Engineers who make Selfish Machines

The Ethics of Socio-Technical Systems

Philip N. Brown
Dept of Computer Science
University of Colorado, Colorado Springs
Traditional Engineering Ethics Training

- Challenger Space Shuttle Accident
- Tacoma Narrows Bridge Collapse
- Hyatt Regency Walkway Collapse
- Bhopal Disaster
- Chernobyl Meltdown
Hyatt Regency Walkway Collapse

- Kansas City, 1981
- Hotel walkways collapse onto crowded dance floor
- 114 killed, 216 injured
- Cause: Systemic neglect of proper design review

Image credit: Public Domain, Dr. Lee Lowery, Jr., P.E.
Hyatt Regency Walkway Collapse

- Kansas City, 1981
- Hotel walkways collapse onto crowded dance floor
- 114 killed, 216 injured
- **Cause:** Systemic neglect of proper design review

**Engineer:**

"Build this"

"Ok, whatever"

Only supports 60% of required load

**Fabricator:**

"too expensive... can we do this?"

Only supports **half** load of red design

Image credit: Public Domain, Wikimedia Commons
Ethics Message: Engineer things that don’t break

Implication: Adequate design is sufficient.
Traditional Engineering Ethics Training

- Challenger Space Shuttle Accident
- Tacoma Narrows Bridge Collapse
- Hyatt Regency Walkway Collapse
- Bhopal Disaster
- Chernobyl Meltdown

Ethics Message: Engineer things that don’t break

Implication: Adequate design is sufficient.
Engineered System
1001101
0010011

Social System

Socio-Technical Paradigm
Socio-Technical Paradigm

Engineered System
1001101
0010011

Social System

Smart Grid

ebay

Bitcoin

Ethereum

lyft

University of Colorado
Colorado Springs
• Why does the feedback loop matter?
• What are ethical implications?
• How to teach this?
Stand on right

Image credit: Public Domain, Pixabay
Escalator Example

Engineered System

Social System

Walk on left

Stand on right

Image credit: Public Domain, Pixabay
Walk on left

Stand on right
Human Intuition: Don’t be that guy!

Walk on left  Stand on right
Human Intuition: Don’t be that guy!

Question: Robot’s escalator policy?

Walk on left

Stand on right
Question: Robot’s escalator policy?
Question: Robot’s escalator policy?

Human Intuition: not this
Escalator Example

Low Impact

Selfish

Stand Left

Systems Perspective: definitely this (maximizes throughput)

For more info: Prof. Lesley Strawderman at Mississippi State University
Escalator Example

Systems Perspective:
definitely this
(maximizes throughput)

For more info: Prof. Lesley Strawderman at Mississippi State University
**Escalator Example**

**Benefits:**
- Selfish: Robot acts like a human
- Altruistic: Everybody waiting to use escalator

**Harms:**
- Selfish: Everybody waiting to use escalator
- Altruistic: Robot looks like a jerk!

**Ethical?**
- Selfish: Bad for “system”
- Altruistic: Good for “system”

For more info: Prof. Lesley Strawderman at Mississippi State University
Traditional Ethics Message: Engineer things that don’t break

Needed Update: Engineer your machines to interact with people
Self-driving car design

Choosing Routes in Highway Networks
Choosing Routes in Highway Networks

Agenda: pose simple model
   Explore ethics in its context
Choosing Routes in Highway Networks

Agenda: pose simple model
   Explore ethics in its context

Question: should self-driving cars be altruistic?
Self-driving car design

Fast, if empty
Congestion-sensitive

15 minutes per car
Travel Time

1 hour
Slow no matter what
Congestion-insensitive
Self-driving car design

15 minutes
Fast, if empty
Congestion-sensitive

15 minutes per car
Travel Time

1 hour
Slow no matter what
Congestion-insensitive

1 hour
Traffic

Total Time
3.25 hours
Self-driving car design

30 minutes
30 minutes

Fast, if empty
Congestion-sensitive

15 minutes per car

Travel Time

1 hour

1 hour

Slow no matter what
Congestion-insensitive

Traffic

Total Time
3.25 hours
3 hours

3.25 hours
3 hours
Self-driving car design

45 minutes
45 minutes
45 minutes

Fast, if empty
Congestion-sensitive

15 minutes per car
Travel Time

Slow no matter what
Congestion-insensitive

1 hour

Traffic

Total Time

3.25 hours
3 hours
3.25 hours

3.25 hours
3 hours
3.25 hours
Self-driving car design

- Fast, if empty
- Congestion-sensitive
- 15 minutes per car
- Travel Time
- Slow no matter what
- Congestion-insensitive

