# Ethical Principles

Module 4 of a course on *Ethical Issues in AI*

*Prepared by*

**John Hooker**

*Emeritus Professor, Carnegie Mellon University*

Chautauqua, June 2024

# Ethical principles

- We must have principles for resolving issues in an objective way.
  - Otherwise we can rationalize anything.
  - **Generalization principle**
  - **Utilitarian principle**
  - **Autonomy principle**

# Basic assumptions

- **Universality of reason**
  - *You don't get to have your own logic.*
- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*

# Basic assumptions

- **Universality of reason**
  - *You don't get to have your own logic.*
- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*
- This is the **deontological** approach to ethics.
  - ***Deontology** = What is required.*
    - Ethical principles represent what is required for the possibility of free action.

# Universality of reason

- What is rational **does not depend on who I am**.
  - *I don't get to have my own logic.*
- The assumption underlies science and all forms of rational inquiry.
  - *Ethics assumes nothing more.*

# Acting for reasons

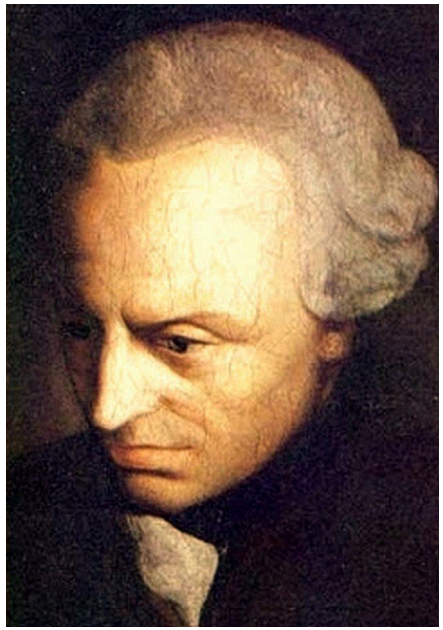
- Basic premise:  We always act for a reason.
  - *Every action has a rationale.*
- Why?
  - *This is how we distinguish **freely chosen action** from mere behavior.*
    - An MRI machine can detect our decisions **before we make them**.
    - If decisions are determined by **biological causes**, how can they be freely chosen?

# Acting for reasons

- Solution:
  - *Free chosen actions have **two kinds of explanation**:*
    - A biological cause
    - A rationale provided by the agent
  - *For example:*
    - A hiccup has **only** a biological explanation.
      Not a freely chosen action.
    - Drinking water to stop hiccups has
      **2 explanations**: a biological cause
      and a rationale. A freely chosen action.

# Acting for reasons

- Dual standpoint theory
  - *Originally proposed by Immanuel Kant.*
    - *Grundlegung zur Metaphysik der Sitten* (1785)
    - *Recent versions: Nagel (1986), Korsgaard (1996), Nelkin (2000), Bilgrami (2006).*
  - *Provides a **basis for ethics**.*
    - Ethical principles are **necessary conditions** for the logical coherence of an action's rationale.



# Generalization principle

# Generalization principle

- My action has a reason behind it.
  - *Why? Every action has a rationale.*
- So if the reason justifies the action for me...
  - *It justifies the action for **anyone to whom the reason applies**.*
  - *Why? Universality of reason.*

# Generalization principle

- Maybe I don't agree with universality of reason.
  - *Why can't I say, "My arguments show that cheating is unethical, but others are free to believe something else."*
    - They are, but I am saying they would be **wrong**.
    - That's what it **means** to believe that cheating is unethical.

# Generalization principle

- Maybe I don't agree with universality of reason.
  - *Why can't I say, "My arguments show that cheating is unethical, but others are free to believe something else."*
    - They are, but I am saying they would be **wrong**.
    - That's what it **means** to believe that cheating is unethical.
  - *Suppose I say, "My calculations show that 7 + 8 = 15, but others are free to believe that 7 + 8 is something else."*
    - They are, but I am saying they would be wrong.
    - That's what it **means** to believe that 7 + 8 = 15.

# Generalization principle

- **Example**
- Suppose I steal a watch from a shop.
- I have 2 reasons:
  - *I want a new watch.*
  - *I won't get caught.*
    - Security at the shop is lax.

# Example - Theft

- So I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*

# Example - Theft

- So I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*
- But I know that if all do this, they will get caught.
  - *The shop will install security.*
  - *My reasons will no longer apply*

# Example - Theft

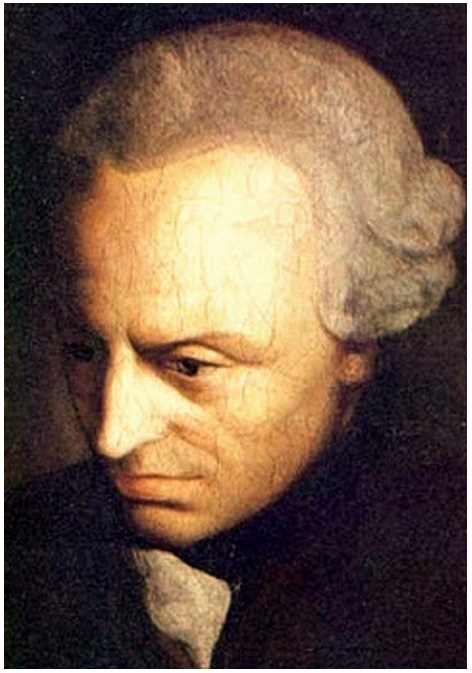
- I am not saying that all these people actually **will** steal watches.
  - *Only that if they did, my reasons would no longer apply.*

