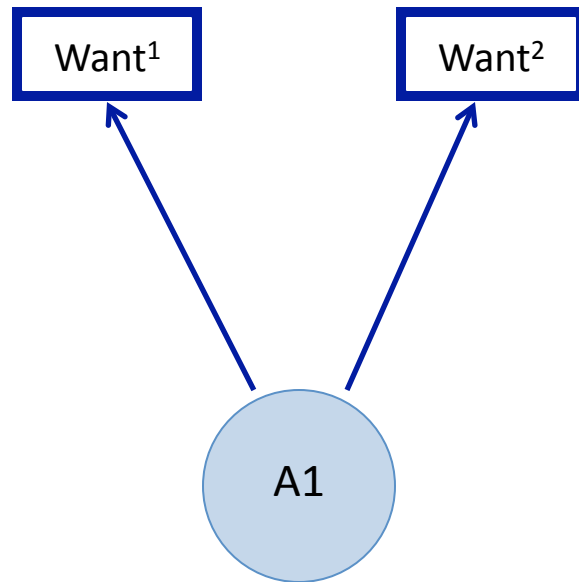
“The drama of existence is not only that existence is divided into choices between desires but that existence is also suspended between the Law that is given me and my nature, which is incapable of submitting to the Law without constraint. It is not freedom which defines the human being. It is obedience which defines him.”

“And God Created Woman,”  
1972/1990, 166

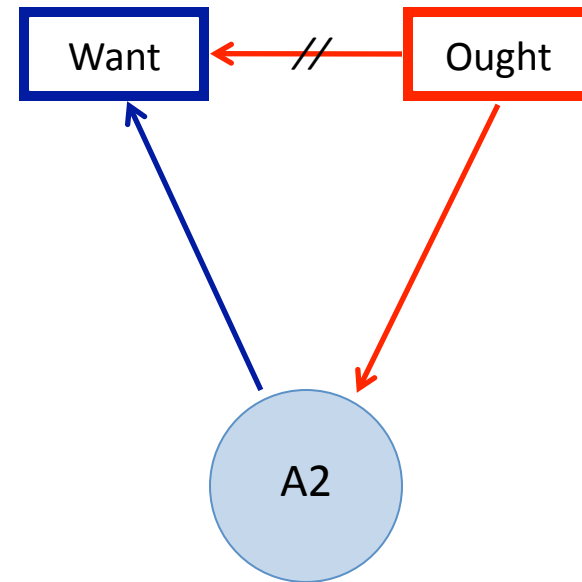


Emmanuel Levinas

## On the Various Meanings of *Ought*



Preferential Dilemma



Moral Dilemma



## On the Various Meanings of *Ought*



Immanuel Kant

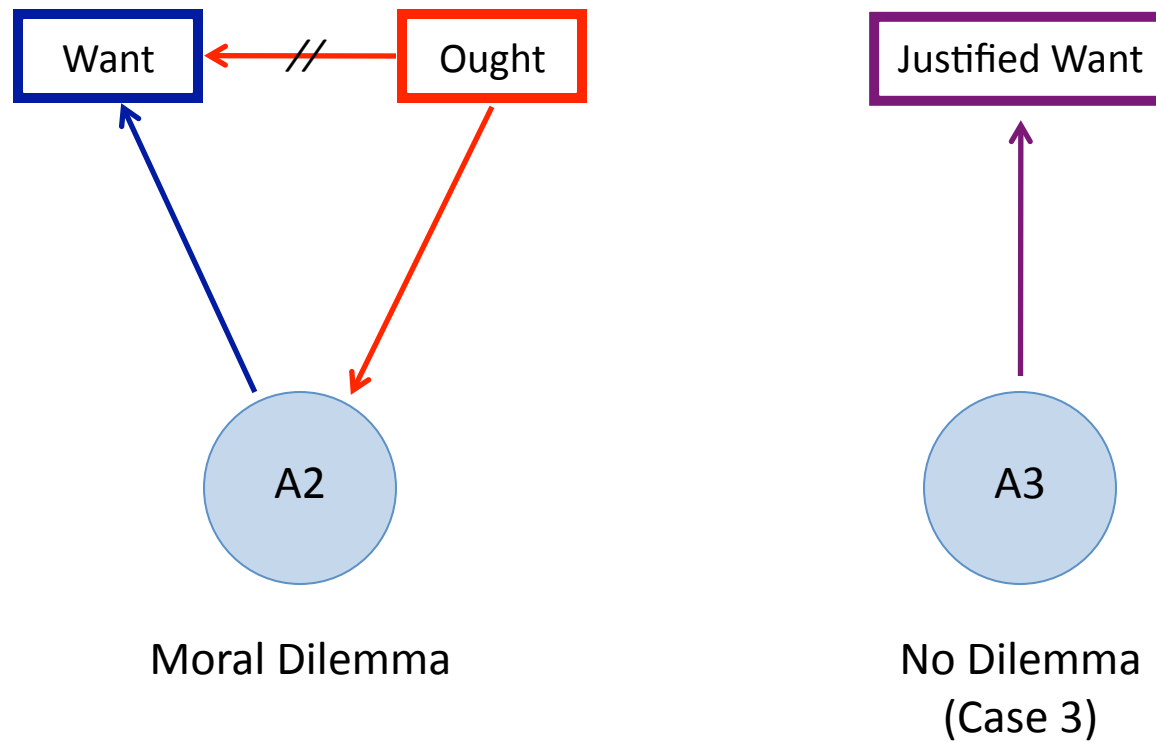
Case	Inclination (want)	Duty (ought)	Moral Status
0	No	No	
1	Yes	No	immoral
2	Yes <sup>1</sup>	Yes	immoral
3	Yes <sup>2</sup>	Yes	amoral
4	No	Yes	moral

