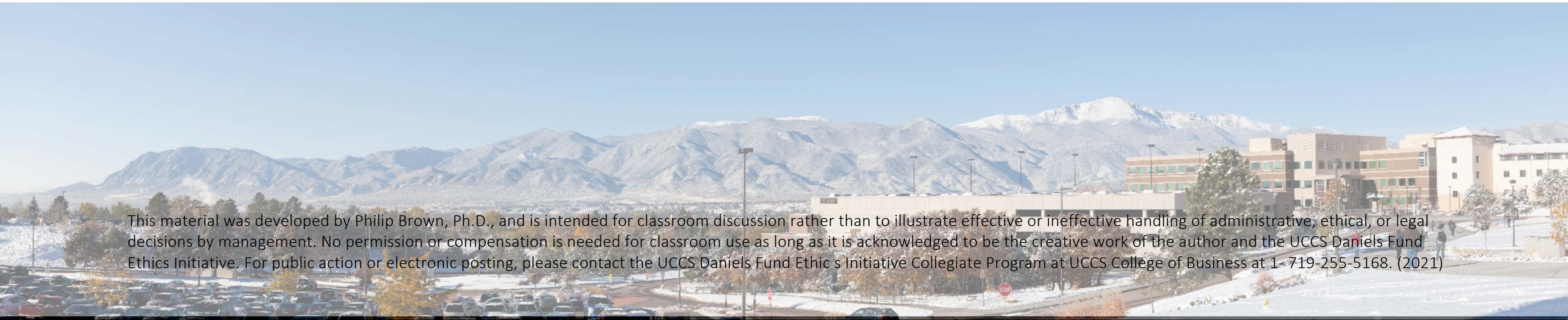
Engineers who make Selfish Machines

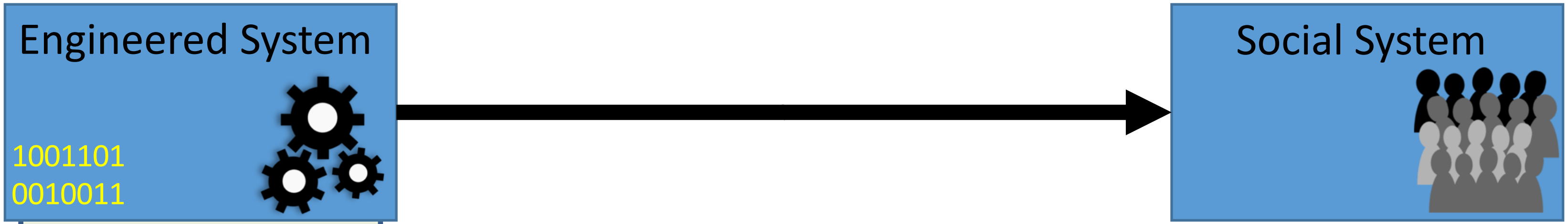
The Ethics of Socio-Technical Systems

Philip N. Brown
Dept of Computer Science
University of Colorado, Colorado Springs



This material was developed by Philip Brown, Ph.D., and is intended for classroom discussion rather than to illustrate effective or ineffective handling of administrative, ethical, or legal decisions by management. No permission or compensation is needed for classroom use as long as it is acknowledged to be the creative work of the author and the UCCS Daniels Fund Ethics Initiative. For public action or electronic posting, please contact the UCCS Daniels Fund Ethic s Initiative Collegiate Program at UCCS College of Business at 1- 719-255-5168. (2021)

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

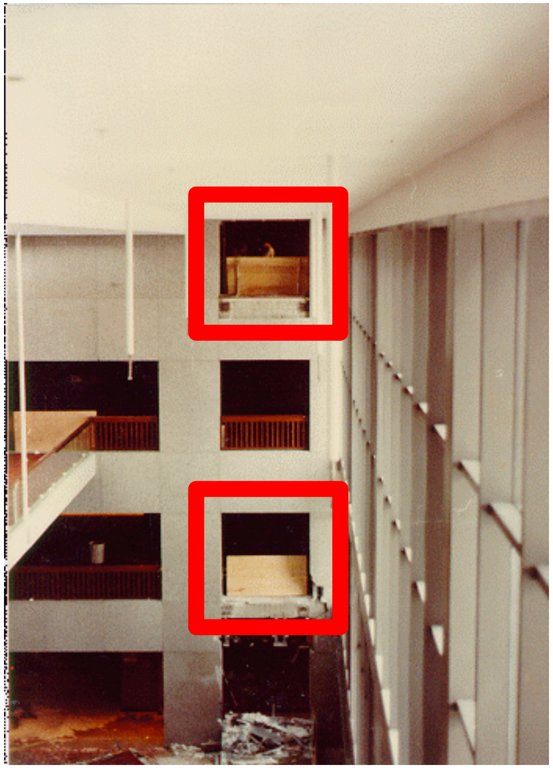
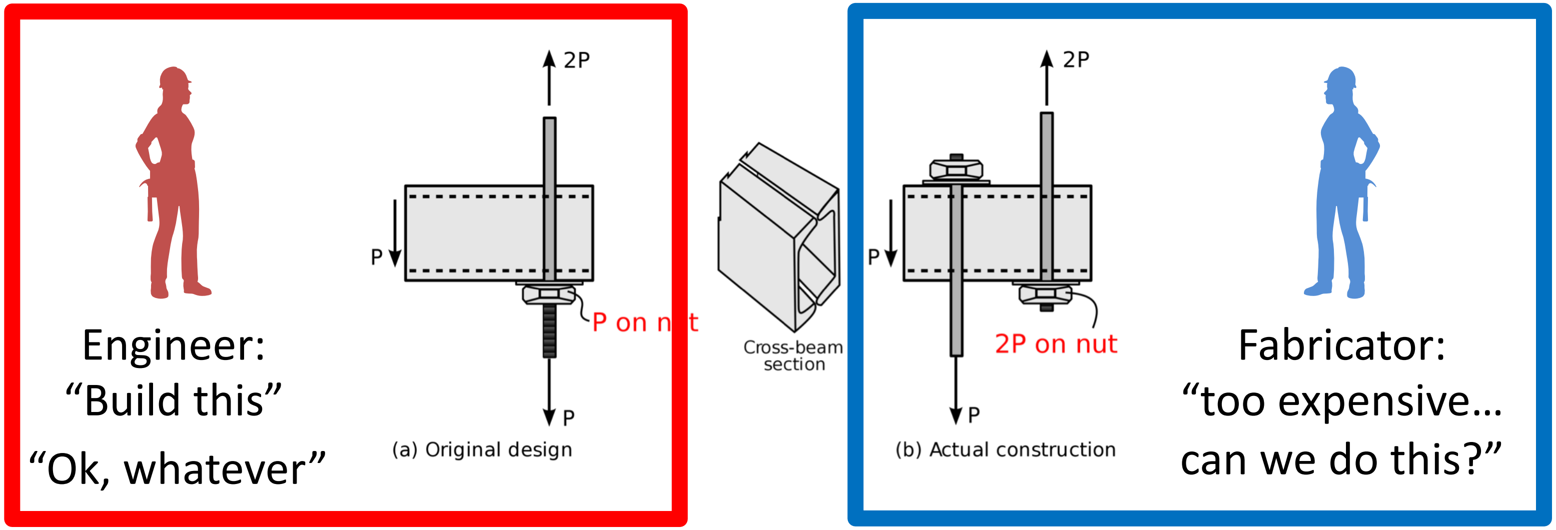


Image credit: Public Domain, Dr. Lee Lowery, Jr., P.E.

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”

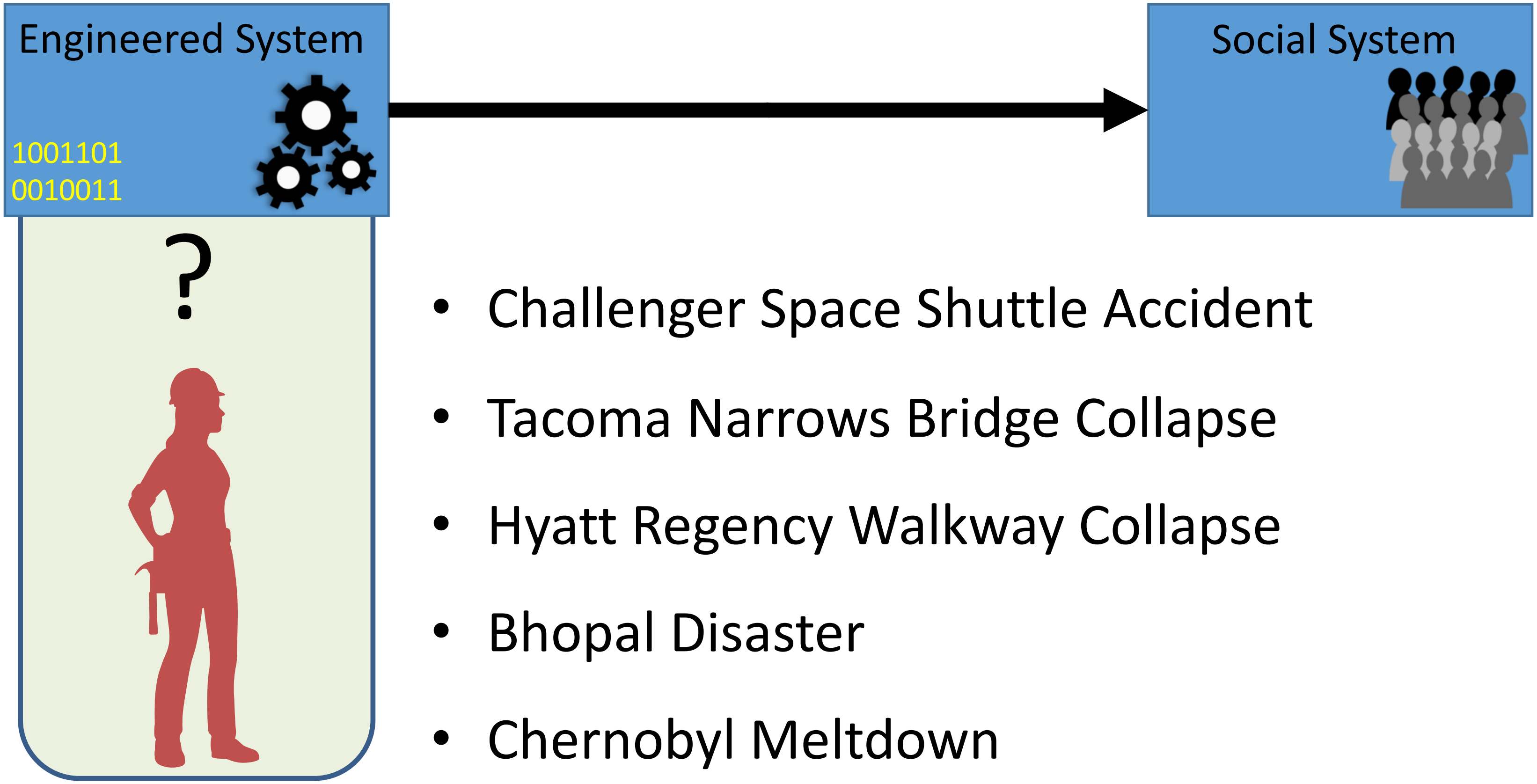
Fabricator:
“too expensive… can we do this?”

Only supports 60% of required load

Only supports **half** load of red design

Image credit: Public Domain, Wikimedia Commons

Traditional Engineering Ethics Training

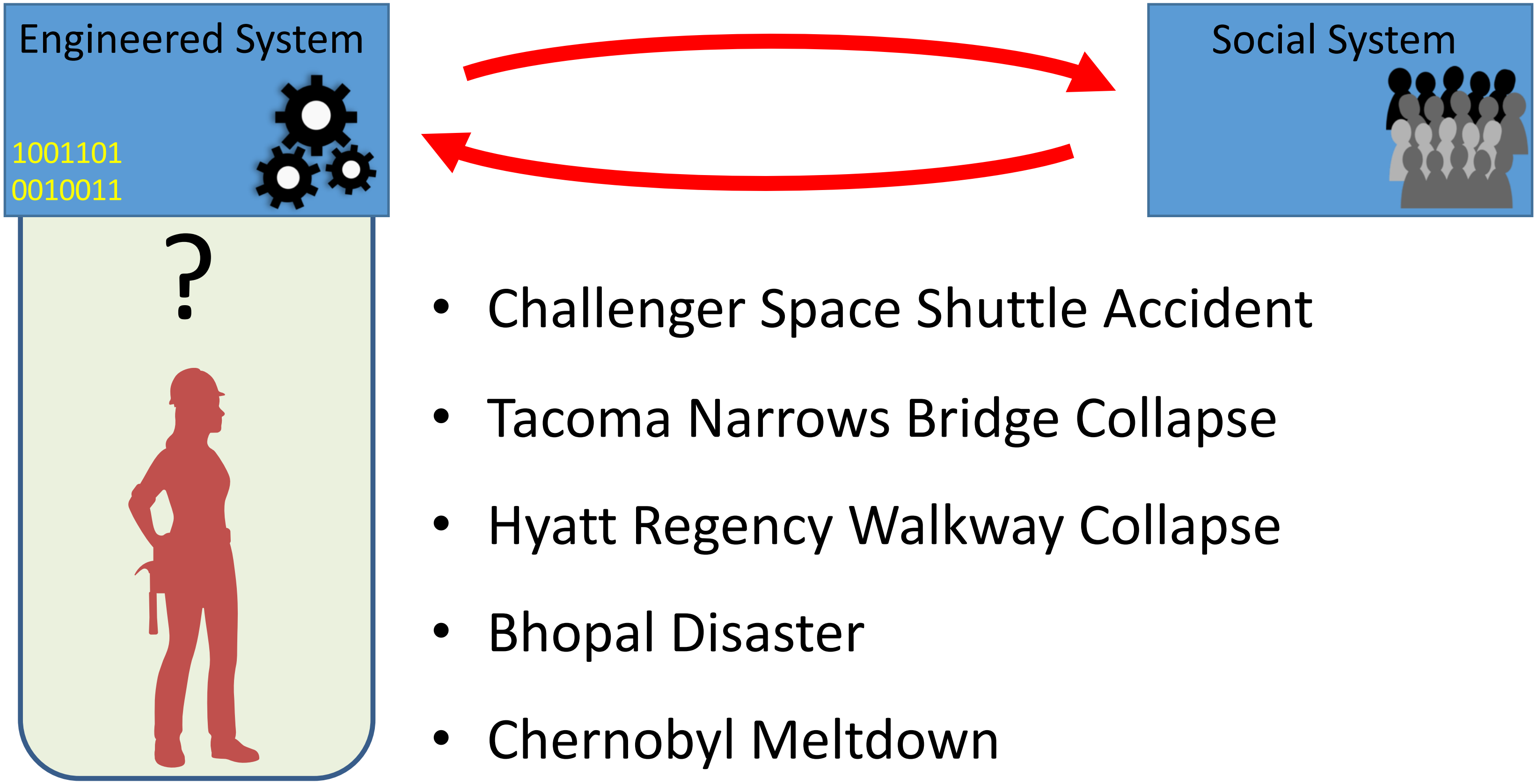


- Challenger Space Shuttle Accident
- Tacoma Narrows Bridge Collapse
- Hyatt Regency Walkway Collapse
- Bhopal Disaster
- Chernobyl Meltdown

Ethics Message: Engineer things that don’t break

Implication: Adequate design is sufficient.

Traditional Engineering Ethics Training

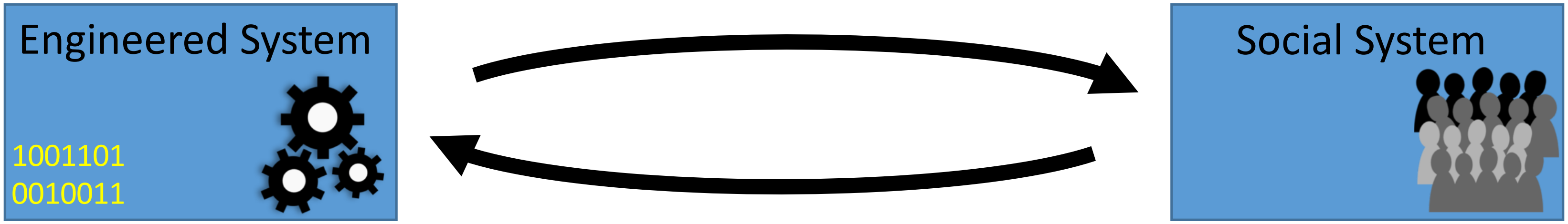


- Challenger Space Shuttle Accident
- Tacoma Narrows Bridge Collapse
- Hyatt Regency Walkway Collapse
- Bhopal Disaster
- Chernobyl Meltdown

Ethics Message: Engineer things that don’t break

Implication: Adequate design is sufficient.