# Example - Theft

- My reasons are **inconsistent** with the assumption that people will act on them.
- I am caught in a contradiction.
  - *I am deciding that these reasons justify theft for **me**.*
  - *But I am **not** deciding that these reasons justify theft for **others**.*
  - *I can't have it both ways.*

# Generalization principle

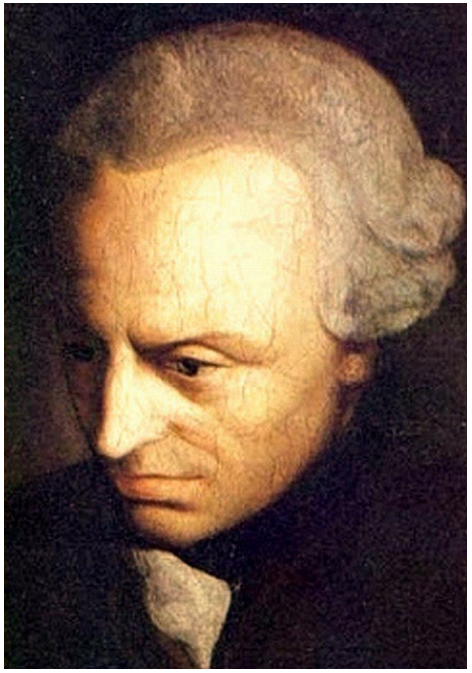
- The principle is:
  - *The **reasons** for an action should be **consistent** with the assumption that **everyone with the same reasons acts the same way**.*



Immanuel Kant
1724-1804

# Generalization principle

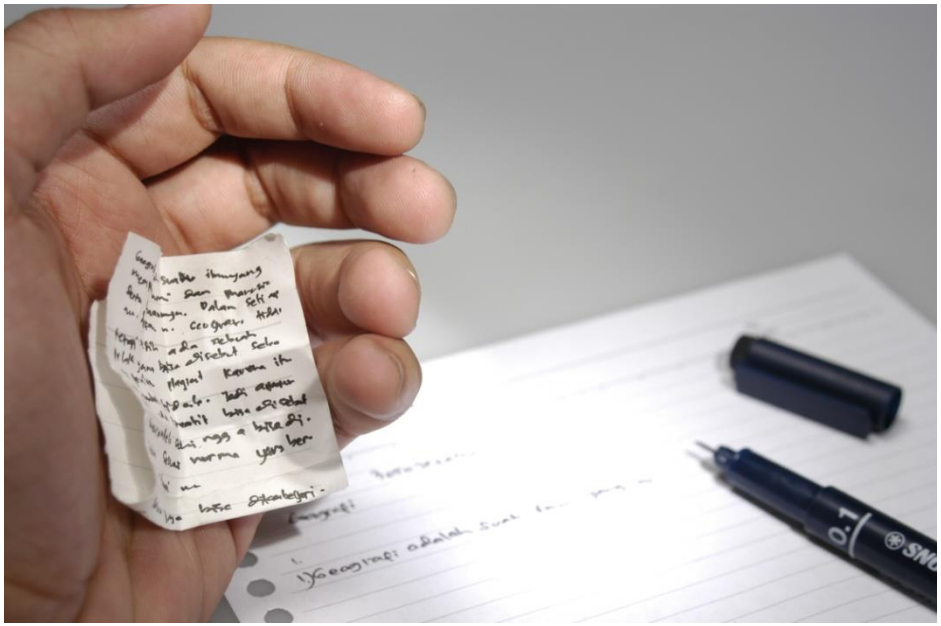
- Or more precisely:
  - *It should be **rational** for me to believe that the **reasons** for my action are **consistent** with the assumption that **everyone with the same reasons acts the same way**.*



