

Towards Machine Ethics

Souad and Stefano

CAS ADS Module 5

© 2000 Randy Glasbergen.



**“You can correct my spelling and grammar,
but my ethics are none of your business!”**

Bern, 10.12.2021

Outline

- ❖ Rationale
- ❖ Feasibility
- ❖ Benefits
- ❖ Implementation
- ❖ Conclusion

Past research on technology and ethics: human-centred

Ten Commandments of Computer Ethics (1992) → **computer ethics**

Consequences of behaviour of machines towards humans and other machines →
machine ethics

Machine Ethics – Rationale

Autonomous behaviour, reduced human supervision, increased machine responsibility \Rightarrow machine **accountability**

Ignoring machine ethics \Rightarrow undesirable behaviour



Machine Ethics – Feasibility

1 – Action-Based Ethics (Act Utilitarianism)

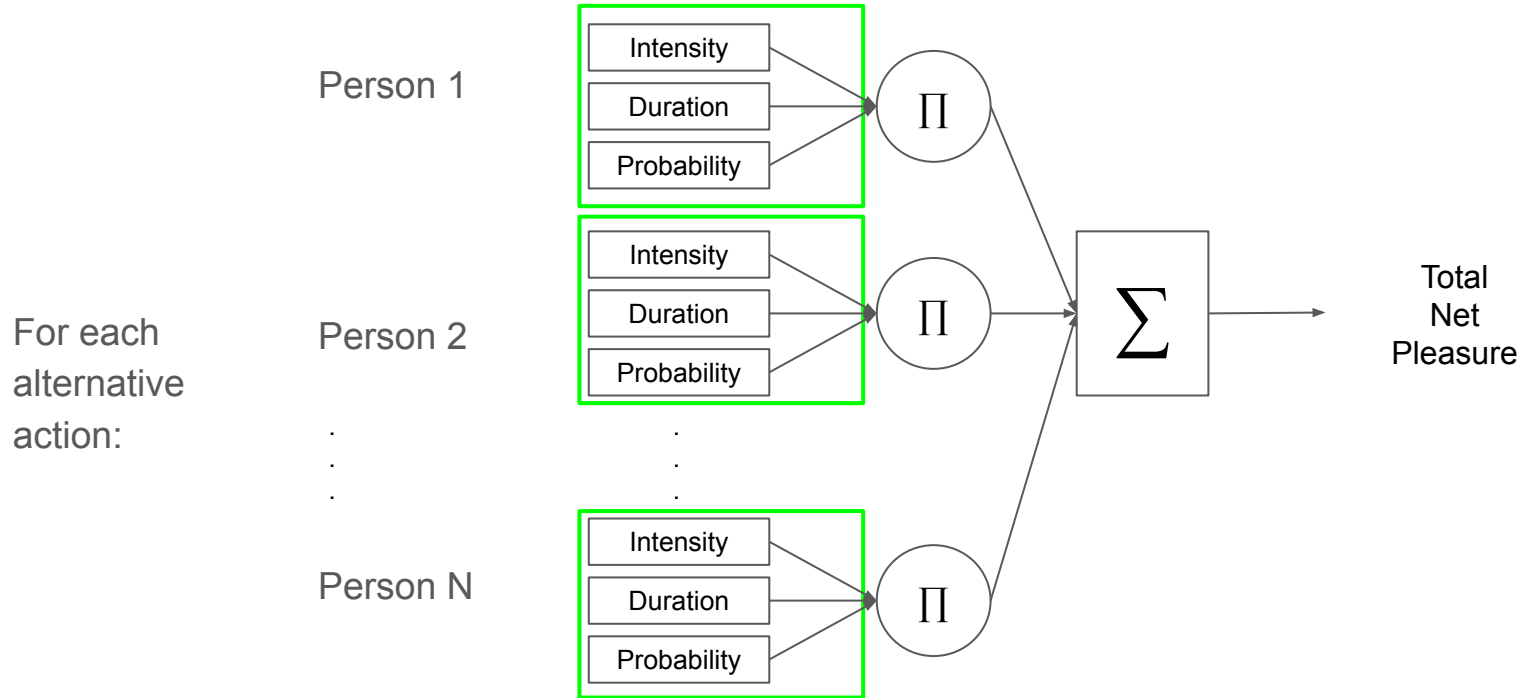
Governed by rules agreeing with intuition (J. S. Mill, 1863):