Socio-Technical Paradigm

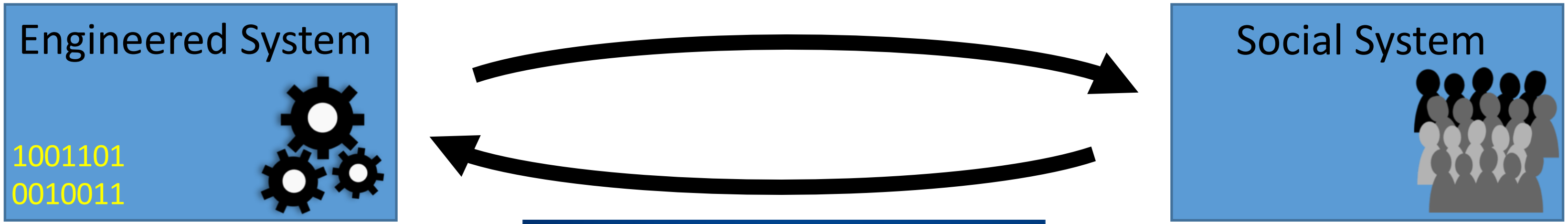


Engineered System

Social System

ebay

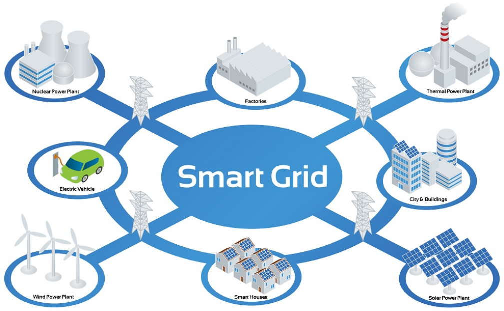
Socio-Technical Paradigm



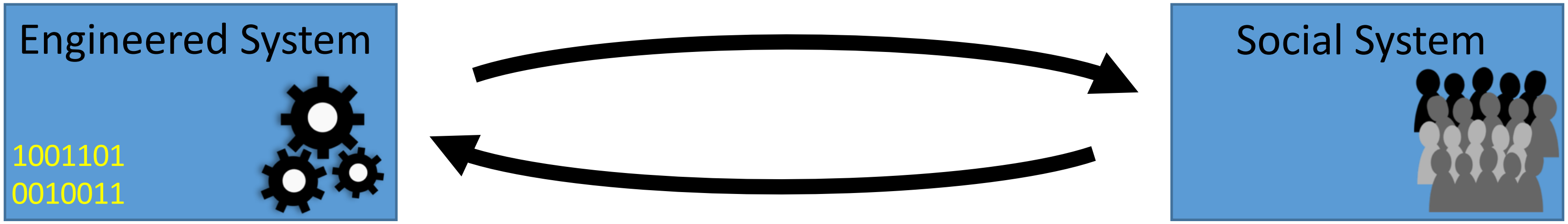
Engineered System

1001101
0010011

Social System

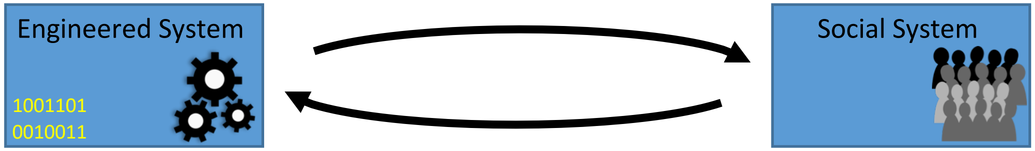


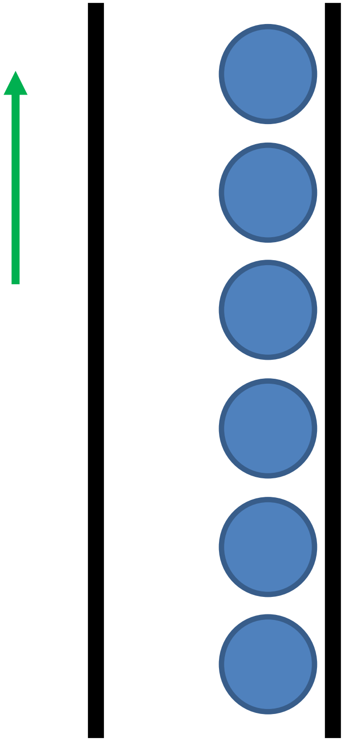
Socio-Technical Paradigm



- Why does the feedback loop matter?
- What are ethical implications?
- How to teach this?

Escalator Example

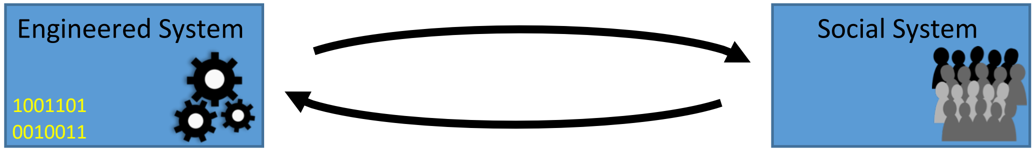




Stand on right

Image credit: Public Domain, Pixabay

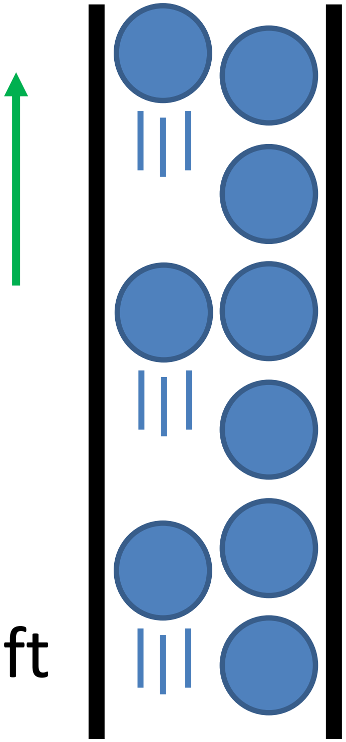
Escalator Example



Engineered System

1001101
0010011

Social System

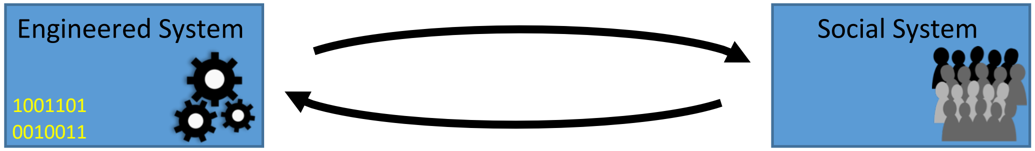


Walk on left

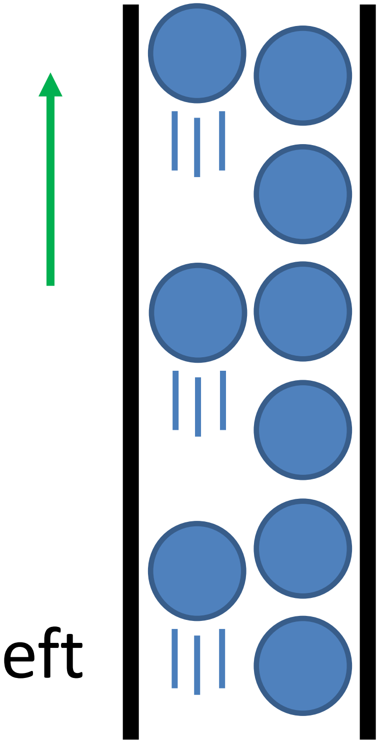
Stand on right

Image credit: Public Domain, Pixabay

Escalator Example

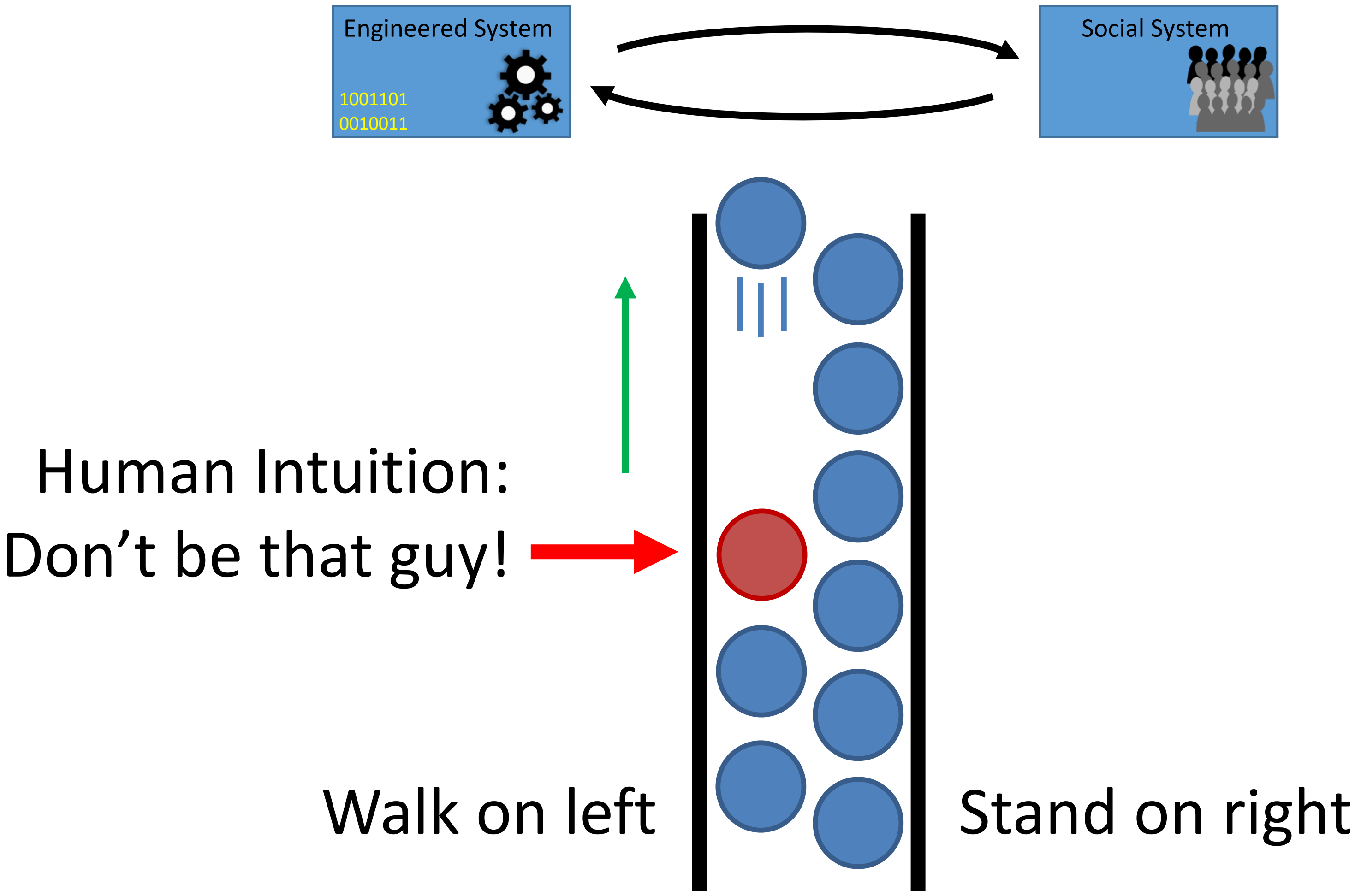

Engineered System

Social System


Walk on left

Stand on right

Escalator Example

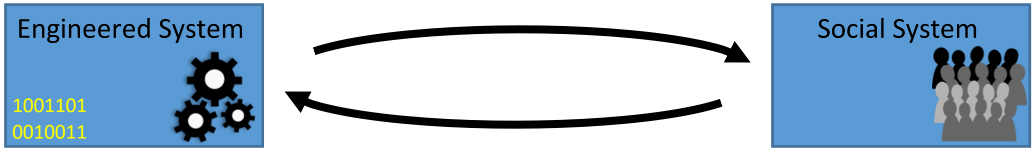


Human Intuition:
Don’t be that guy!

Walk on left

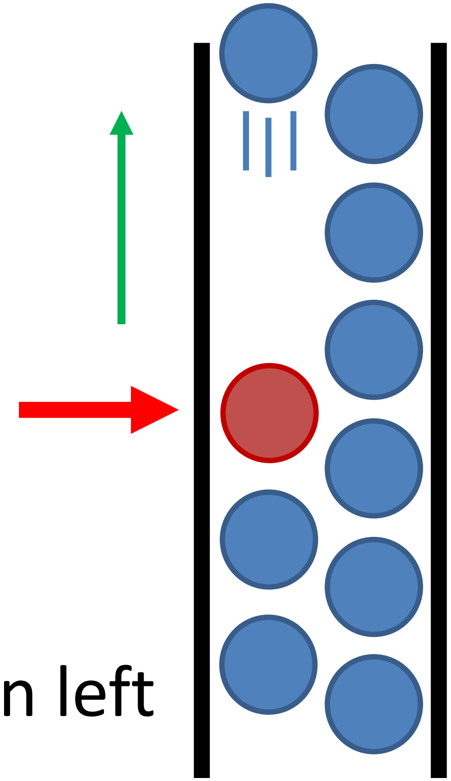
Stand on right

Escalator Example



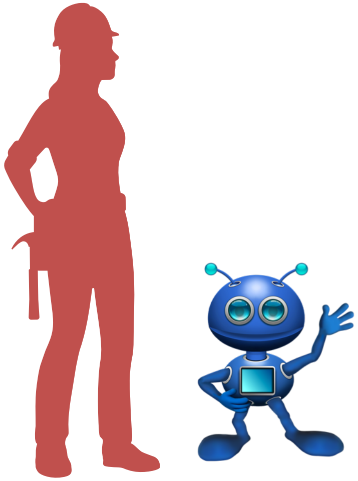
Human Intuition:

Don’t be that guy!