Immanuel Kant
1724-1804

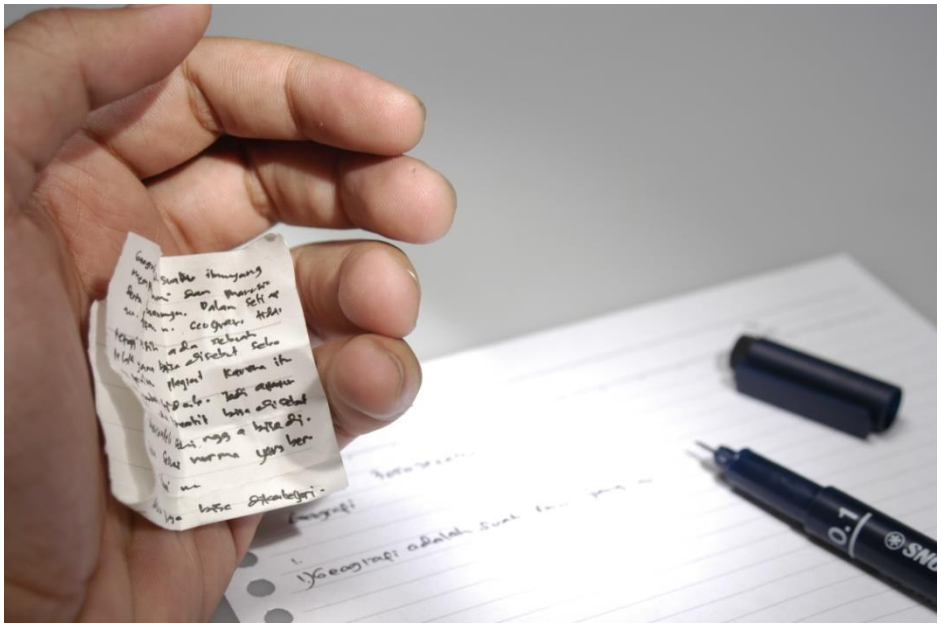
# Example - Cheating

- What is wrong with cheating on an exam?
- My reasons:
  - *I will get a better grade.*
  - *Which means I will get a better job.*



# Example - Cheating

- Nearly all students have these reasons.
- If they all cheat…
  - *Everyone will have a top grade.*
  - *Good grades won't get me a better job.*



# Example - Agreements

- **Breaking an agreement** violates generalization principle.
  - *If I break it merely for convenience or profit.*
  - *An agreement (or contract) is a mutual promise.*

# Example - Agreements

- Suppose everyone broke agreements when convenient.
  - *It would be impossible to **make** agreements in the first place.*
  - *And therefore impossible to achieve my purposes by **breaking** them!*
  - *The whole point of having an agreement is that you keep it when you **don't want** to keep it.*

# Example - Lying

- Lying for mere convenience violates the generalization principle.
  - *If the reason for lying implies that people will believe the lie.*
  - *If everyone lied when convenient, no one would believe the lies.*
    - The possibility of **communication** presupposes a certain amount of credibility.

# Example - Lying

- Lying can be generalizable, depending on the reasons.
  - *Workers in an Amsterdam office building lied to Nazi police, 1940-42.*
    - They denied knowing the whereabouts of Anne Frank's family, who they knew were hiding in the building.
    - Their purpose was to avoid revealing a Jewish family's location to the Nazi regime.

# Example - Lying

- Lying can be generalizable, depending on the reasons.
  - *This is generalizable*
    - If everyone lied to avoid revealing a Jewish family's location to the Nazi regime, it would still be possible to accomplish this purpose by lying.
    - It would not be necessary for the police to believe the lies.

# Alternate formulations

- The generalization principle can be formulated:
  - *It should be rational for me to believe that I could **achieve my purposes** if everyone with my reasons acted the same way.*
    - I would not achieve my purpose by cheating if everyone cheated to get a better job.

# Alternate formulations

- The principle can also be formulated:
  - *It should be rational for me to believe that the **practice or institution** that makes achieving my purposes possible would **not be undermined** if everyone with my reasons acted the same way.*
    - The institution of grades would be undermined if everyone cheated to get a better job.

# Example - Theft

- Suppose I steal whenever it is convenient and profitable for me.  Why is this unethical?
  - *Because it is illegal?*
    - Suppose it were legal.  Would this make it OK?

# Example - Theft

- Suppose I steal whenever it is convenient and profitable for me. Why is this unethical?
  - *Because it is illegal?*
    - Suppose it were legal. Would this make it OK?
  - *It would undermine the institution of **property**.*
    - I steal something to have possession and use of it.
    - That is, to make it **my property**.
    - If everyone stole for convenience, there would be no institution of property.
    - When I steal something, others will steal from me 5 minutes later.

# Example - Deception

- One can **deceive** without lying.
  - *For example, if your doctor deliberately neglects to mention a serious diagnosis.*
    - There is no lying, only deception.
    - Deception = **causing someone to believe something you know is false**.

# Example - Deception

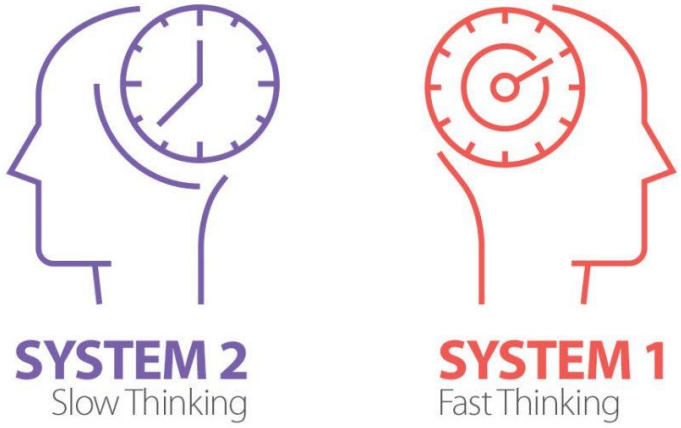
- One can **deceive** without lying.
  - *For example, if your doctor deliberately neglects to mention a serious diagnosis.*
    - There is no lying, only deception.
    - Deception = **causing someone to believe something you know is false**.
  - *Deception, merely for convenience, is **not generalizable**.*
    - It would not deceive if generalized.