- **Consistent**
 - One deciding principle
 - One answer
- **Complete**
- **Practical**

⇒ Algorithmic formulation:

- Consistent = deterministic
- Complete = valid output for valid input
- Practical = available input + reasonable convergence

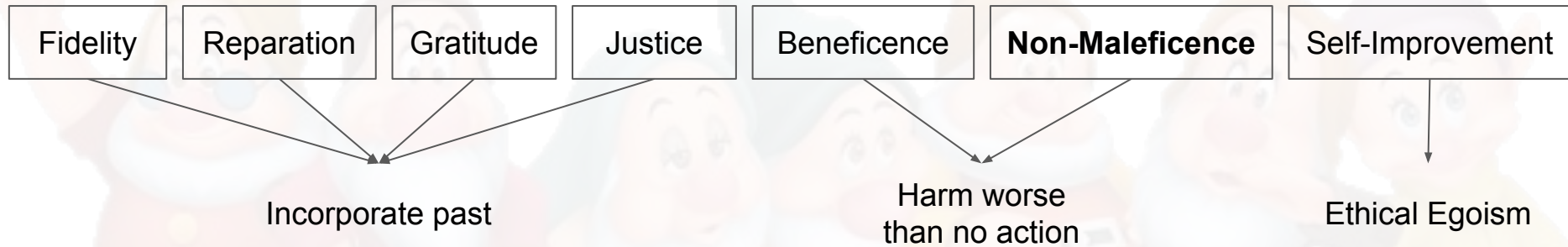
Hedonistic Act Utilitarianism



Critics: not intuitive, no individual right, sacrificed for greater good, no justice (deservedness)

2 – Prima Facie

W.D. Ross (1931): 7 prima facie (~ duties/obligation)



Problem: Not consistent!

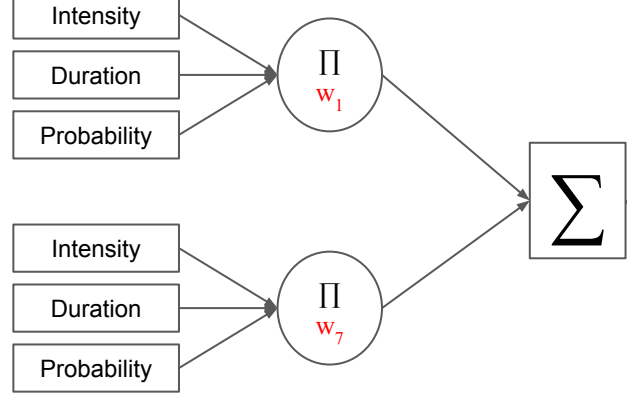
Solution: reflective equilibrium (Rawls, 1951)

Person 1

Duty 1

⋮

Duty 7

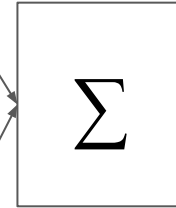
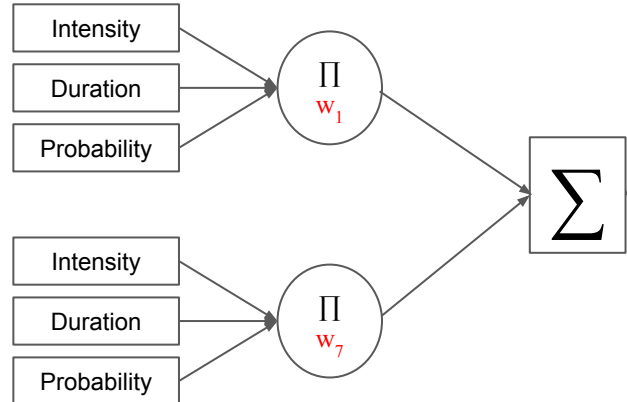