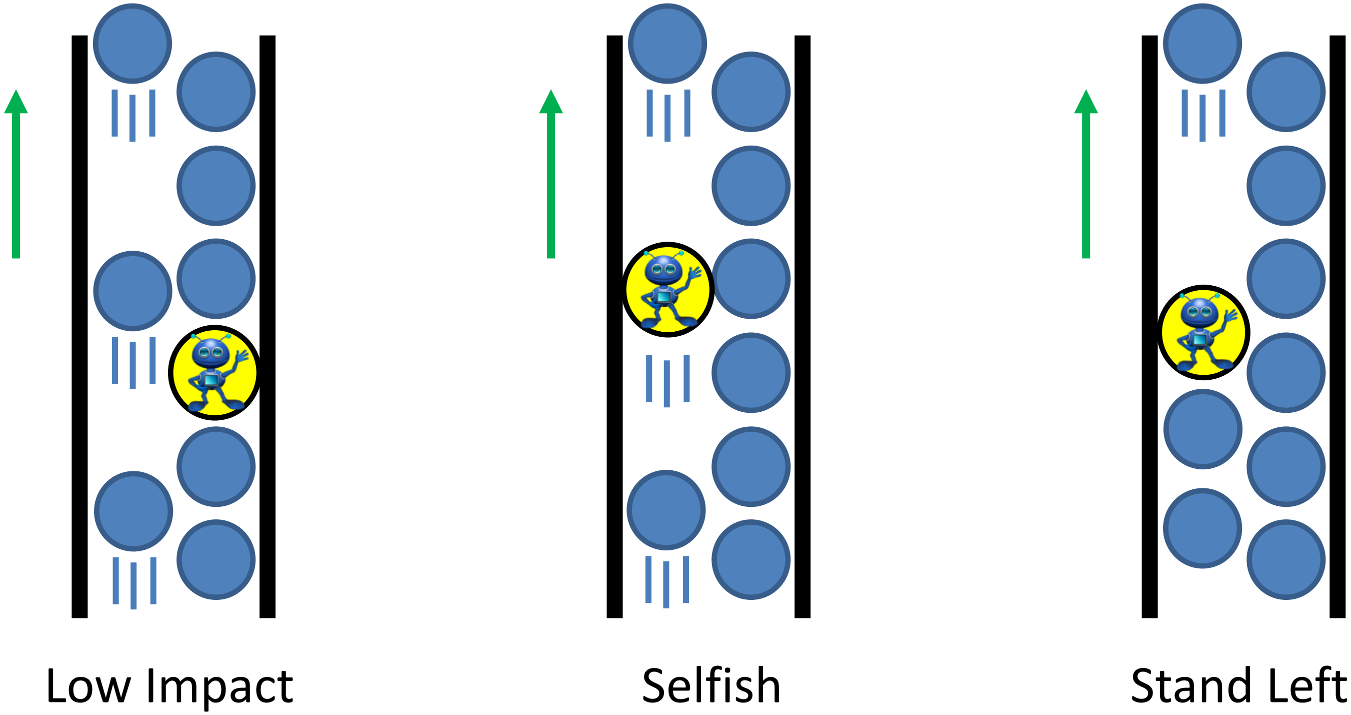
Walk on left

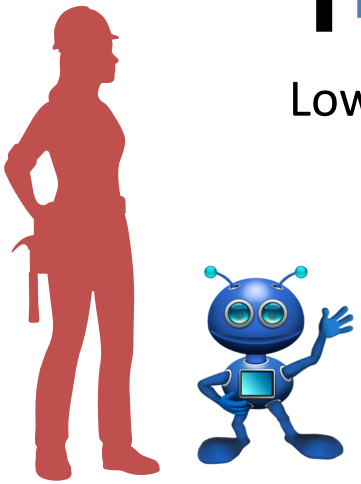
Stand on right



Question: Robot’s escalator policy?

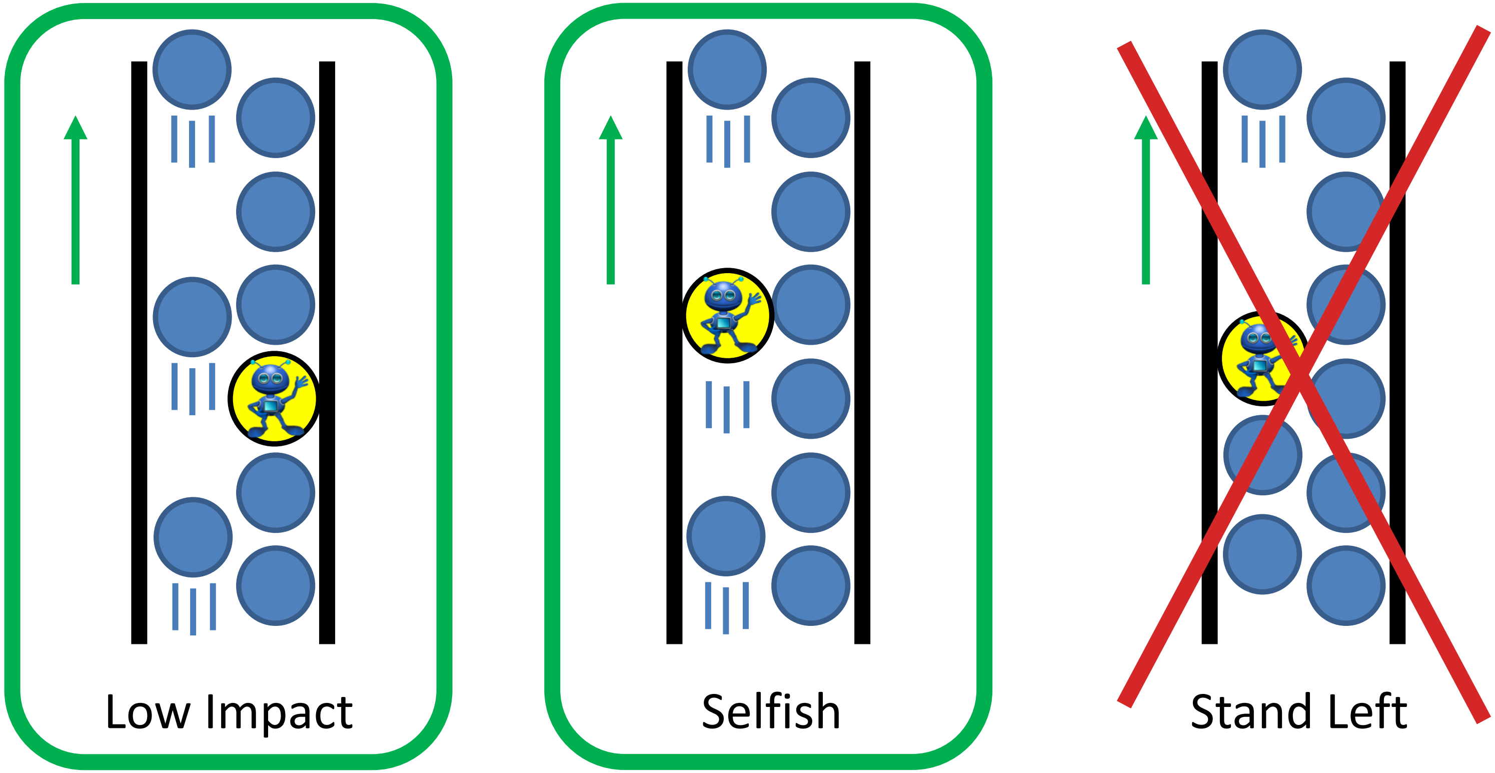
Escalator Example



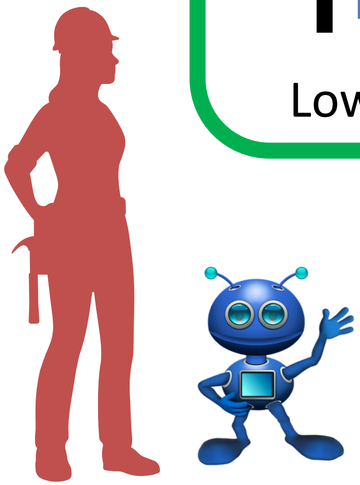


Question: Robot’s escalator policy?

Escalator Example

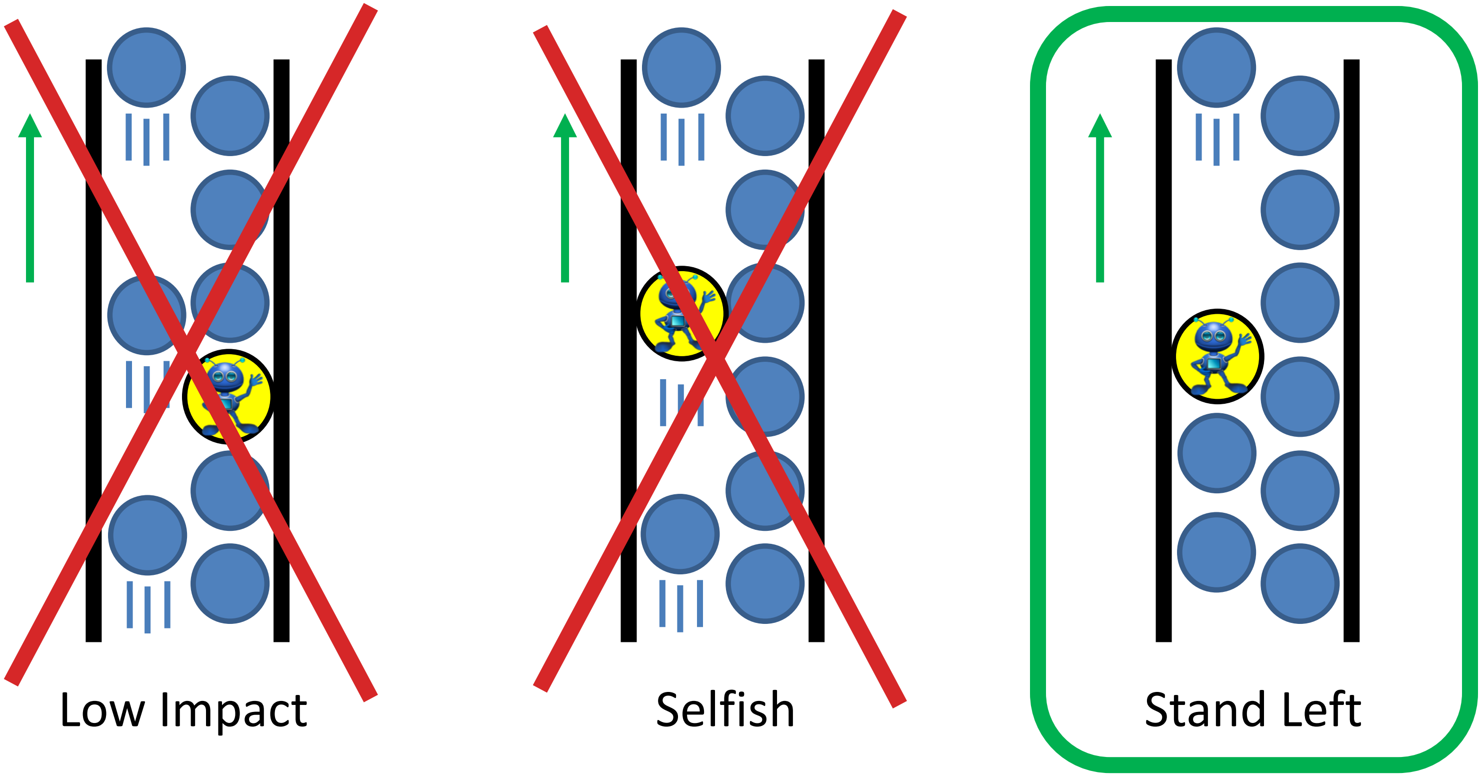


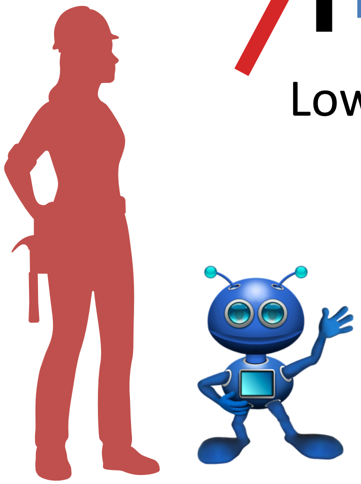
Human Intuition:
not this



Question: Robot’s escalator policy?

Escalator Example

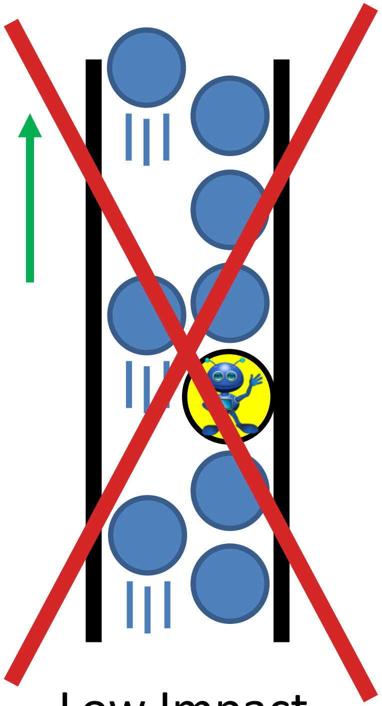




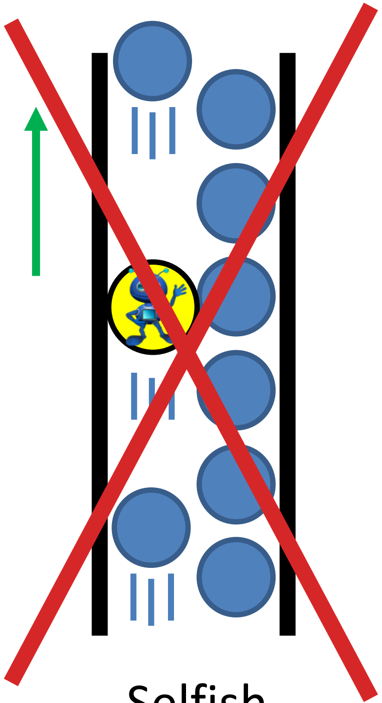
Systems Perspective: definitely this (maximizes throughput)

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

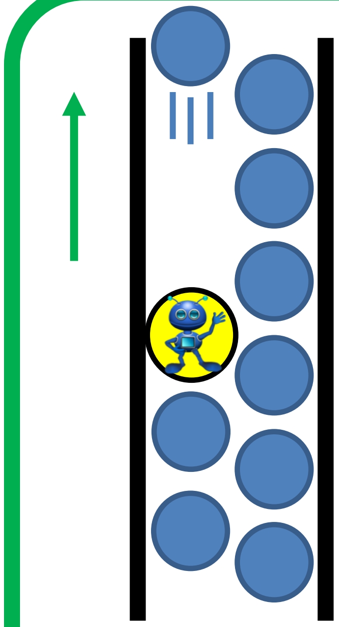
Escalator Example



Low Impact

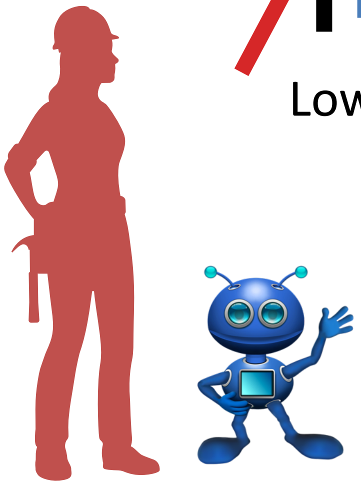


Selfish



Altruistic

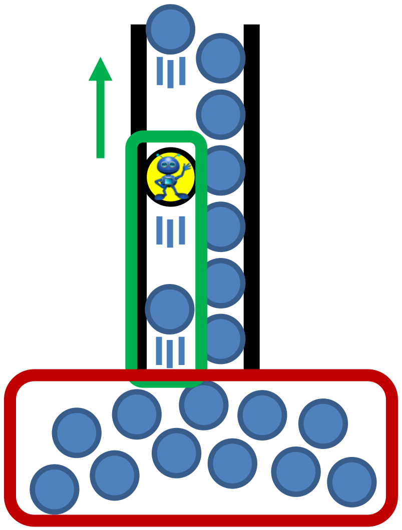
Systems Perspective: definitely this (maximizes throughput)



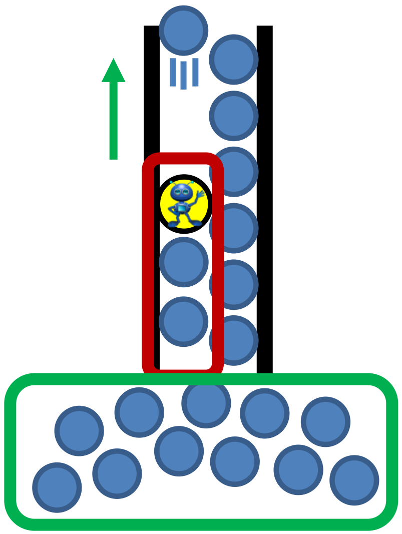
For more info: Prof. Lesley Strawderman at Mississippi State University

Escalator Example

Selfish



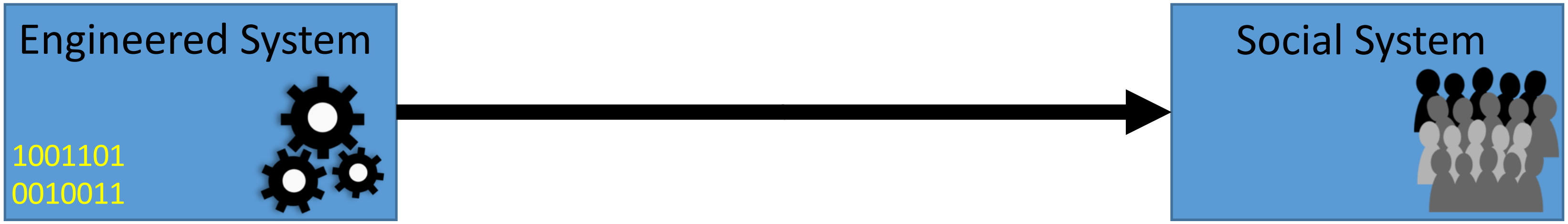
Altruistic



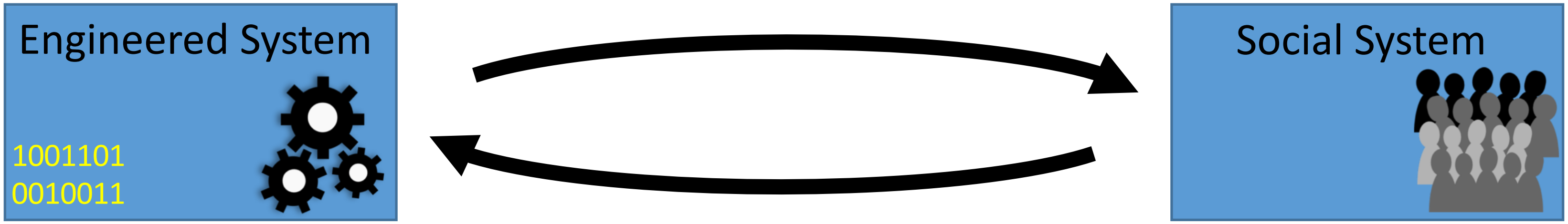
Benefits:	Robot a few people behind	Everybody waiting to use escalator
Harms:	Everybody waiting to use escalator	Robot a few people behind
Ethical?	Bad for “system” Robot acts like a human	Good for “system” Robot looks like a jerk!