<sup>1</sup> = mediate inclination; <sup>2</sup> = immediate inclination

**Mill doesn't care about the distinctions between 2, 3 & 4.**



## On the Various Meanings of *Ought*



# On the Various Meanings of *Ought*

## Kantian Agents

- B1) beings that are driven solely by inclination, such as animals.
- B2) beings that act solely out of reason and, therefore, duty, such as divine intellects.
- B3) beings so constituted that their inclinations always accord with duty, such as perfect human beings. (Case 3)
- B4) beings with inclinations that, at least some of the time, disagree with duty, but who nonetheless follow the dictates of duty. (Case 4)
- B5) beings with inclinations that, at least some of the time, disagree with duty, but who do what they want without regard for duty. (Case 1 & 2)



# On the Various Meanings of *Ought*

## What We Morally **Need** Our “Moral” Agents To Be

B4') beings with inclinations that, at least some of the time, disagree with duty, but who nonetheless follow the dictates of duty, *even though it is possible for them not to do so.*

**Is this morally (or logically) coherent?**

For an extended discussion on this point, see Beavers, 2009.



# On the Various Meanings of *Ought*

## What We **Want** Our “Moral” Agents To Be

B3) beings so constituted that their inclinations always accord with duty, such as perfect human beings.

**But . . .**

Such beings would be *designed to be good*, in which case they could not pass the courage test mentioned earlier, nor could they experience the moral dilemma that creates the necessary situation for moral choice that would allow them to be held responsible and accountable for their actions. They would not, in other words, be *moral* beings . . .

. . . at least if Kant is right. But supposing he isn't. What then?



## To Be Perfectly Honest . . . With Ourselves

- P1) *Moral* Interiority (i.e., conscience, choice as a response to genuine moral dilemmas, and hence, moral culpability and responsibility) is either a sufficient condition for ethics, a necessary condition, or both.
- P2) If ethical agency is evaluated on the basis of behavior and is achievable without interiority (as defined above), then interiority is not a necessary condition for ethics.
- P3) If a day will come when we commonly and rightfully speak of machines as being moral without being contradicted, then the condition in P2 will be satisfied.
- C) Hence, on that day, ethics will be defined so that moral interiority is merely a sufficient, but not necessary condition for ethics.



## To Be Perfectly Honest . . . With Ourselves

### Corollary:

On that day, conscience, moral choice, moral culpability and responsibility will be *inessential* components of ethics.

### Question:

What is ethical about ethics without these things? That is, how does ethics differ from any other preferential decision procedure, civic law or social custom?

This, of course, is to ask what is it that *ought*, morally and not preferentially, implies.



# To Be Perfectly Honest . . . With Ourselves

Traditionally understood, ought implies ...

- *Can*
  - Not must, thus, it implies *might not* (See Beavers, 2009)
  - *Implementability* (See Beavers, 2011)
  - *Moral Interiority* (See Beavers 1990, 1995 & 2001)
    - A desire that is challenged on the basis of some X, thus,
    - A moral dilemma that enables moral freedom
    - That makes my actions worthy of praise or blame
    - That makes me able to account for them, and,
    - Hence, makes me responsible for them.

Otherwise merely appearing or acting good is sufficient and necessary for justifying moral action, and the result is *moral methodological behaviorism*.



# To Be Perfectly Honest . . . With Ourselves

## An Open Question:

If the current direction of ethics is pushing us toward a *moral methodological behaviorism*, how long will it be before we accept a *moral metaphysical behaviorism* in which ethics is simply good, or appropriate behavior, however achieved?