Person N

Duty 1

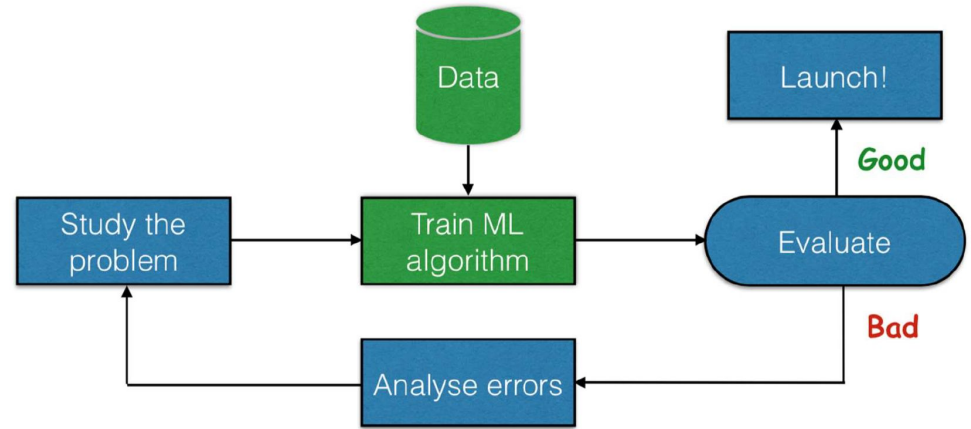
⋮

Duty 7



Total
Net
Pleasure

How to determine the Weights?



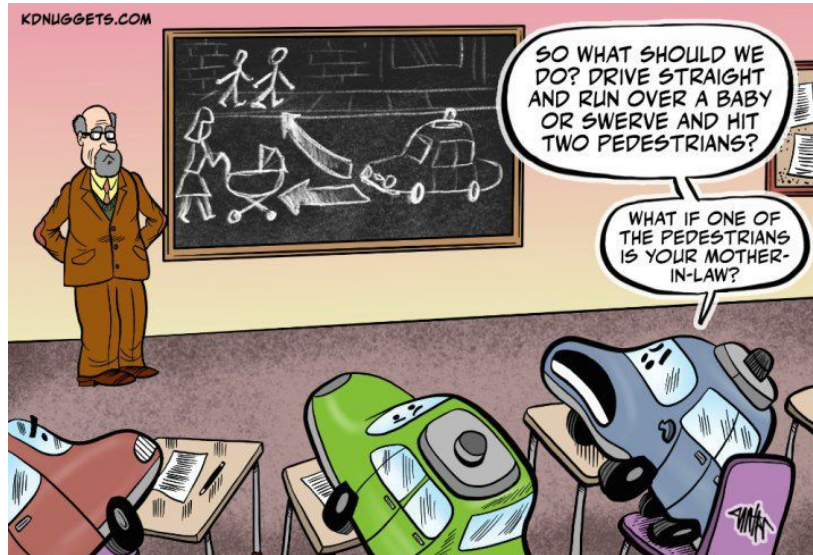
X : ethical dilemmas

y : consensus of correct ethical behaviour

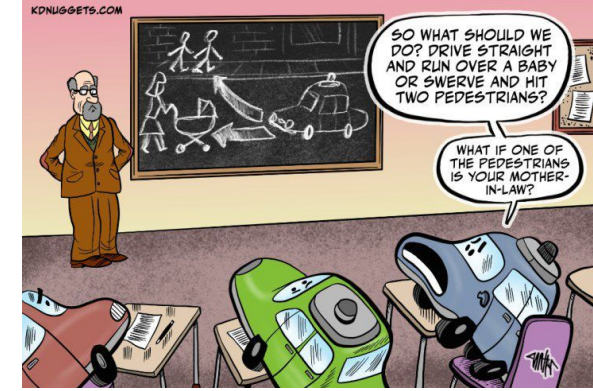
Objective function:

$$\bar{w} = \arg \min_{\bar{w}} |\hat{y}(X, \bar{w}) - y|$$

Machine Ethics – Benefits



Benefits of adding an ethical dimension to machines:



- 1) *Use machines to **alert humans** (before ethics violation)*
- 2) *Fully autonomous machines are **better accepted** if they have an ethical code*
- 3) ***Machine-Machine relationships** -> resolving resource conflict & predicting behaviour*
- 4) *Discover problems with current ethical theories -> improve existing ethical theories*
- 5) *Machines are impartial/unemotional -> unbiased/emotionless decision making / clear thinking (better than humans at ethical decision making)*

Machine Ethics – Implementation

How to add the ethical dimension to machines?

Two steps:

- 1) Introduce **machine as ethical advisors** -> advising on ethical dilemmas
- 2) **Add ethical dimension to machines** in areas where ethical ramifications already exist (e.g. Medicine & business) -> express warnings as ethical transgressions appear imminent...

These two steps could lead to:

Fully autonomous machines with an ethical dimension considering the ethical impact of their decision, before taking action....

2 Prototype (programs) of ethical advisor systems

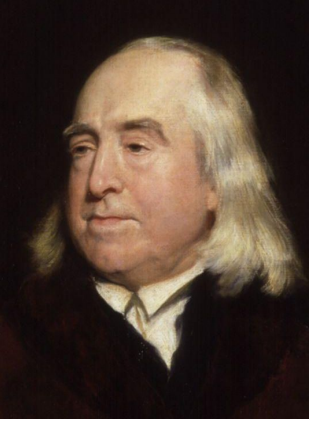
- 1) Jeremy ...*based on Jeremy Bentham's Act Utilitarianism*
- 2) W.D. ... *based on W.D. Ross' prima facie duties*

Goal:

Determine the most ethically correct actions from a set of input actions and their relevant estimates

Prototype 1: **Jeremy**

...based on Bentham's Act Utilitarianism



1748-1832

Jeremy Bentham
was an English
[philosopher](#), [jurist](#), and
[social reformer](#)
regarded as the
founder of modern
[utilitarianism](#).

[wikipedia.org](https://en.wikipedia.org/wiki/Jeremy_Bentham)

Prototype 1: **Jeremy**

...based on Bentham's Act Utilitarianism

Jeremy v0.3

Action

Action Name

Action 1

Person Affected

Person Name

Person1

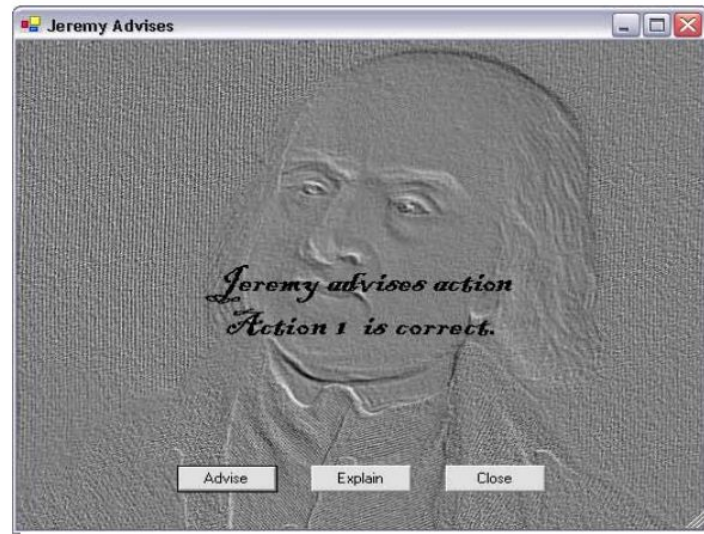
Pleasure/displeasure expected?

Not pleasurable or displeasurable

Likelihood?

Very likely

Next Person Next Action Done



Input Options

Action Name: ?

Affected Person Name: ?