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

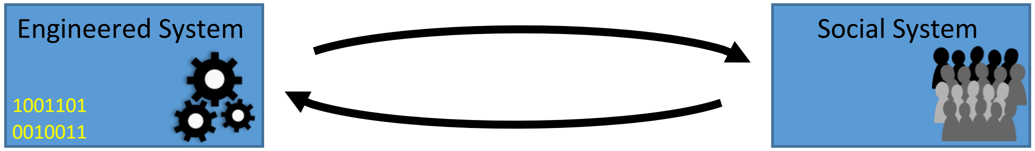
Broader Picture



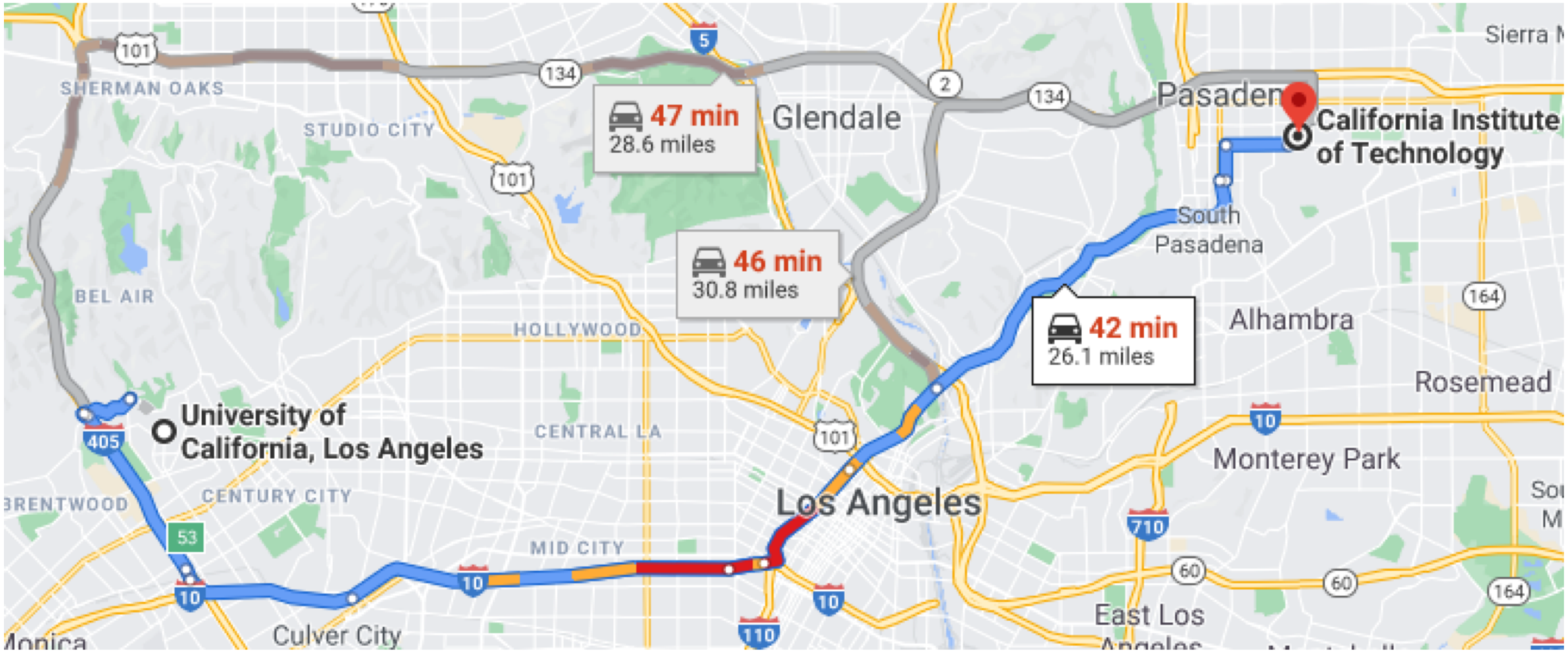
Traditional Ethics Message: Engineer things that don’t break

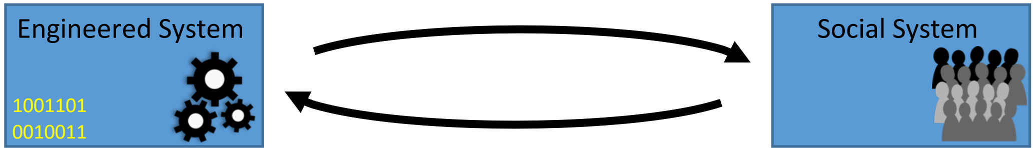


Needed Update: Engineer your machines to ***interact*** with people

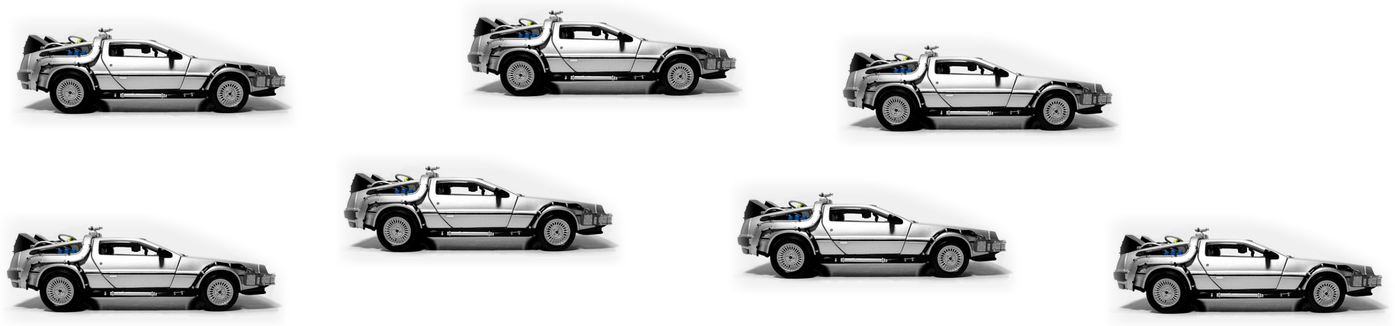


Choosing Routes in Highway Networks



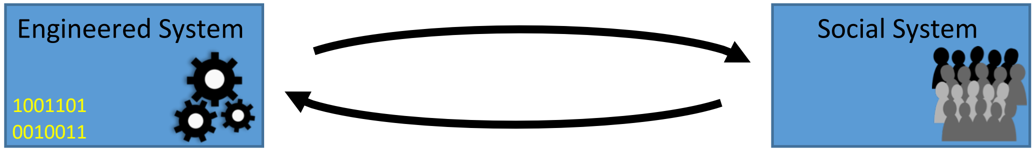


Choosing Routes in Highway Networks



Agenda: pose simple model

Explore ethics in its context



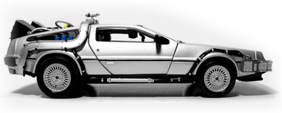
Choosing Routes in Highway Networks

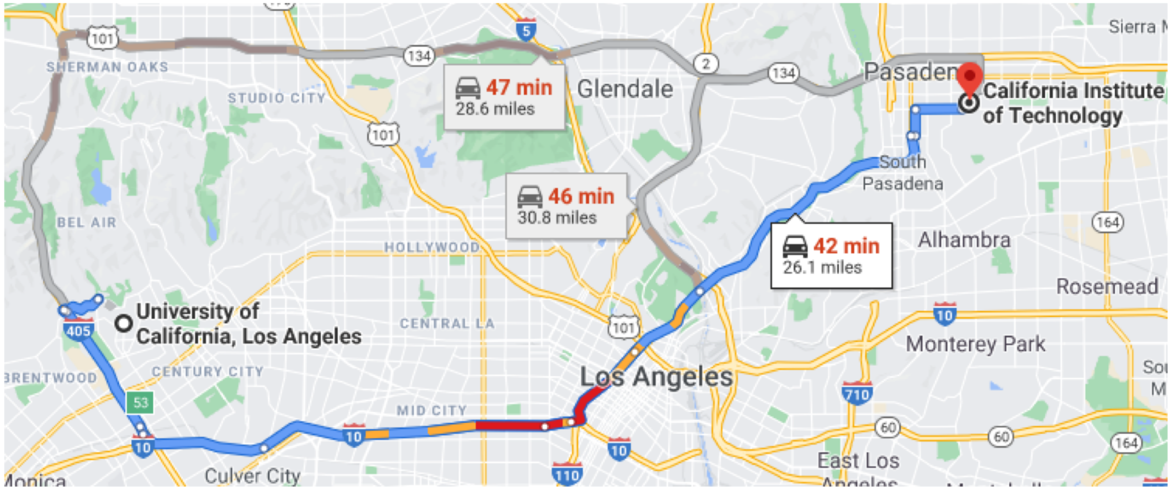
Agenda: pose simple model

Explore ethics in its context

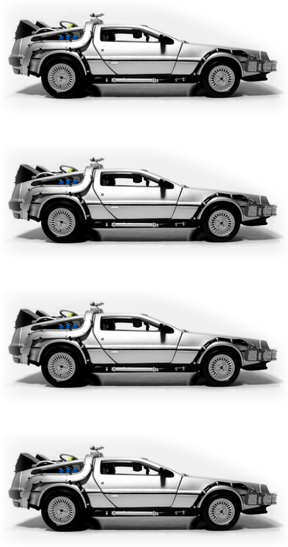
Question: should self-driving cars be altruistic?