# Human decision making

- A flaw in rationality-based ethics?
  - *Most of our actions are automatic.*
    - We can't devise a rationale for everything we do.
    - We are creatures of habit.
  - *This doesn't mean we are unethical most of the time.*

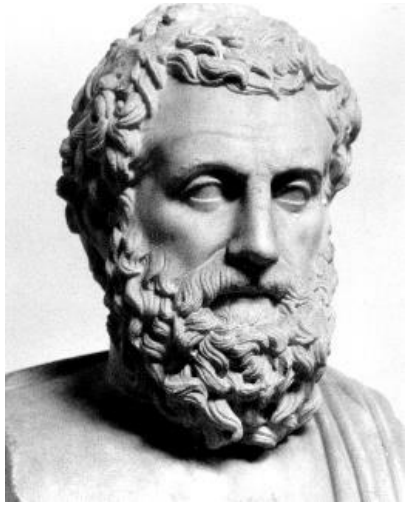
# Human decision making

- A flaw in rationality-based ethics?
  - *Most of our actions are automatic.*
    - We can't devise a rationale for everything we do.
    - We are creatures of habit.
  - *This doesn't mean we are unethical most of the time.*
  - ***Dual process theory** agrees.*
    - **System 1 thinking** is fast and unconscious.
    - **System 2 thinking** is slow and based on conscious reasoning.
    - We usually rely on System 1.



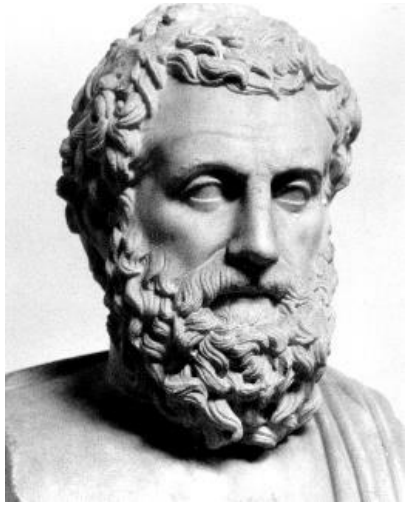
# Human decision making

- Ethicists are well aware of this
  - *Going back at least to Aristotle.*
  - *We allows habits to **continue**.*
    - If I continue smoking, I **make a decision** not to break the habit.



# Human decision making

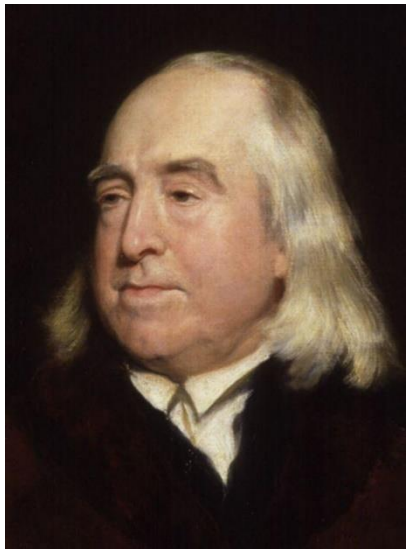
- Ethicists are well aware of this
  - *Going back at least to Aristotle.*
  - *We allows habits to **continue**.*
    - If I continue smoking, I **make a decision** not to break the habit.
  - *We can **invoke system 2 thinking** when needed.*
    - This is where **ethics** comes into play.



# Utilitarian principle

# Importance of utility

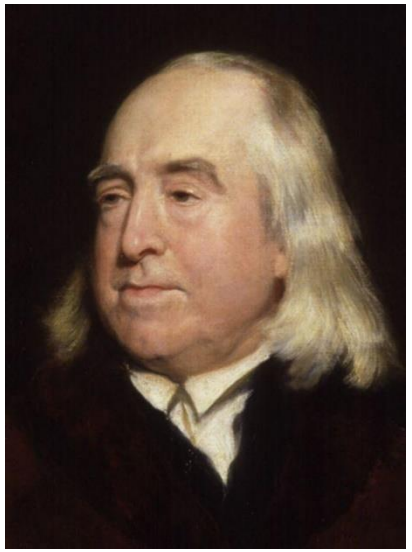
- The **utilitarian principle** is based on the idea that one should try to make things better.
  - *Or as Jeremy Bentham put it, one's actions should* ***maximize utility***
    - That is, create the **greatest good for the greatest number**
      *"On the principle of utility" (1780)*



Jeremy Bentham
Father of utilitarianism

# Importance of utility

- The **utilitarian principle** is based on the idea that one should try to make things better.
  - *Or as Jeremy Bentham put it, one's actions should* ***maximize utility***
    - That is, create the **greatest good for the greatest number**
      *"On the principle of utility" (1780)*
    - For example, Bentham believed that criminal penalties should be designed to reduce crime rather than exact retribution.