Traffic:
- Total Time:
  - 3.25 hours
  - 3 hours
  - 3.25 hours
  - 4 hours
Self-driving car design

1 hour
1 hour
1 hour
1 hour
1 hour

15 minutes per car

Travel Time

1 hour

Best option!
(Pareto optimal)

Traffic

Total Time

3.25 hours

3 hours

3.25 hours

4 hours
Self-driving car design

30 minutes

30 minutes

1 hour

1 hour

15 minutes per car

Travel Time

S

D

3.25 hours

3 hours

3.25 hours

4 hours

Traffic

Total Time

Best option!
(Pareto optimal)
Self-driving car design

**Travel Time**
- 30 minutes
- 30 minutes
- 1 hour
- 1 hour

**Total Time**
- 3.25 hours
- 3 hours
- 3.25 hours
- 4 hours

**Best option! (Pareto optimal)**

**Selfish traffic is like this!**

**Incentive to switch!**

Altruistic self-driving cars?
Altruistic self-driving cars?

Altruism: act like there is 2x actual traffic

Best option! (Pareto optimal)

Selfish traffic is like this!

Traffic

Total Time

3.25 hours

3 hours

3.25 hours

4 hours
Self-driving car design

1 hour
1 hour
1 hour
1 hour

Travel Time

15 minutes per car

S

D

Total Time

3.25 hours

3 hours

3.25 hours

4 hours

Best option! (Pareto optimal)

Selfish traffic is like this!

Traffic

3.25 hours

3 hours

3.25 hours

4 hours
Self-driving car design

45 minutes
45 minutes
45 minutes

15 minutes per car

S

Travel Time

D

1 hour

1 hour

Total Time
3.25 hours
3 hours
3.25 hours

Traffic

Best option!
(Pareto optimal)

4 hours

3.25 hours
Self-driving car design

Traffic

Total Time

45 minutes
45 minutes
90 minutes
1 hour

1 hour

15 minutes per car

Travel Time

3.25 hours
3 hours
3.25 hours
4 hours

Best option!
(Pareto optimal)
Self-driving car design

30 minutes
30 minutes

Travel Time
15 minutes per car

S

D

1 hour
1 hour
1 hour

Best option!
(Pareto optimal)

Traffic

Total Time
3.25 hours
3 hours
3.25 hours
4 hours

30 minutes
1 hour
30 minutes
1 hour
1 hour

3.25 hours
3 hours
3.25 hours
4 hours

University of Colorado
Colorado Springs
Altruistic self-driving cars:

- Improve congestion
- Even if only some are altruistic
- Without making others worse off
- Unambiguously Ethical?

Traffic

Total Time

3.25 hours

3 hours

3.25 hours

4 hours
Altruistic self-driving cars:

- Improve congestion
- Even if only some are altruistic
- Without making others worse off
- Unambiguously Ethical?

Traffic

Total Time

3.25 hours  3 hours  3.25 hours  4 hours
Self-driving car design

119 minutes

15 minutes per car

1 hour
Self-driving car design

119 minutes

15 minutes per car

1 hour

Selfish Traffic

Total Time

8 hours
Altruistic self-driving cars?

120 minutes

119 minutes

15 minutes per car

1 hour

Selfish Traffic

8 hours

Total Time

8 hours
Altruistic self-driving cars?

119 minutes

15 minutes per car

1 hour

Selfish Traffic

8 hours

Total Time

8 hours
Altruistic self-driving cars?

Selfish Traffic

119 minutes

15 minutes per car

1 hour

1 Altruistic Car

Total Time

8 hours

9 hours

Self-driving car design

Total Time

8 hours
Altruistic self-driving cars?

119 minutes

15 minutes per car

8 hours

9 hours

Total Time

Selfish Traffic

1 Altruistic Car

1 hour
Altruistic self-driving cars?

- Selfish Traffic: 8 hours
- 1 Altruistic Car: 9 hours
- 4 Altruistic Cars: 12 hours

Total Time: 119 minutes
15 minutes per car
Conclusions

Decision design for socio-technical systems

**Machine Policy**

- Selfish: Not Optimal
- Altruistic: Inconsiderate?

If **Ethical** means **utilitarian**, then sometimes altruism is good!

If **Ethical** means **risk-averse**, then machines should be selfish!
Traditional Ethics Message: Engineer things that don’t break

Needed Update: Engineer machines to interact well
What I’m doing
• Teaching CS4730/5730: Algorithmic Game Theory
• Designing CS4740: Social and Engineering Networks
  • Mathematics of interaction
• Public Lectures on responsible technological interaction

What the community can do
• Engineering curricula need integrated social science
• Assume that interaction $\rightarrow$ counterintuitive outcomes

**Needed Update:** Engineer machines to *interact* well