Self-driving car design



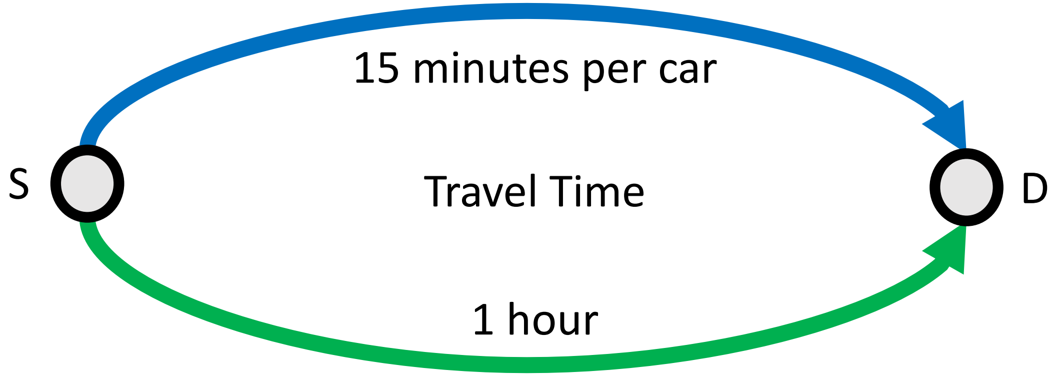


Self-driving car design



Fast, if empty

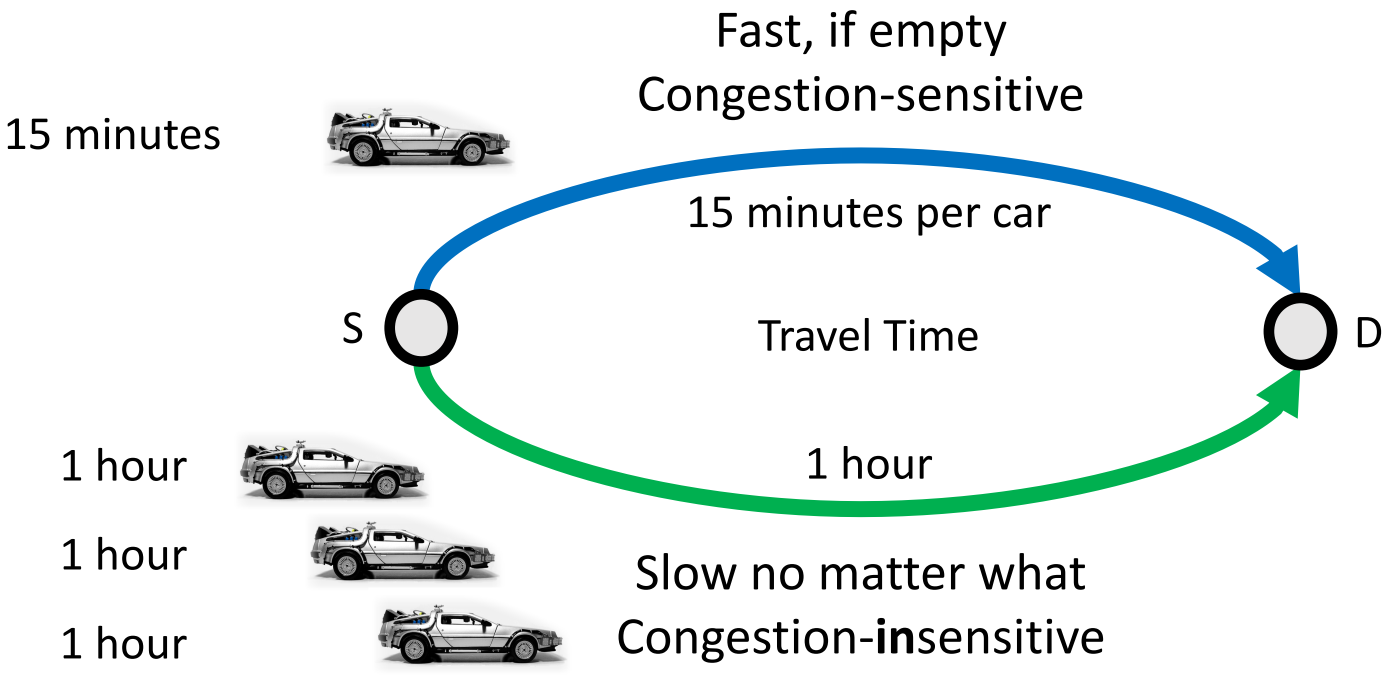
Congestion-sensitive



Slow no matter what

Congestion-**in**sensitive

Self-driving car design



15 minutes

Fast, if empty
Congestion-sensitive

15 minutes per car

S

Travel Time

D

1 hour

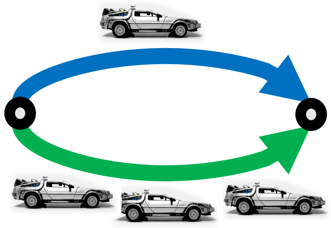
1 hour

1 hour

1 hour

Slow no matter what
Congestion-**in**sensitive

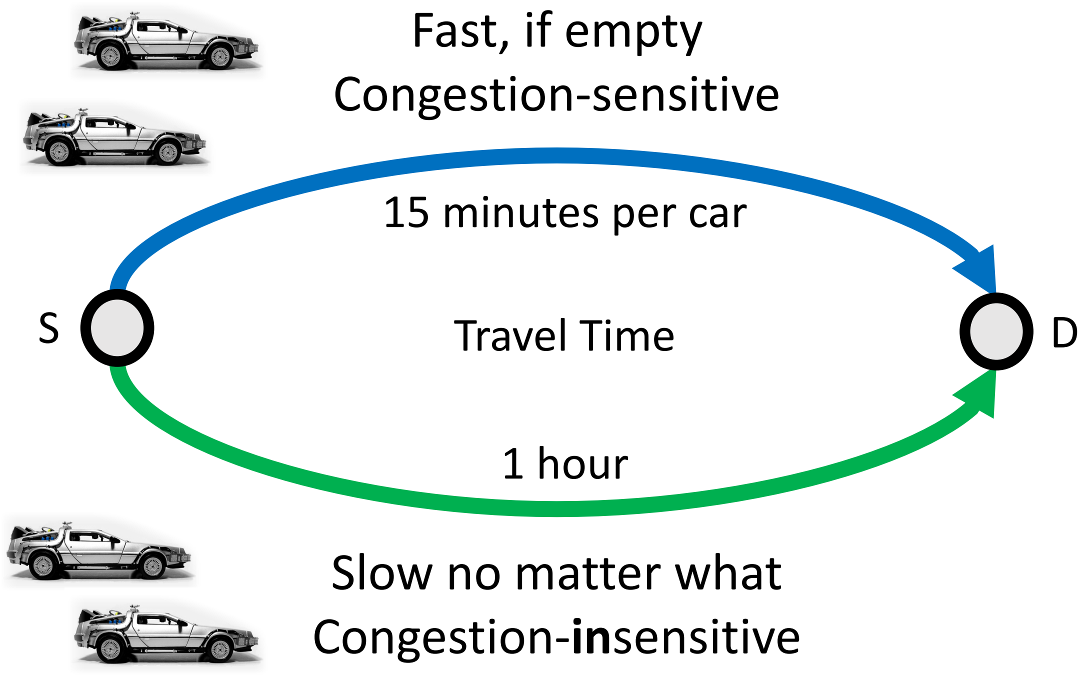
Traffic



Total
Time

3.25 hours

Self-driving car design



30 minutes

30 minutes

Fast, if empty

Congestion-sensitive

15 minutes per car

S

Travel Time

D

1 hour

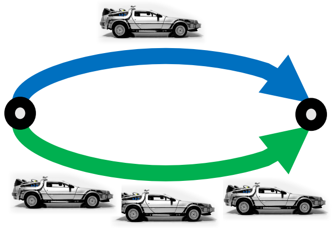
1 hour

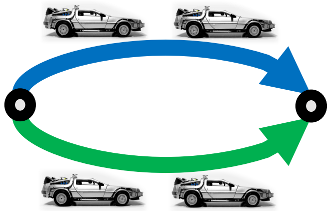
1 hour

Slow no matter what

Congestion-**in**sensitive

Traffic



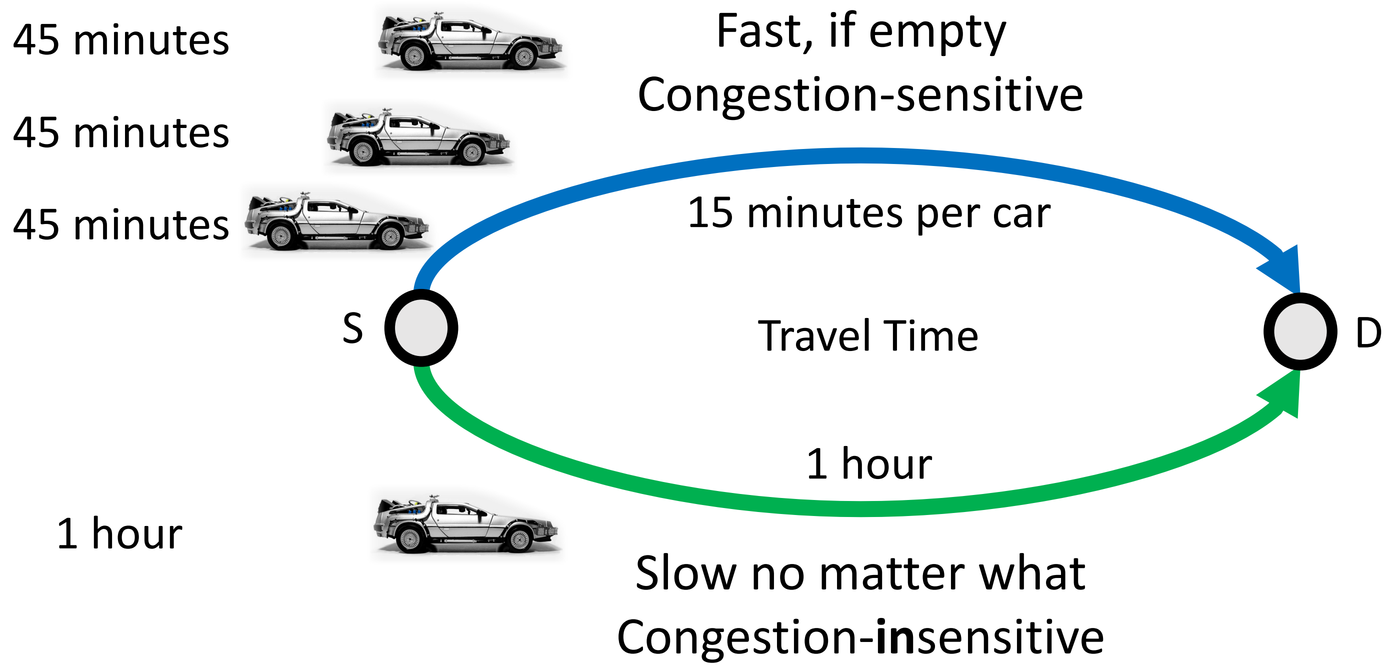


Total Time

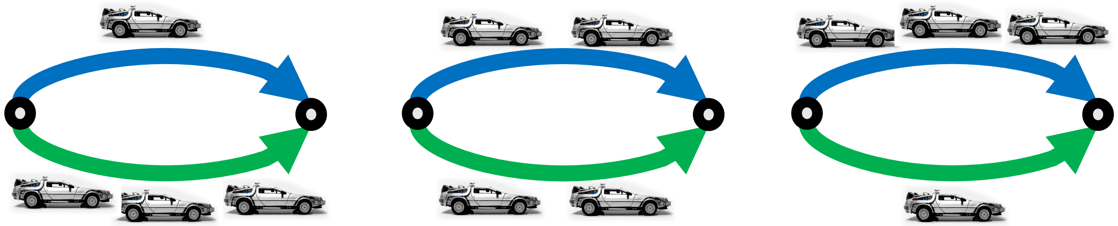
3.25 hours

3 hours

Self-driving car design

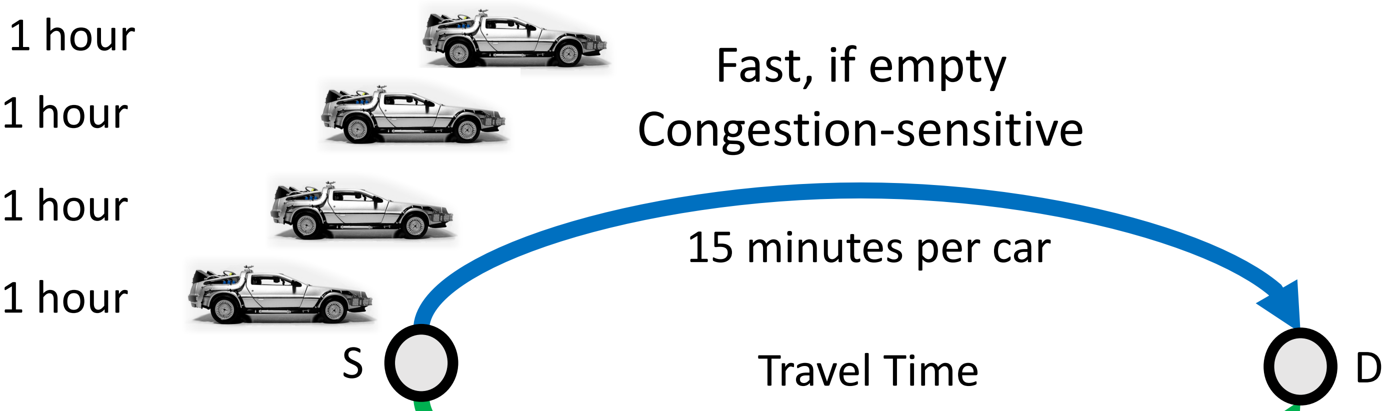


Traffic



Total Time	3.25 hours	3 hours	3.25 hours

Self-driving car design



1 hour

1 hour

1 hour

1 hour

Fast, if empty

Congestion-sensitive

15 minutes per car

S

Travel Time

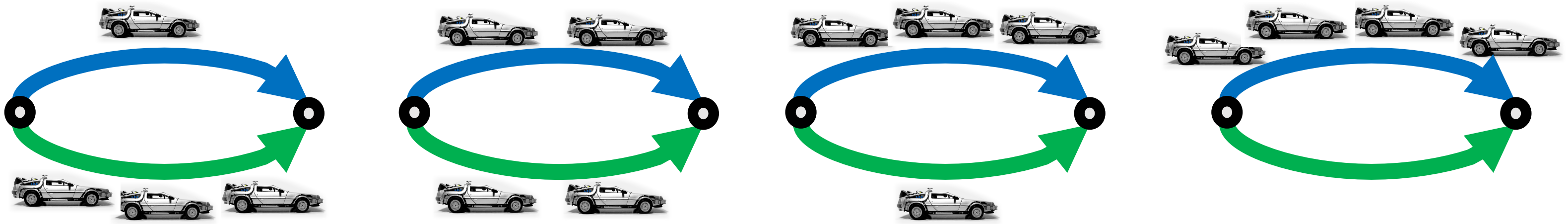
D

1 hour

Slow no matter what

Congestion-**in**sensitive

Traffic