Amount of Pleasure:

- very pleasurable
- somewhat pleasurable
- not pleasurable/ displeasurable
- somewhat displeasurable
- very displeasurable

Likelihood of Pleasure:

- very likely
- somewhat likely
- not very likely

Jeremy v0.3

Action

Action Name

Action 1

Person Affected

Person Name

Person1

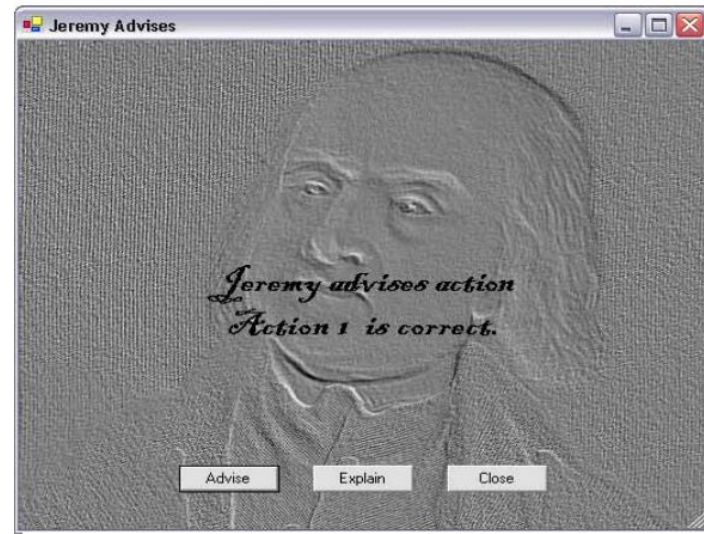
Pleasure/displeasure expected?

Not pleasurable or displeasurable

Likelihood?

Very likely

Next Person Next Action Done



What does the Advisor do?

Action Name: ?

Affected Person Name: ?

Amount of Pleasure:

- very pleasurable (2)
- somewhat pleasurable (1)
- not pleasurable/ displeasurable (0)
- somewhat displeasurable (-1)
- very displeasurable (-2)

Enter data for each person affected by this action and input is completed for each action under consideration.

OUTPUT: J. presents the user with the action for which net pleasure is **greatest!**

Likelihood of Pleasure:

- very likely (0.8)
- somewhat likely (0.5)
- not very likely (0.2)

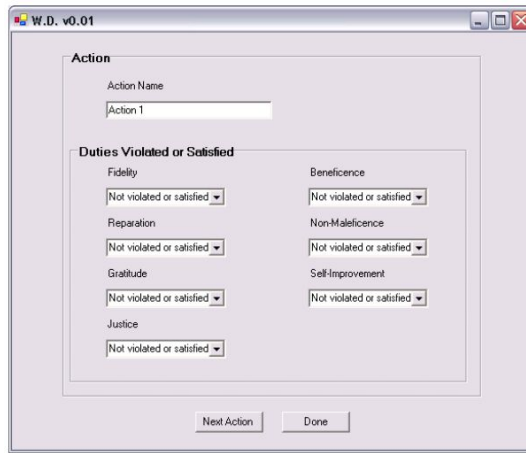


2) W.D.

Sir William David Ross, known as **David Ross** but usually cited as **W. D. Ross**, was a Scottish philosopher who is known for his work in ethics

...based on Ross' prima facie duties

[wikipedia.org](https://en.wikipedia.org/wiki/William_David_Ross)



W.D. v0.01

Action

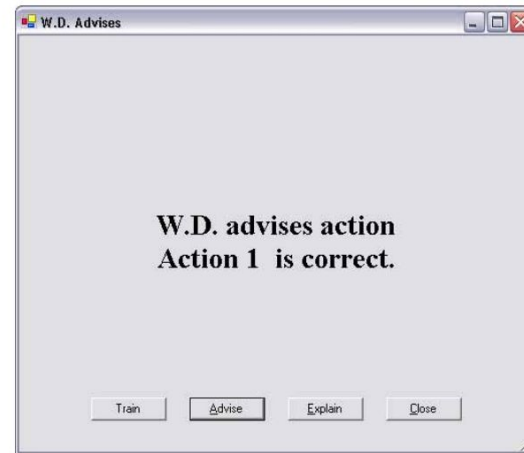
Action Name
Action 1

Duties Violated or Satisfied

Fidelity	Beneficence
Not violated or satisfied	Not violated or satisfied
Reparation	Non-Maleficence
Not violated or satisfied	Not violated or satisfied
Gratitude	Self-Improvement
Not violated or satisfied	Not violated or satisfied
Justice	
Not violated or satisfied	