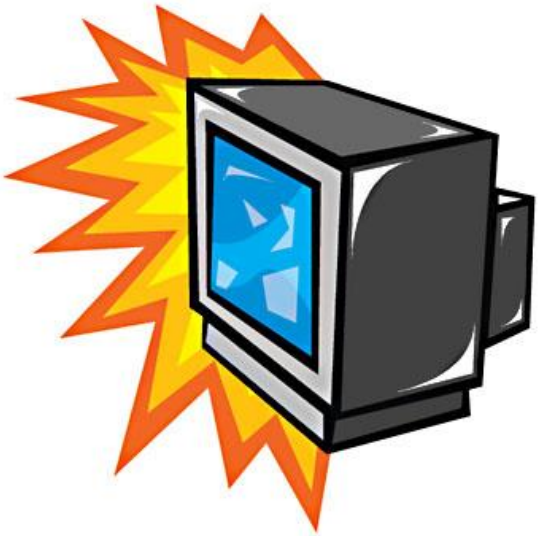


Jeremy Bentham
Father of utilitarianism


Bentham’s skeleton dressed in his clothes, with wax head, in student center of University College London

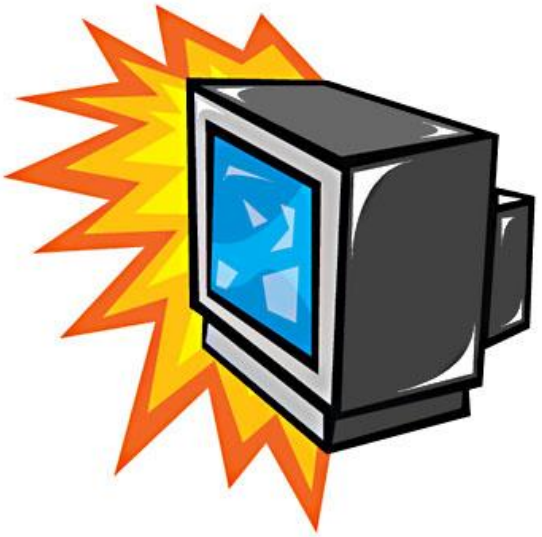
# Importance of utility

- For example, suppose I listen to loud TV in my hotel room at 2 am.
  - *Keeping other guests awake.*
    - Why is this unethical?
    - Let's say it doesn't violate hotel rules
    - **So it satisfies the generalization principle**.



# Importance of utility

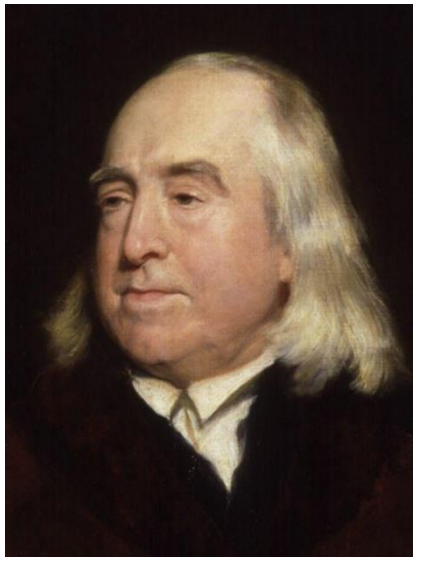
- For example, suppose I listen to loud TV in my hotel room at 2 am.
  - *Keeping other guests awake.*
    - Why is this unethical?
    - Let's say it doesn't violate hotel rules
    - **So it satisfies the generalization principle**.



- Problem: it reduces net utility.
  - *Maybe it makes me a little happier.*
  - *But it substantially reduces utility of other guests.*

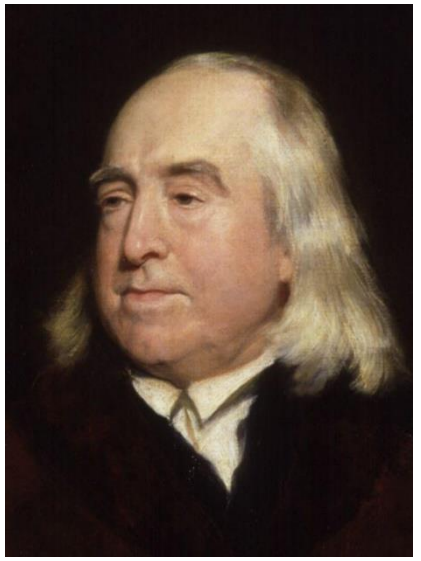
# Basic argument

- Step 1: An action is often a **means to an end**.
  - *You may want to achieve some goal.*
  - *Maybe your ultimate goal is happiness.*
    - This was Bentham's suggestion.
  - *Whatever it is, let's call it **utility**.*
    - It's what you regard as **inherently valuable**, as the **end** to which your actions are a **means**.