Total Time

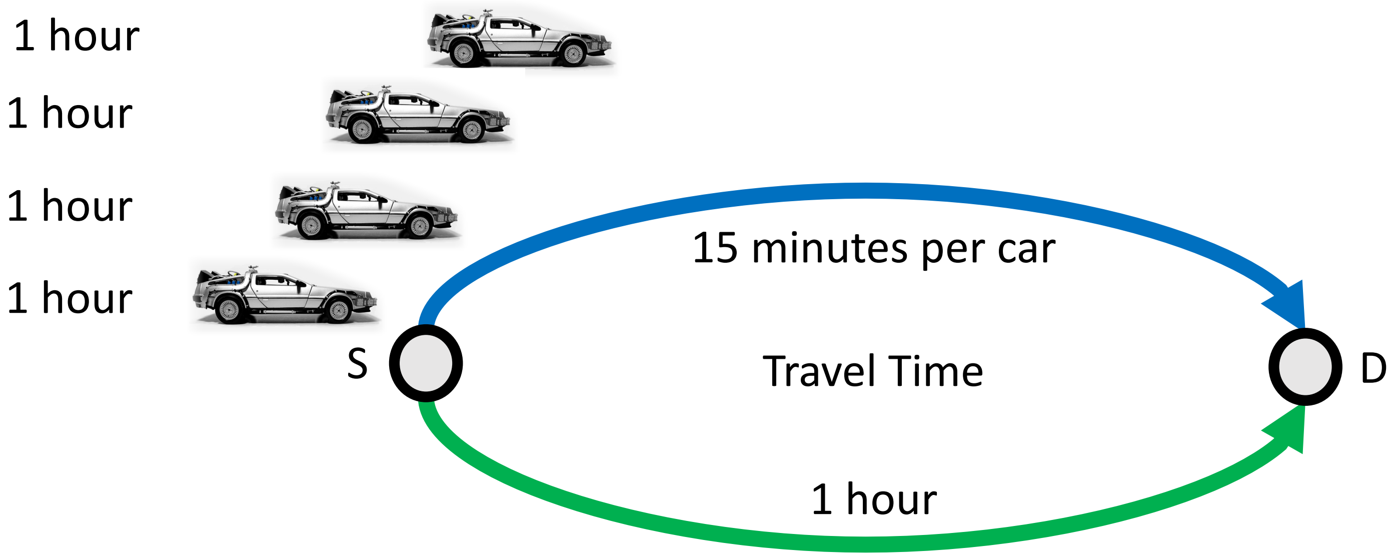
3.25 hours

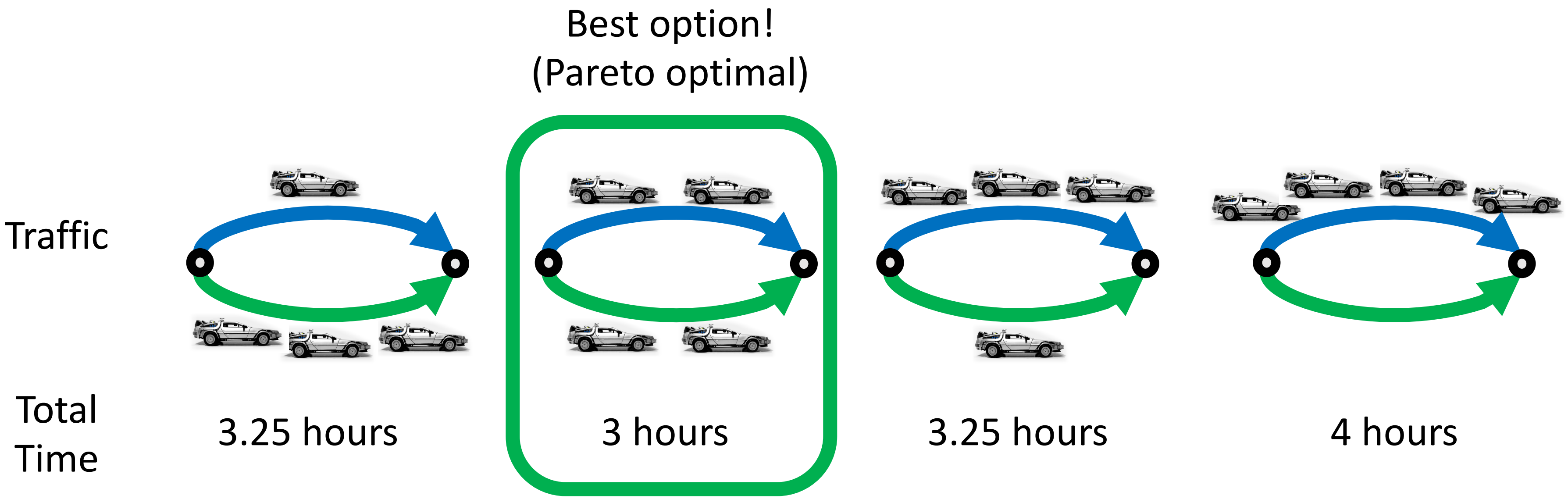
3 hours

3.25 hours

4 hours

Self-driving car design





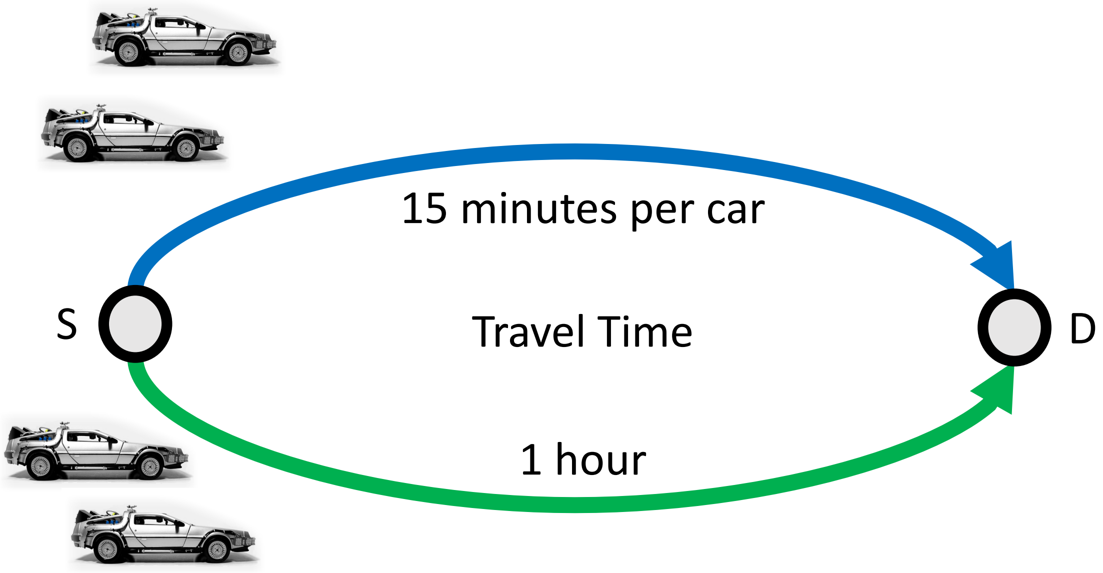
Self-driving car design

30 minutes

30 minutes

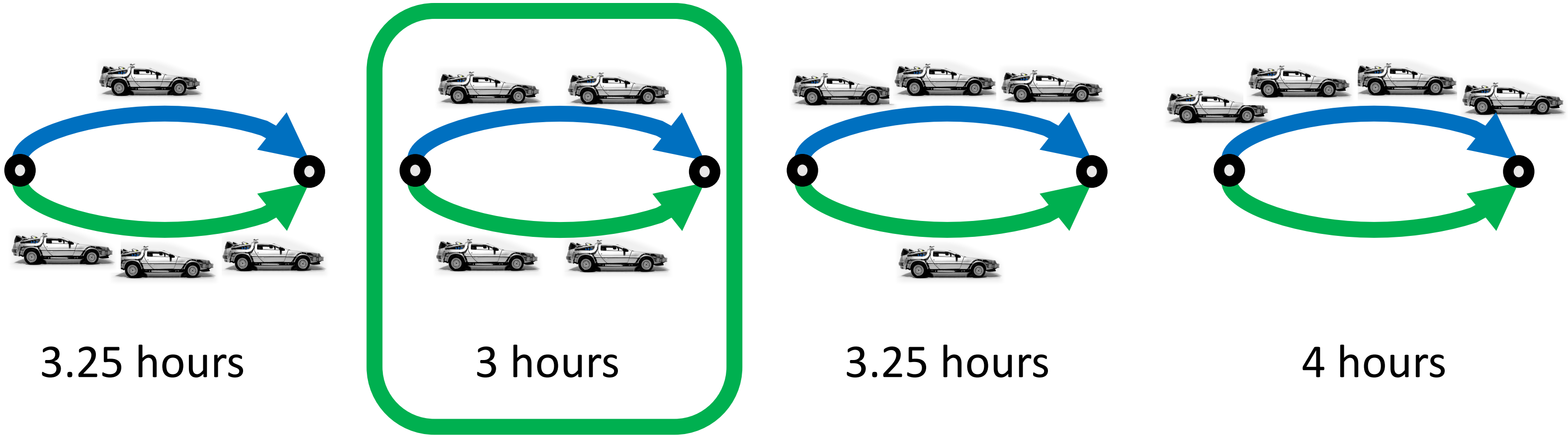
1 hour

1 hour



Best option!
(Pareto optimal)

Traffic



Total Time

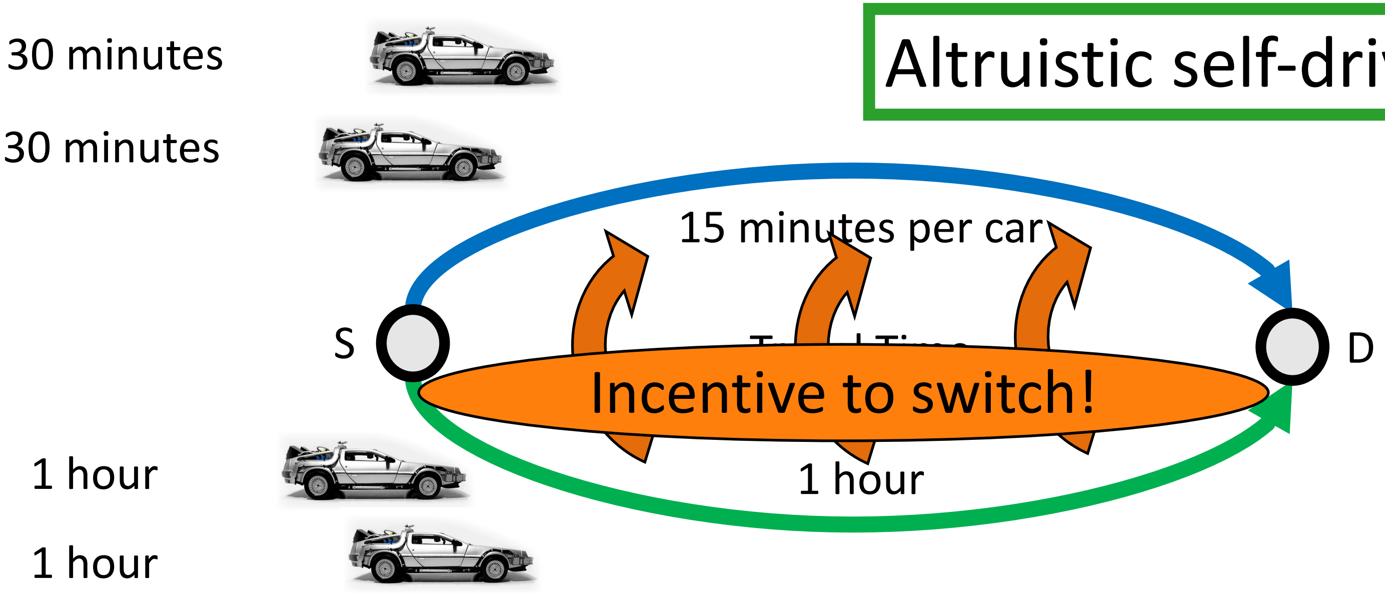
3.25 hours

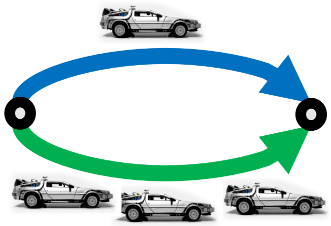
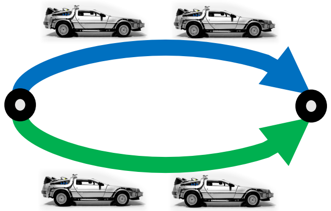
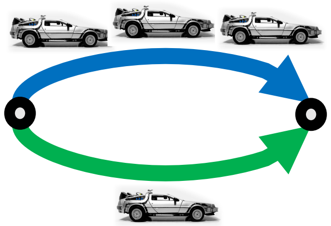
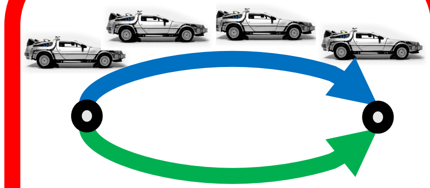
3 hours

3.25 hours

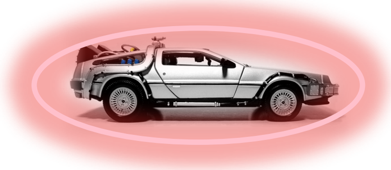
4 hours

Altruistic self-driving cars?



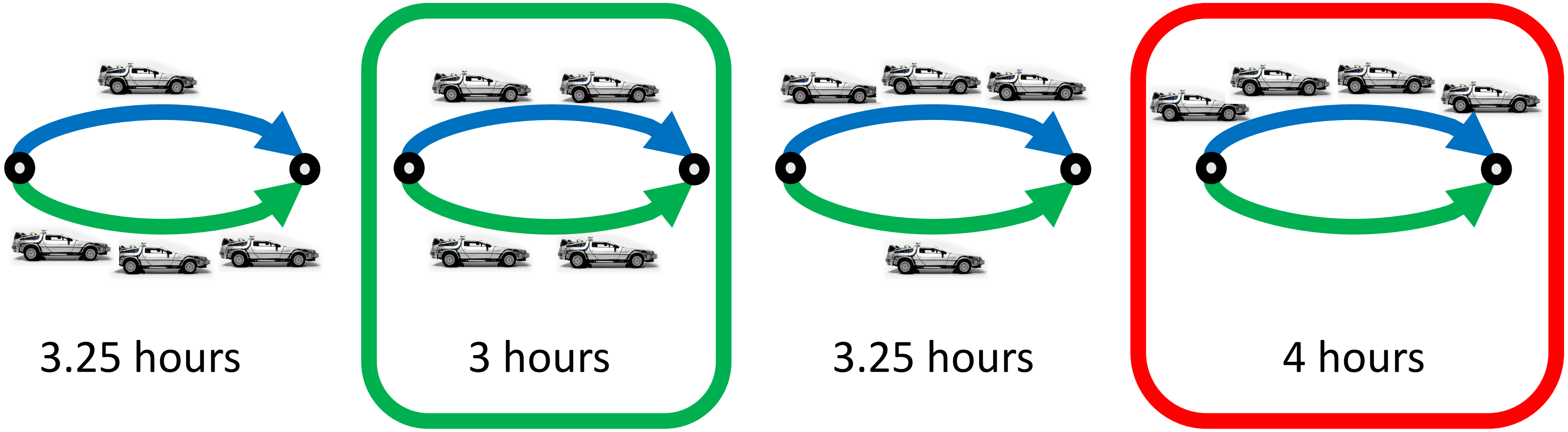
		Best option! (Pareto optimal)		Selfish traffic is like this!
Traffic				
Total Time	3.25 hours	3 hours	3.25 hours	4 hours

Altruistic self-driving cars?



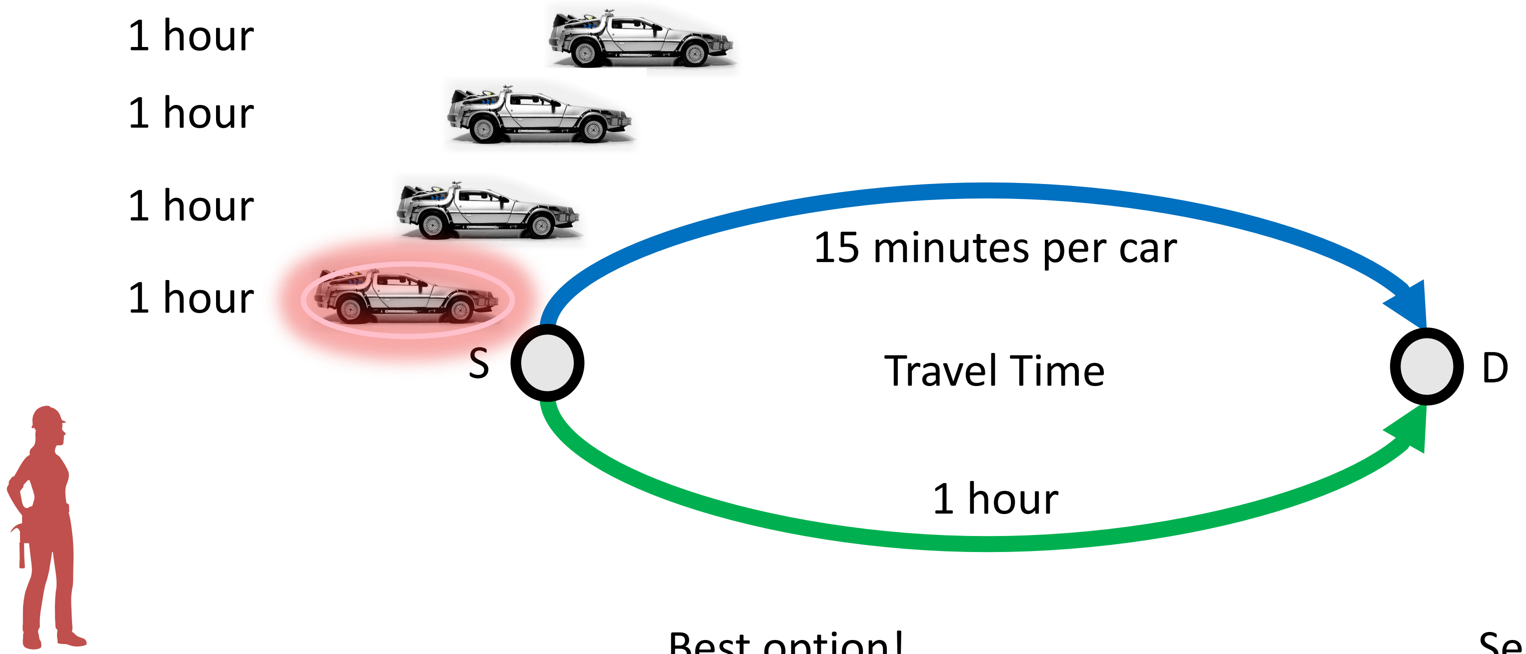
Altruism: act like there is 2x actual traffic

Traffic



		Best option! (Pareto optimal)		Selfish traffic is like this!
Total Time	3.25 hours	3 hours	3.25 hours	4 hours

Self-driving car design



1 hour

1 hour

1 hour

1 hour

S

15 minutes per car

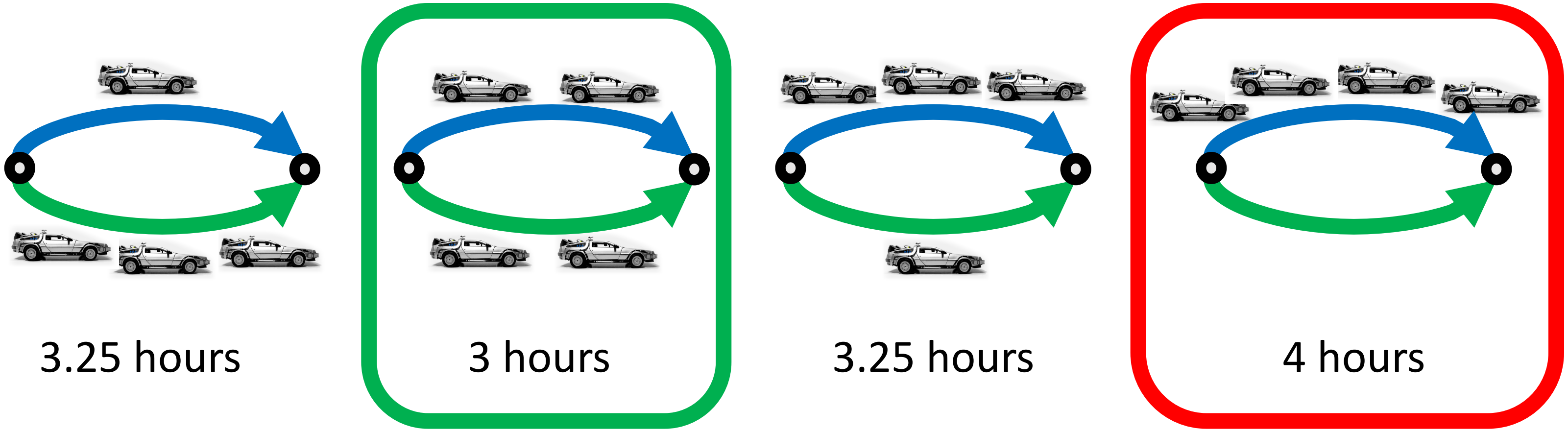
Travel Time

D

1 hour

Best option!
(Pareto optimal)

Selfish traffic
is like this!