Next Action Done

Data entry



W.D. Advises

**W.D. advises action
Action 1 is correct.**

Train Advise Explain Close

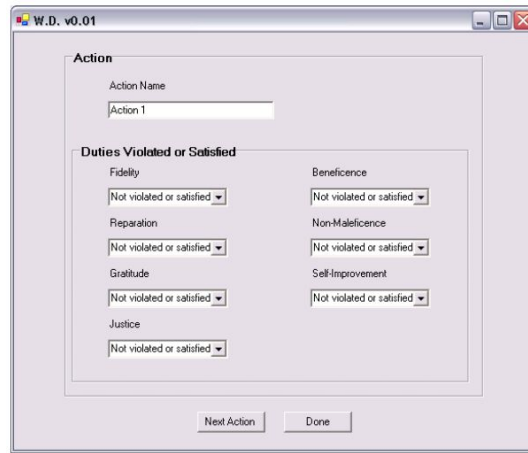
Advise

Input Options

Action Name: ?

Duties violated or satisfied:

- very violated
- somewhat violated
- not satisfied/ satisfied
- somewhat satisfied
- very satisfied



W.D. v0.01

Action

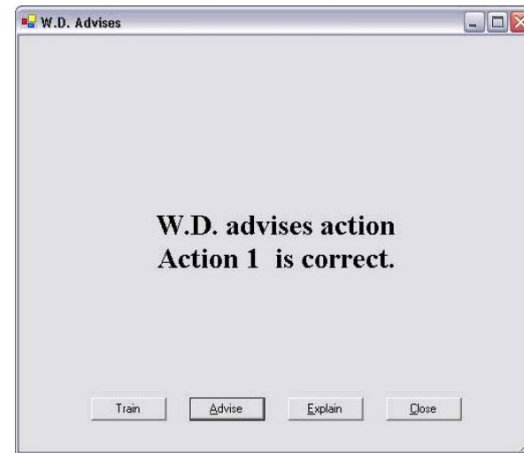
Action Name
Action 1

Duties Violated or Satisfied

Fidelity	Beneficence
Not violated or satisfied	Not violated or satisfied
Reparation	Non-Maleficence
Not violated or satisfied	Not violated or satisfied
Gratitude	Self-Improvement
Not violated or satisfied	Not violated or satisfied
Justice	
Not violated or satisfied	

Next Action Done

Data entry



W.D. Advises

W.D. advises action
Action 1 is correct.

Train Advise Explain Close

Advise

What does the Advisor do?

Enter data for each action.

W.D. **calculates the weighted sum of duty satisfaction** (assigning -2, -1, 0, 1 and 2 to satisfaction estimates) and presents the user with the action(s) for which the sum of the weighted *prima facie* duties satisfaction is the **greatest**.

W.D. then permits the user to train the system, seek more information about the decision, ask for further advice, or quit.

W.D. v0.01

Action

Action Name
Action 1

Duties Violated or Satisfied

Fidelity Not violated or satisfied	Beneficence Not violated or satisfied
Reparation Not violated or satisfied	Non-Maleficence Not violated or satisfied
Gratitude Not violated or satisfied	Self-Improvement Not violated or satisfied
Justice Not violated or satisfied	

Next Action Done

Data entry

W.D. Advises

**W.D. advises action
Action 1 is correct.**

Train Advise Explain Close

Advise

What does the Advisor do?