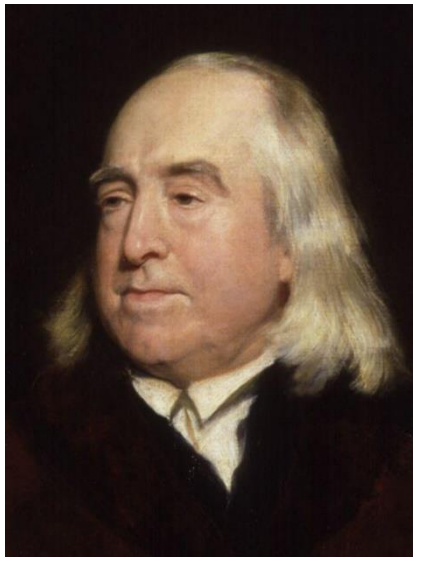
# Basic argument

- Step 2:  If I regard something as inherently valuable...
  - *I must regard it as inherently valuable for **anyone** (not just me).*
  - *…due to the **universality of reason**.*



# Basic argument

- Step 3:  My actions should take everyone's happiness as seriously as my own.
  - *Bentham thought this means* ***maximizing total net utility****.*
    - This is adequate for most purposes.
    - We will go with it for now.



# Utilitarian principle

- An act is ethical only if I can rationally believe that **no other act**…
  - *creates **more net expected utility****…*
  - *and **satisfies other ethical principles**.*

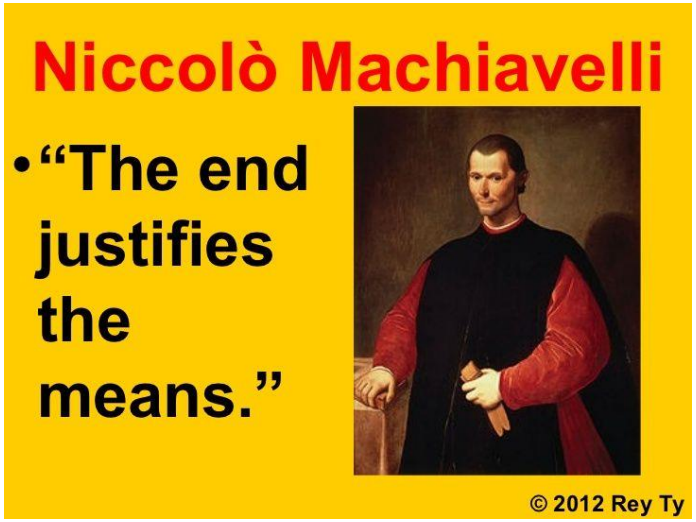
**counting **everyone's** utility.*

# Utilitarian principle

- Why consider only actions that satisfy other ethical principles?
  - *Because behavior that doesn't satisfy other ethical principles is **not action**.*
    - And so is not a freely chosen option.
  - *So utility can never "override" the other principles.*

# Utilitarian principle

- Does the **end** justify the **means**?
  - *Only an end **can** justify a means.*
  - *But only it can do so **only if**…*
    - The means satisfies the **generalization** and **autonomy** principles.
    - The disutility of the means **doesn't offset** the utility of the end.



# Measuring utility

- What if it's hard to predict the outcome of a decision?
  - *I am not required to have a crystal ball.*
  - *I need only make a **rational** determination, given the evidence.*
  - *If there is no way to tell, the utilitarian principle is satisfied by default.*

# Measuring utility

- But I must make a reasonable effort to research the issue.
  - *The same effort I would invest in decisions that affect me.*
    - For example, it is **irrational** to buy real estate without checking into it.
  - *Find **optimal tradeoff** between research and need to act.*
    - This is “satisficing,” a concept introduced by Herbert Simon.
    - Don’t waste time researching minor decisions.

# Autonomy principle

# Autonomy

- Fundamental obligation: **respect autonomy**.
  - *This rules out murder, coercion, slavery, etc.*
- Why this obligation?
  - *Will explain shortly…*

# Autonomy

- Autonomy = self-law
  - *I act **autonomously** when I freely make up my own mind about what to do, based on coherent reasons I give for my decision.*
  - *An **agent** is someone who can act autonomously.*
    - Sometimes called a “moral agent.”
  - ***Agency** is the exercise of autonomy.*

# Autonomy

- Autonomous vs. programmed
  - *An "autonomous car" is **not** autonomous in this sense.*
  - *It is only **programmed**.*
    - …and therefore independent of real-time human control.

# Action plans

- To make things more precise…
  - *An action has the form of an **action plan**.*
    - **If** the reasons for my action apply, **then** do it.
    - Example: "If I want to catch the bus, and the bus stop is across the street, and no cars are coming, then cross the street."