Traffic

Total Time

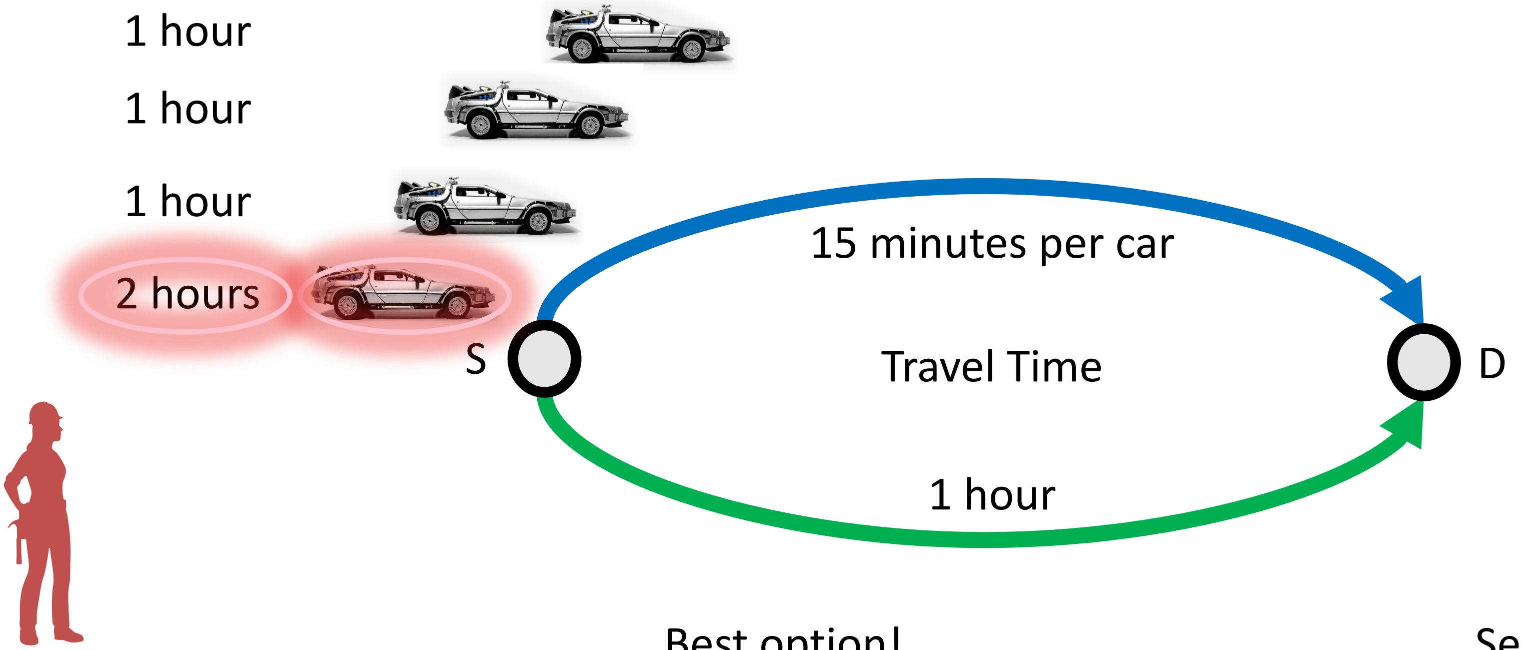
3.25 hours

3 hours

3.25 hours

4 hours

Self-driving car design



1 hour

1 hour

1 hour

2 hours

S

15 minutes per car

Travel Time

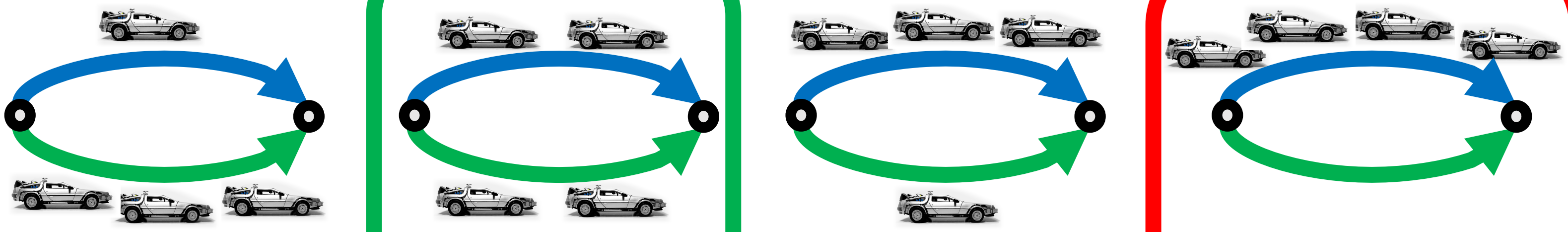
1 hour

D

Best option!
(Pareto optimal)

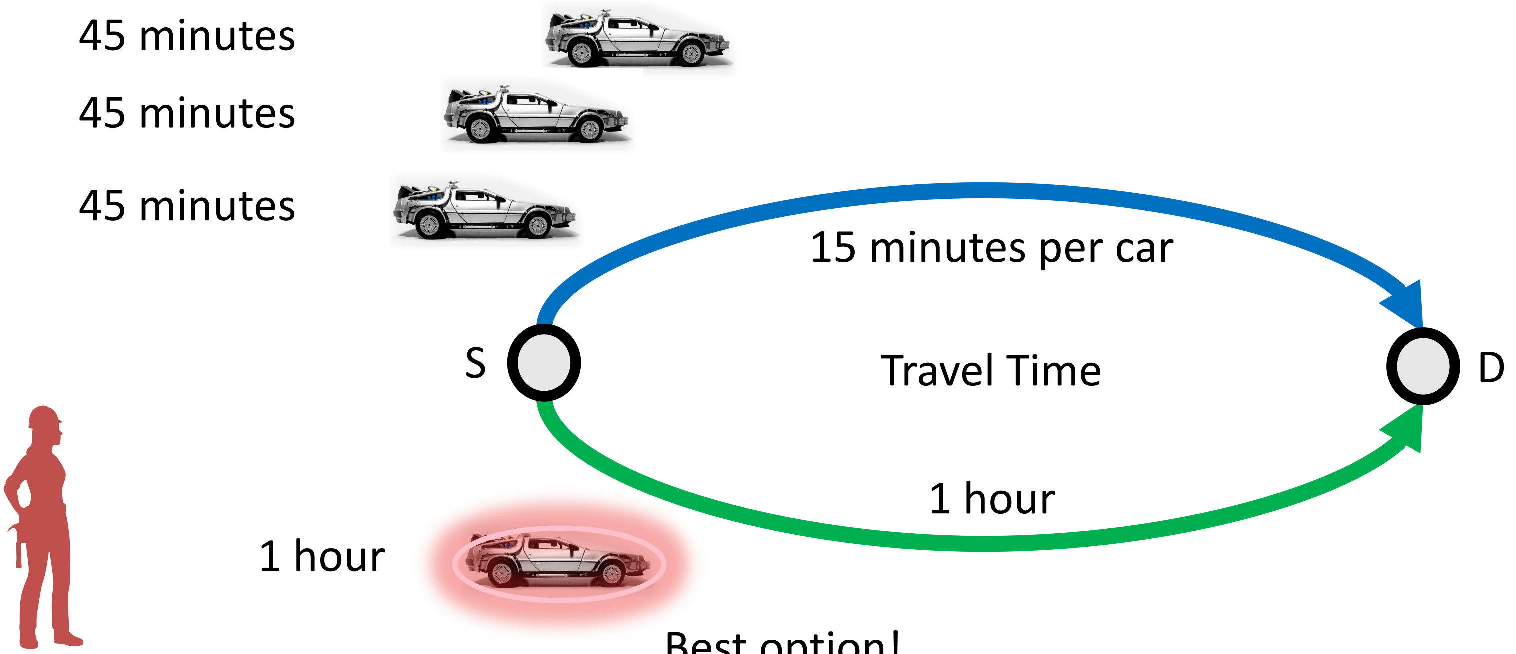
Selfish traffic
is like this!

Traffic



Total Time	3.25 hours	3 hours	3.25 hours	4 hours

Self-driving car design



45 minutes

45 minutes

45 minutes

15 minutes per car

S

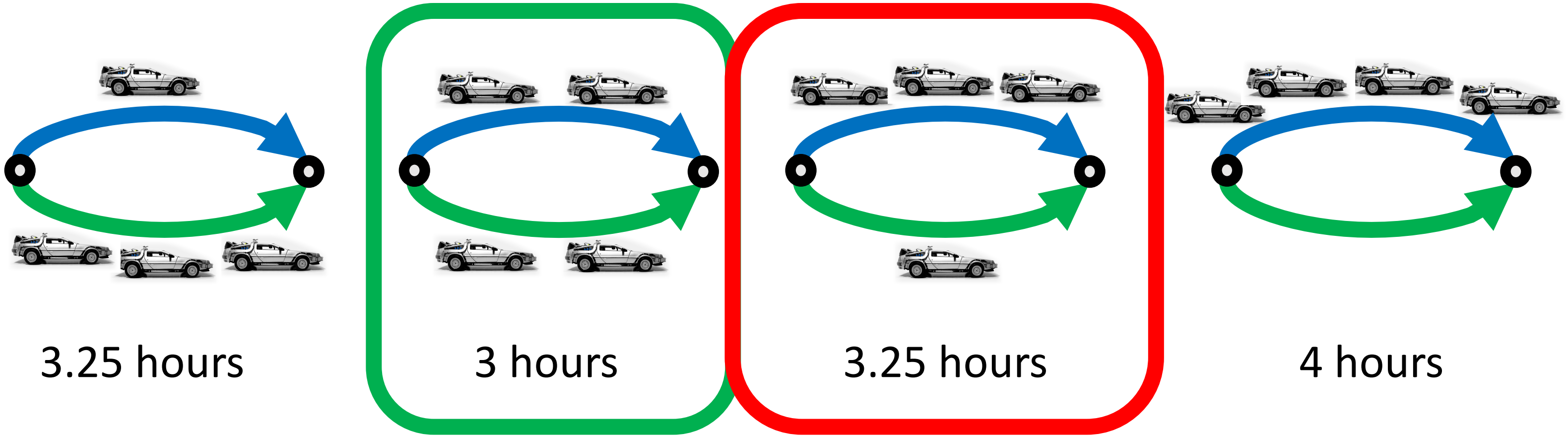
Travel Time

D

1 hour

1 hour

Best option!
(Pareto optimal)



Traffic

Total Time

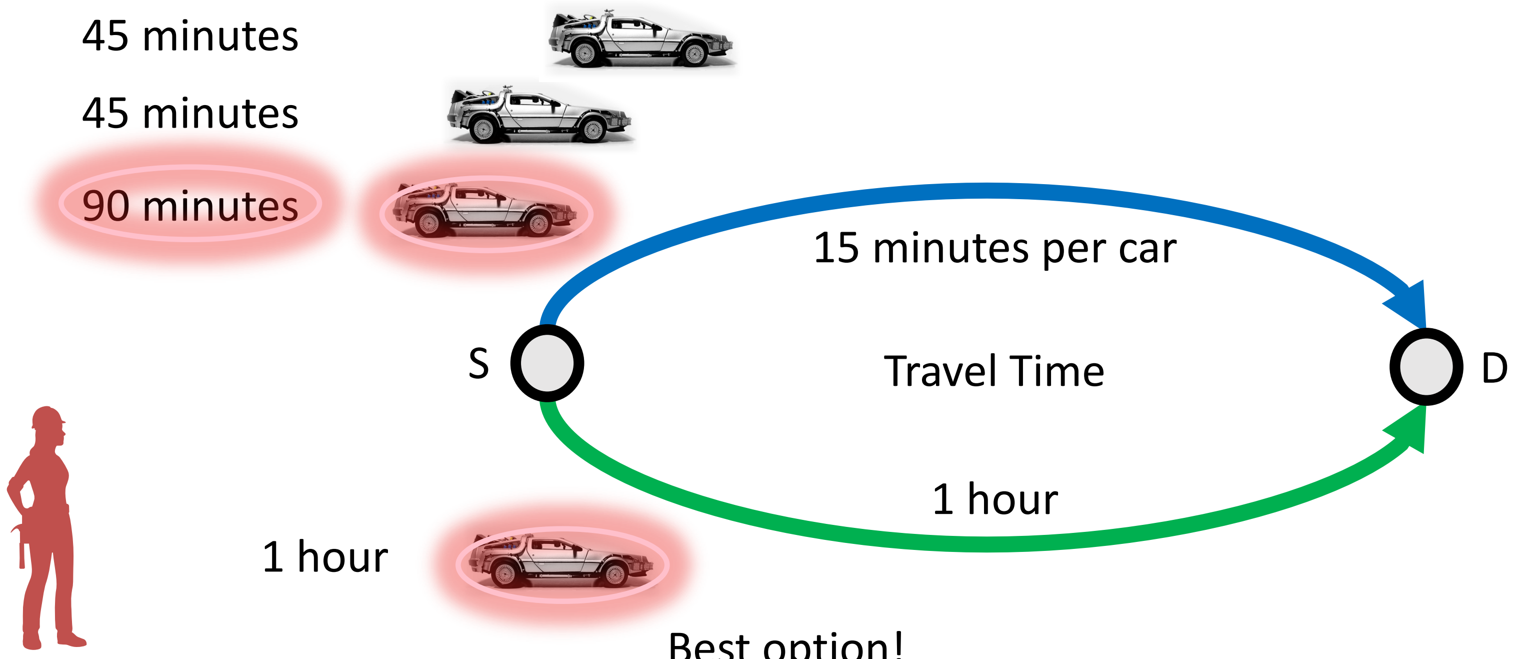
3.25 hours

3 hours

3.25 hours

4 hours

Self-driving car design



45 minutes

45 minutes

90 minutes

S

15 minutes per car

Travel Time

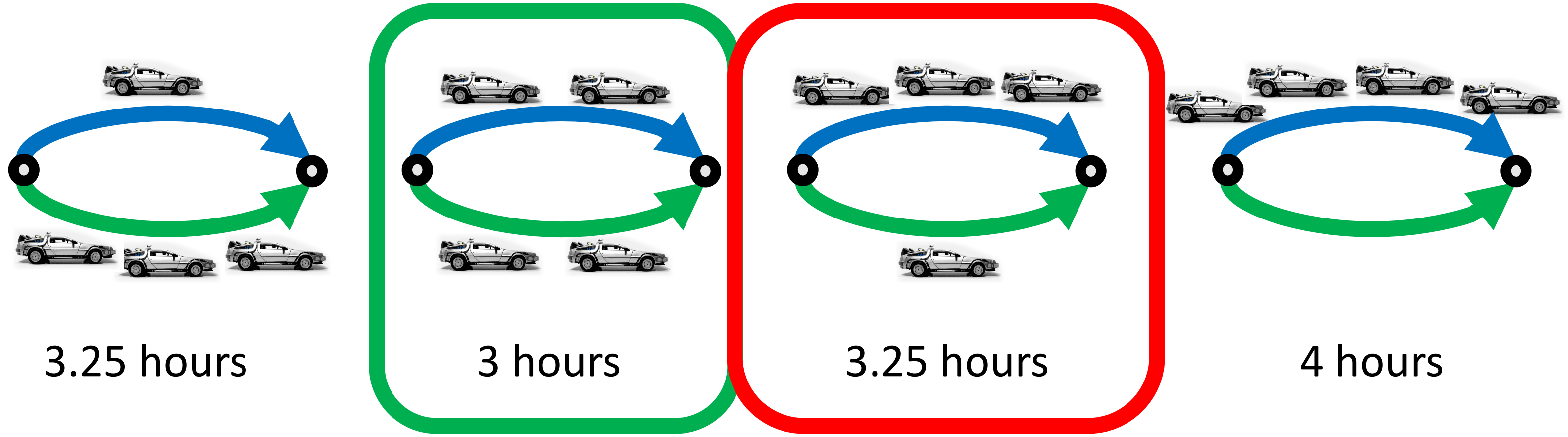
D

1 hour

1 hour

Best option!
(Pareto optimal)