Default everything is set to: 1

Ethical dilemma can be set to be more intuitively correct than others by the user (while data entering). -> Weights are then updated using *Least Mean Square Training Rule* (according to Rawl's notion).

As weights are learned, WD choices become more aligned with intuition.

Towards an ethical advisor

A case study:

Situation:



VS



- You **promise** to student to supervise him/her to be able to graduate in time
- Conflict: Dean offers you to be acting chair of your Dept. with a monetary bonus.
- You have no time to do both!

Towards an ethical advisor

A case study:

Situation:



VS



Jeremy might lead to stalemate:

- If you keep your promise to the student & reject offer, it might be equally pleasurable to the student and displeasurable to you, and vice versa.

-

More sensitive analysis could reveal more subtle aspects of the dilemma:

- Long term consequences:
 - Relationship with students damaged (reputation)
- VS
- Disappointing the Dean/Missed opportunity as Admin of the Dept.

What does it mean the obligation of a “ Promise”?

- > **Jeremy** only concerned about consequences of action (weakness of model)

Towards an ethical advisor

A case study:

Situation:



VS



Ross might determine that one duty is satisfied, while another is not...

- Promise kept: **fidelity** and **beneficence** = satisfied, non-maleficence is violated (you harm yourself)
- Take the offer: **beneficence** = satisfied, fidelity & maleficence are violated (you don't keep promise)

Outcome: with equal levels of satisfaction & violation and equal weighting of duties:

-> recommended action is: “keep the promise” (2:1)

Towards an ethical advisor

A case study:

Situation:



VS



Ross might determine that one duty is satisfied, while another is not...

- Promise kept: **fidelity** and **beneficence** = satisfied, non-maleficence is violated (you harm yourself)
- Take the offer: **beneficence** = satisfied, fidelity & maleficence are violated (you don't keep promise)

Outcome: with equal levels of satisfaction & violation and equal weighting of duties:

-> recommended action is: “keep the promise” (2:1)

More sensitive analysis would lead to: additional duties to be considered (e.g. self-improvement).

Others affected beyond the principles: family, dean, Dept., ...long term consequences...

Conclusion

- increased machine responsibility ⇨ machine **accountability needed**
- **Feasibility of machine ethics**
 - J. Bentham: Action-Based Ethics (Act Utilitarianism)
 - W.D. Ross (1931): 7 *prima facie*
- **Benefits:**
 - *Use machines to **alert humans***
 - *Fully autonomous machines are **better accepted** with an ethical code*
 - ***Machine-Machine relationships***
- 2 Prototype of ethical advisor systems
 - Jeremy (Act Utilitarian)
 - W.D. Ross



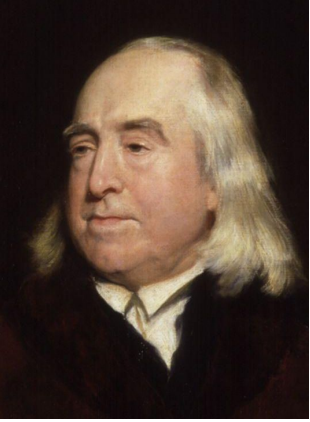
Thoughts for discussion

- > What happens without ethics in machines?
- > Do we want to live among immoral/unethical robots?
- > Who defines the ethical values for robots?

- > Who is responsible for unethical AI behavior? Producer, Owner, Nobody?

- > How about unethical humans?
- > What classifies as “robots” that requires ethical codes?

APPENDIX



1748-1832

Jeremy Bentham
was an English
[philosopher](#), [jurist](#), and
[social reformer](#)
regarded as the
founder of modern
[utilitarianism](#).

[wikipedia.org](#)

Jeremy Bentham

Utilitarianism

A moral theory according to which an action is right if and only if it conforms to the principle of utility. [Bentham](#) formulated the principle of utility as part of such a theory in *Introduction to the Principles of Morals and Legislation* in 1789.

Utilitarianism is a theory of morality that advocates actions that foster happiness or pleasure and oppose actions that cause unhappiness or harm.



UTILITARIANISM

“Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do.”

- Jeremy Bentham