# Coercion

- Coercion violates my autonomy if it **interferes with my action plan**.
  - *I start to cross the street to catch a bus, **no cars are coming**, and you pull me off the street.*
  - *This interferes with my action plan.*
  - *A violation of autonomy.*

# Coercion

- Coercion does **not** violate my autonomy if it is **consistent with my action plan**.
  - *I start to cross the street to catch a bus, and you **pull me out of the path of a car**.*
  - *This is **consistent** with my action plan.*
  - *Not a violation of autonomy.*

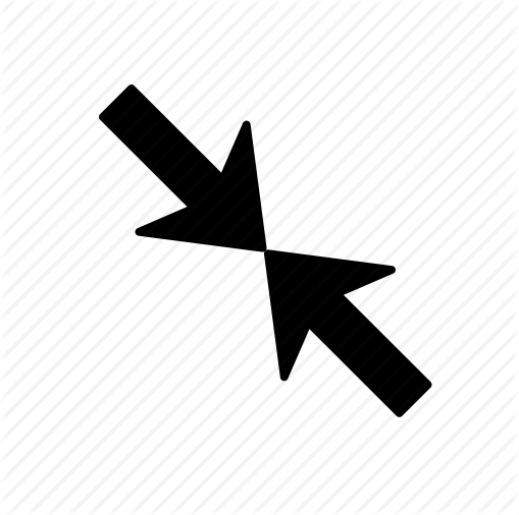
# Autonomy principle

- An action plan is **unethical** if the agent is **rationally constrained to believe*** that…
  - *it **interferes** with the **ethical action plans** of some collection of other agents **without informed or implied consent**.*

**it is irrational not to believe…*

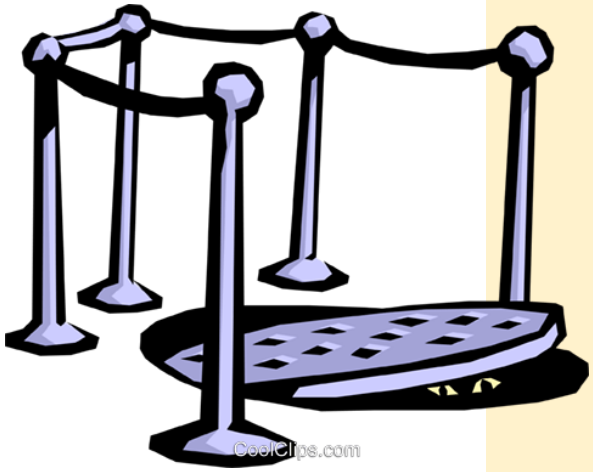
# Autonomy principle

- Why?
  - Let's say I interfere with **your** ethical action plan.
  - If **I were you**, I would be interfering with **my own** action plan, which makes no sense.
  - But the universality of reason says that when choosing an action, **it should not matter** whether I am me or I am you.
  - So interfering with an ethical action plan is **self-contradictory**.



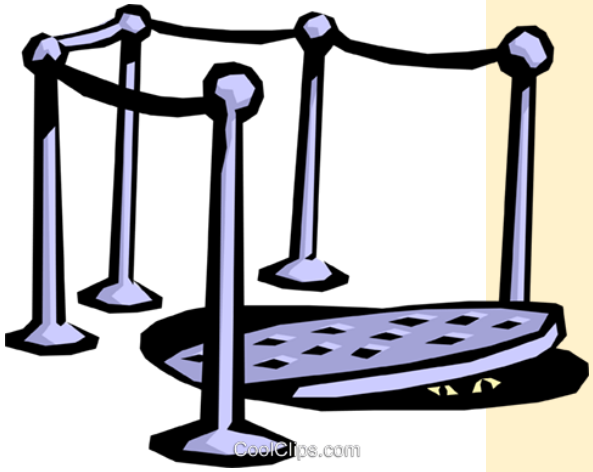
# Autonomy principle

- I must be **rationally constrained to believe** there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - If someone falls into a manhole I leave unprotected while working under a busy street, this is **not** a violation of autonomy.
    - It is only **probable** that someone will fall in.
    - My neglect violates the **utilitarian** principle.



# Autonomy principle

- I must be **rationally constrained to believe** there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - However, suppose I replace the manhole cover with one that will **collapse** when someone steps on it (a booby trap).
    - And it is on **5th Ave NYC**.
    - I must believe this will interfere with someone's action plans.
    - This **violates autonomy**.



# Autonomy principle

- Suppose my commanding officer orders me to torture a prisoner.
  - ***Futility argument:*** *results are the same if I refuse, as someone else will obey the orders.*
    - This shows that the torture passes **utilitarian** test.

Abu Ghraib Prison, Iraq

# Autonomy principle

- Suppose my commanding officer orders me to torture a prisoner.
  - ***Futility argument:*** *results are the same if I refuse, as someone else will obey the orders.*
    - This shows that the torture passes **utilitarian** test.
  - ***Autonomy argument:*** *My torture violates autonomy of the prisoner.*

# Informed consent

- Coercion does **not** violate autonomy if there is **informed consent**.
  - *I attend a concert where there are strict rules against recording the performance.*
  - *Yet I record it anyway.*
  - *Ushers compel me to leave.*
  - *This does not violate my autonomy*
    - I gave informed consent.
    - My action plan was, "If I am not kicked out for doing so, I will record the performance."
    - The ushers did not interfere with this action plan.