**Should we be concerned?**



## To Be Perfectly Honest . . . With Ourselves

### Another Question:

But don't we already use behavioral criteria in assessing the ethical success of human beings?

### An Answer:

Yes (mostly), but the question, *and perhaps the question of the computability of ethics in general*, confuses the epistemic question of knowing what is good with the metaphysical question of how to be good.

“There is a knowledge that presumptuously wants to introduce into the world of spirit the same law of indifference under which the external world sighs. It believes that it is enough to know what is great—no other work is needed. But for this reason it does not get bread; it perishes of hunger while everything changes to gold.” (Kierkegaard, 1843/1983, 27-28)



## To Be Perfectly Honest . . . With Ourselves



Søren Kierkegaard

### From the Concluding Unscientific Postscript

“As soon as subjectivity is eliminated, and passion eliminated from subjectivity, and the infinite interest eliminated from passion, there is in general no decision at all ... All decisiveness, all essential decisiveness, is rooted in subjectivity. A contemplative spirit, and this is what the objective subject is, feels nowhere any infinite need of a decision, and sees no decision anywhere. This is the *falsum* that is inherent in all objectivity.”

(1846/1941, 33)



# Acknowledgements

I wish to thank Colin Allen, Susan Anderson, Larry Colter, Dick Connolly, Christopher Harrison, Deborah Johnson, Jim Moor, Dianne Oliver and Wendell Wallach for the many conversations and debates that led to the formulation of the views presented here.

# References

- Allen, C., Varner G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12: 251-261.
- Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4): 15-26.
- Beavers, A. (1990). Freedom and autonomy: The Kantian analytic and a Sartrean critique. *Philosophy and Theology* 5: 151-168.
- Beavers, A. (1995). *Levinas beyond the horizons of Cartesianism: An inquiry into the metaphysics of morals*. New York, NY: Peter Lang. <http://faculty.evansville.edu/tb2/PDFs/lbhc.pdf>.
- Beavers, A. (2001). Kant and the problem of ethical metaphysics. In M. New (Ed.), *In proximity: Emmanuel Levinas and the Eighteenth Century* (pp. 285-302). Lubbock, TX: Texas Tech University Press. <http://faculty.evansville.edu/tb2/PDFs/KantEM.pdf>.



## References

- Beavers, A. (2009). Between angels and animals: The question of robot ethics, or is Kantian moral agency desirable? Association for Practical and Professional Ethics, Eighteenth Annual Meeting, Cincinnati, Ohio, March 5th-8th, 2009. <http://faculty.evansville.edu/tb2/PDFs/Robot%20Ethics%20-%20APPE.pdf>.
- Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In P. Lin, G. Bekey & K. Abney (Eds.), *Robot ethics: The ethical and social implication of robotics*. Cambridge, MA: MIT Press, forthcoming. <http://faculty.evansville.edu/tb2/PDFs/Moral%20Machines%20and%20the%20Threat%20of%20Ethical%20Nihilism.pdf>.
- Dennett, D. (2006, May). Computers as prostheses for the imagination. The International Computers and Philosophy Conference, Laval, France.
- Kant, E. (1785/1981). *Grounding for the metaphysics of morals*. Trans. J. Ellington. Indianapolis: Hackett Publishing Company.
- Kierkegaard, S. (1843/1983). *Fear and trembling / Repetition*. Trans. H. Hong & E. Hong. Princeton, NJ: Princeton University Press.
- Kierkegaard, S. (1843/1988). *Either/Or*. Vol. 1 & 2. Trans. H. Hong & E. Hong. Princeton, NJ: Princeton University Press.
- Kierkegaard, S. (1845/1988). *Stages on life's way*. Trans. H. Hong & E. Hong. Princeton, NJ: Princeton University Press.



## References

- Kierkegaard, S. (1846/1941). Concluding unscientific postscript. Trans. D. Swenson & W. Lowrie. Princeton, NJ: Princeton University Press.
- Levinas, E. (1961/1967). *Totality and infinity: An essay on exteriority*. A. Lingis (Trans.). Pittsburgh, PA: Duquesne University Press.
- Levinas, E. (1972/1990). And God created woman. In A. Aronowicz (Ed.), *Nine Talmudic readings by Emmanuel Levinas*. Bloomington, IN: Indiana University Press, 1990.
- Mill, J. S. (1861/1979). *Utilitarianism*. Indianapolis, IN: Hackett Publishing Company.
- Schrag, C. (1994). Note on Kierkegaard's teleological suspension of the ethical. In *Philosophical papers: Betwixt and between*. Albany, NY: SUNY Press.
- Turing, A. (1950). Computing machinery and intelligence. *Mind* 59: 433-460.
- Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford, UK: Oxford University Press.