# Informed consent

- Coercion does **not** violate autonomy if there is **informed consent**.
  - *My employer tells me I must transfer to another city or be fired.*
  - *This seems inconsistent with my action plan.*
  - *But by taking the job, I implicitly agreed to abide by the company's business decisions.*
  - *So my action plan is consistent with the company's decision.*

# Limits on autonomy

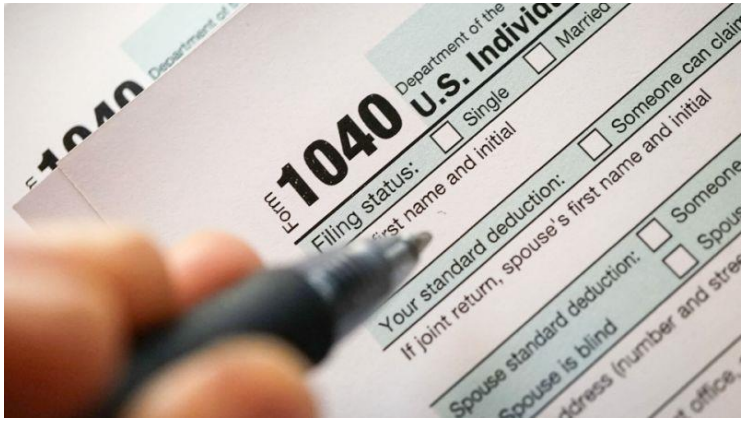
- The autonomy principle doesn't require you to allow people to **do anything they want**.
  - *You can interfere with **unethical** action.*
    - Because unethical action is **not really action**, and so there is no interference with an action plan.

# Limits on autonomy

- The autonomy principle doesn't require you to allow people to **do anything they want**.
  - *You can interfere with **unethical** action.*
    - Because unethical action is **not really action**, and so there is no interference with an action plan.
    - You can defend yourself, because attacking you is unethical.
    - You can keep people off your property, because trespassing when forbidden by owner is illegal and therefore unethical.

# Interference principle

- ***More*** *than minimal coercion is problematic.*
  - I can't lock you in a closet to prevent you from cheating on your income tax.
  - This interferes with many ethical actions.



# Limits on autonomy

- A restaurant can ethically refuse to serve me unless I wear a mask.
  - *This does not interfere with **my** action plan.*
    - I cannot have an action plan of being served. This is not my decision to make.
    - I can only have an action plan of eating in the restaurant **if served**.
  - *A government mandate is another issue.*

# Limits on autonomy

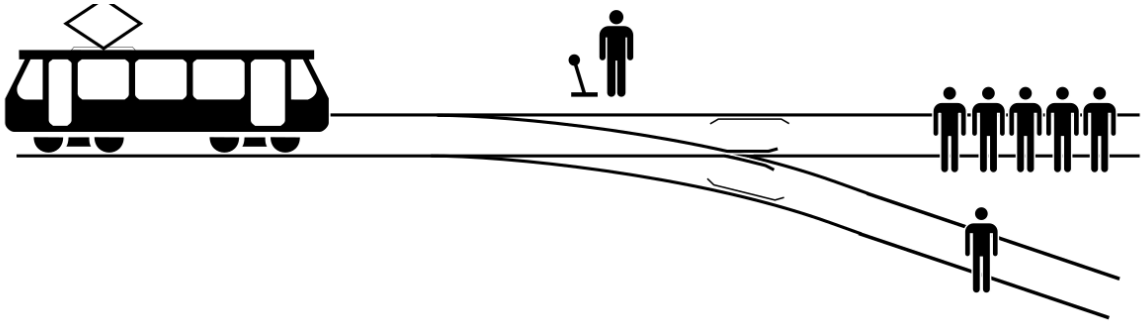
- However, a prison guard cannot ethically refuse to serve me food and water.
  - *True, I cannot have an action plan of being served food and water.*
    - But deprivation of necessities interferes with any and all of my ethical action plans.
    - It therefore violates autonomy.

# Limits on autonomy

- My boss cannot ethically require me to contribute to a political party I don't support.
  - *I gave **no informed consent** to this.*
  - *But there is **no violation of autonomy**.*
    - I cannot have an action plan of being employed while making no political contributions.
  - *But this **violates generalizability**.*
    - It is a breach of the employment agreement, which implicitly promises that my duties will be related to the operation of the business.

# Trolley car dilemmas

- Often introduced to stimulate discussion.
  - *Allow trolley to kill 5 people, or pull switch and kill one person?*
    - Gives impression that ethical dilemmas cannot be resolved.



# Trolley car dilemmas

- Often introduced to stimulate discussion.
  - *Allow trolley to kill 5 people, or pull switch and kill one person?*
    - Gives impression that ethical dilemmas cannot be resolved.
  - *Analysis? Autonomous action is impossible.*
    - Ethics tells us to avoid these situations in war, police work, medical triage, etc.