Traffic



Total Time

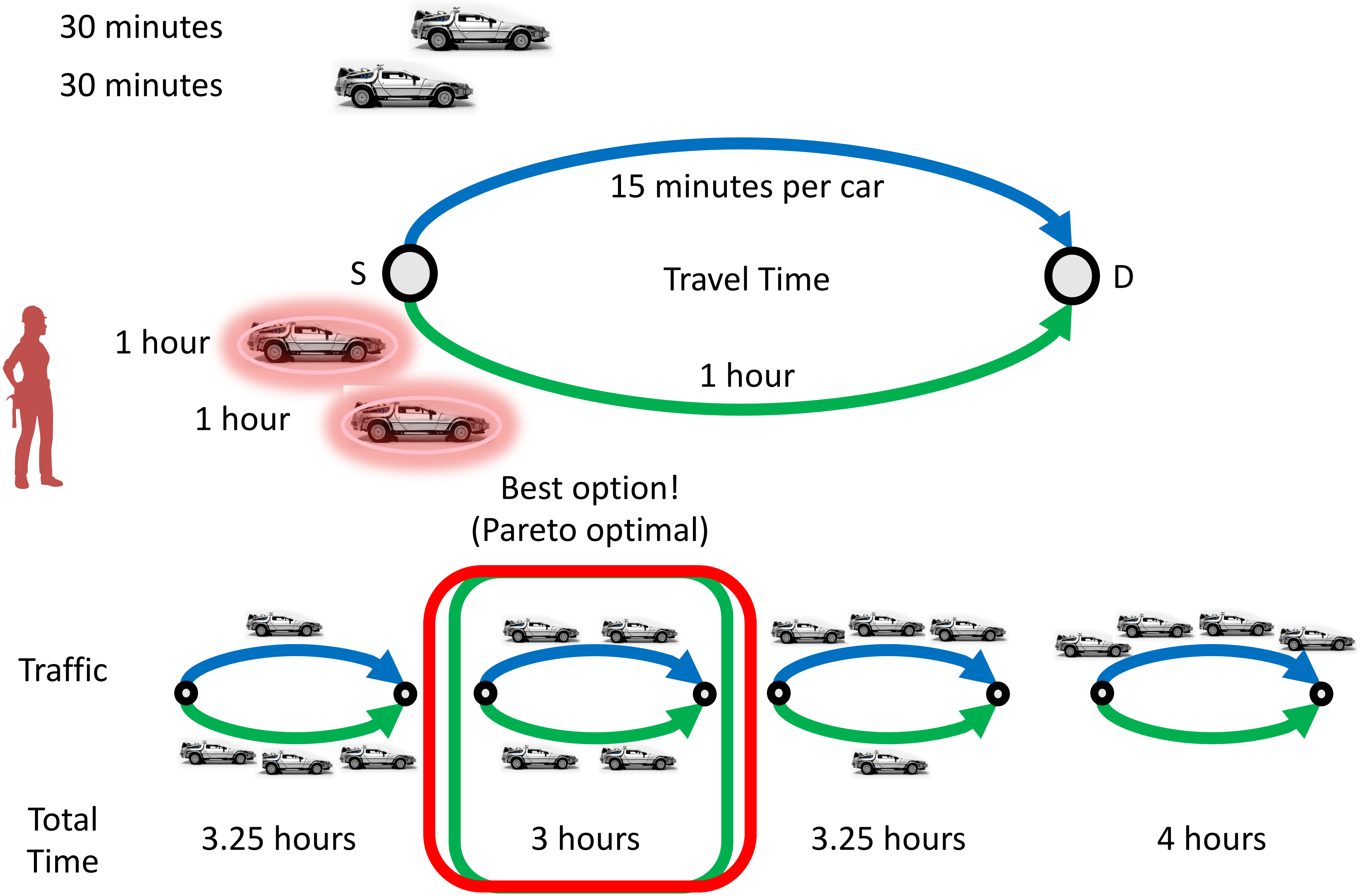
3.25 hours

3 hours

3.25 hours

4 hours

Self-driving car design



30 minutes

30 minutes

15 minutes per car

S

Travel Time

D

1 hour

1 hour

1 hour

Best option!
(Pareto optimal)

Traffic

Total Time

3.25 hours

3 hours

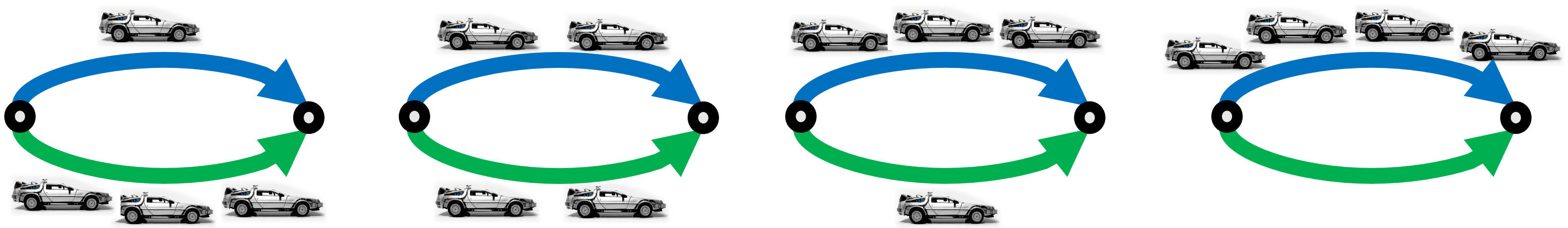
3.25 hours

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

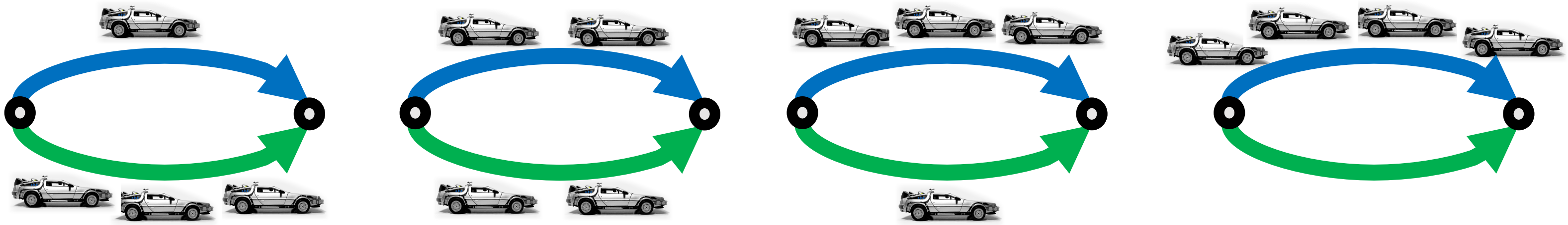
3.25 hours

4 hours

Altruistic self-driving cars:

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

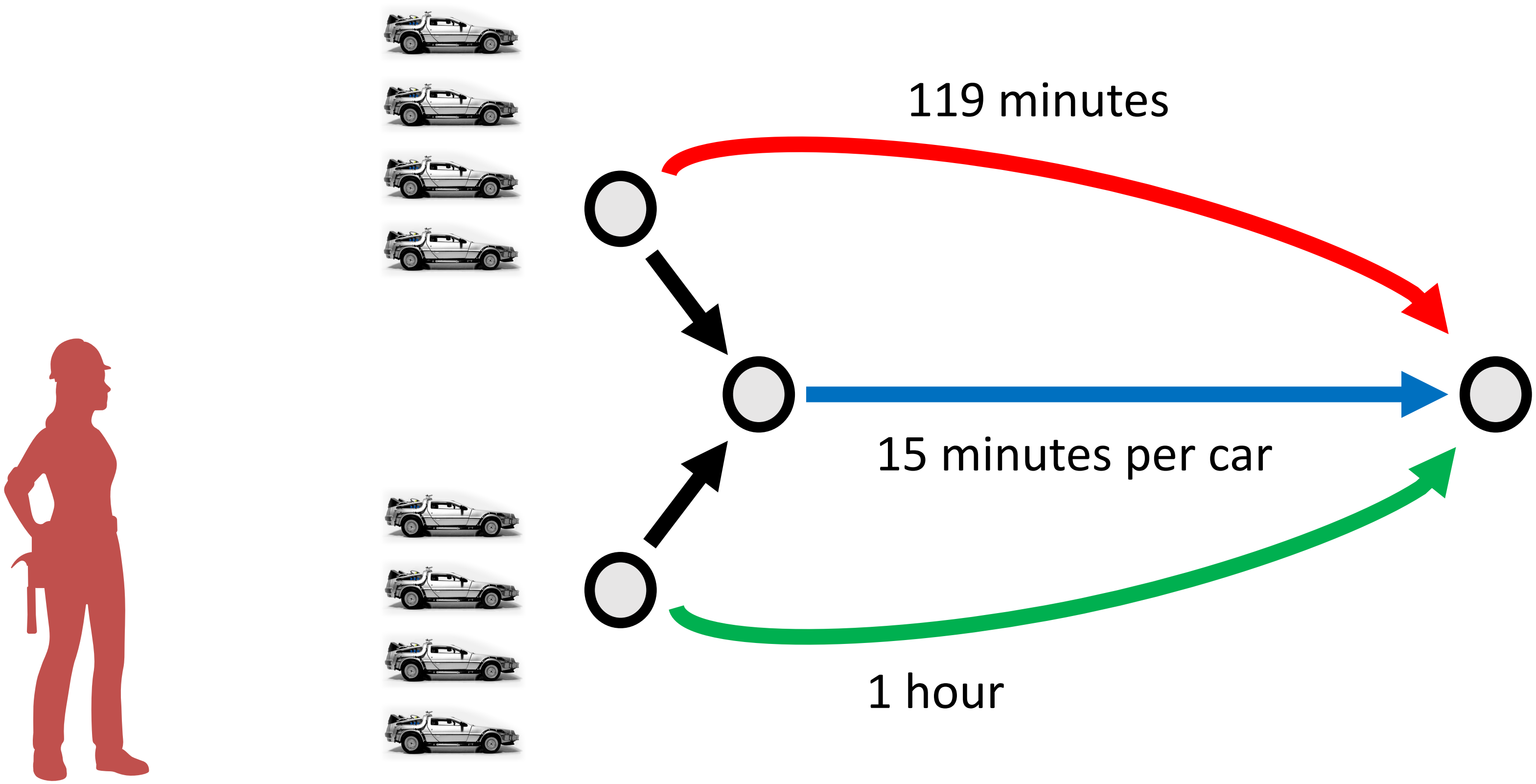
Traffic



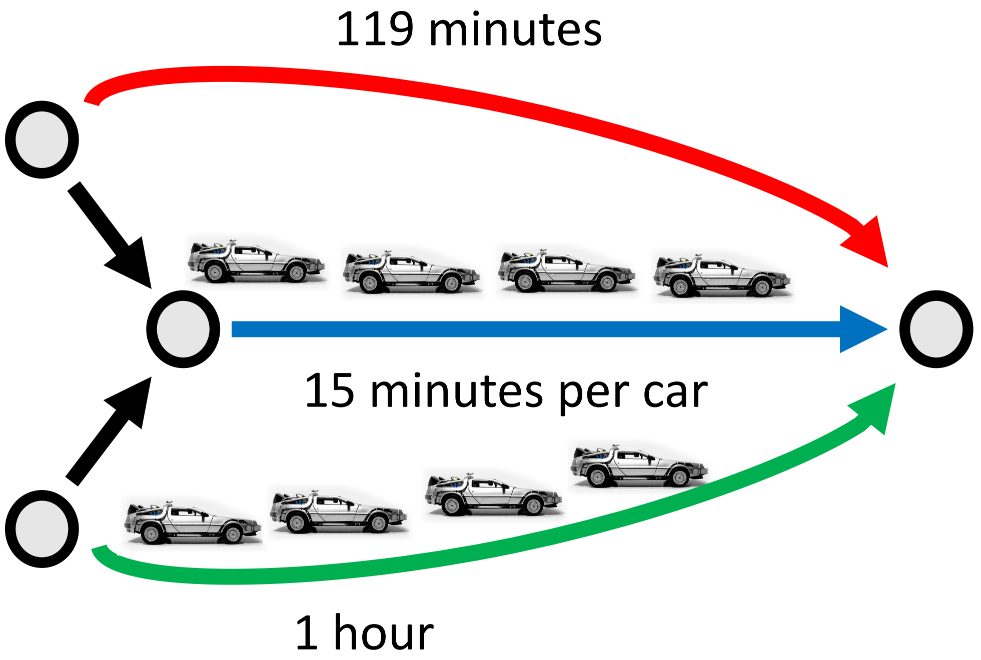
Total Time

3.25 hours	3 hours	3.25 hours	4 hours

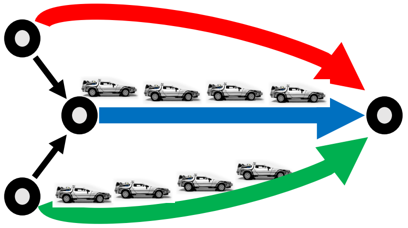
Self-driving car design



Self-driving car design



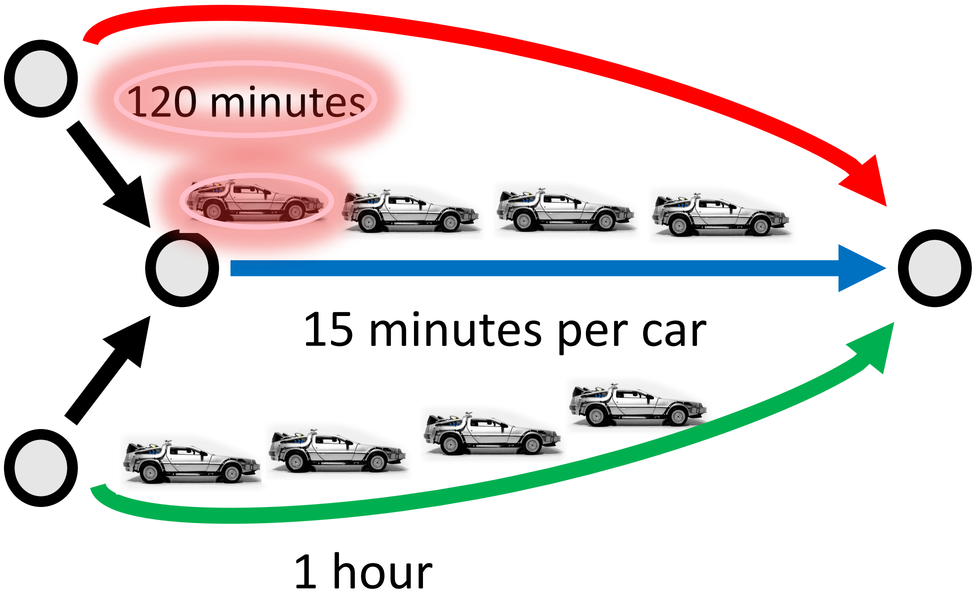
Selfish Traffic



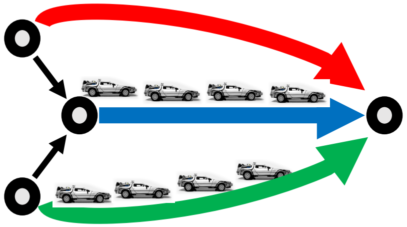
Total Time

8 hours

Altruistic self-driving cars?



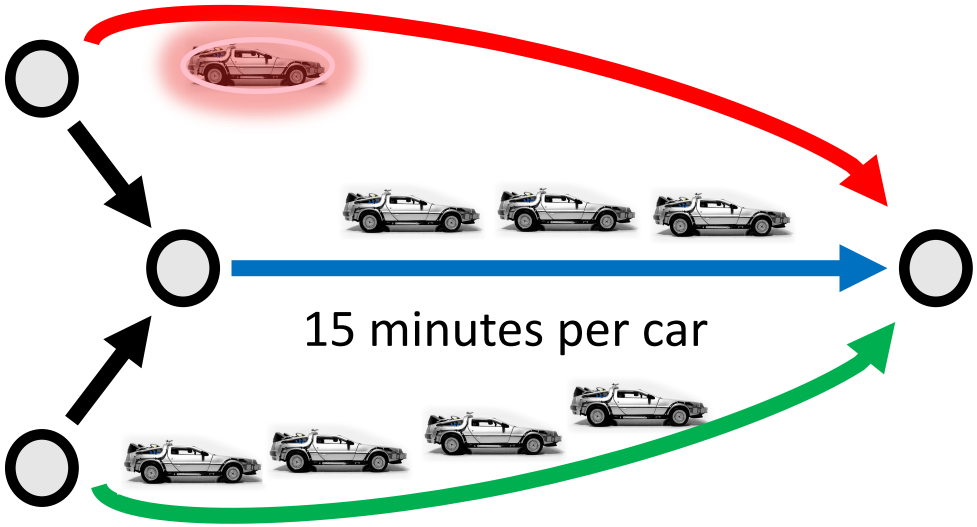
Selfish Traffic



Total Time

8 hours

Altruistic self-driving cars?

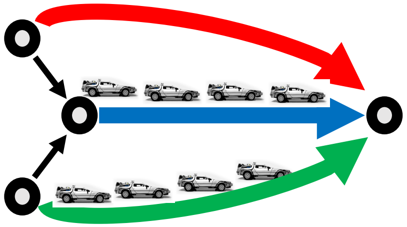


119 minutes

15 minutes per car

1 hour

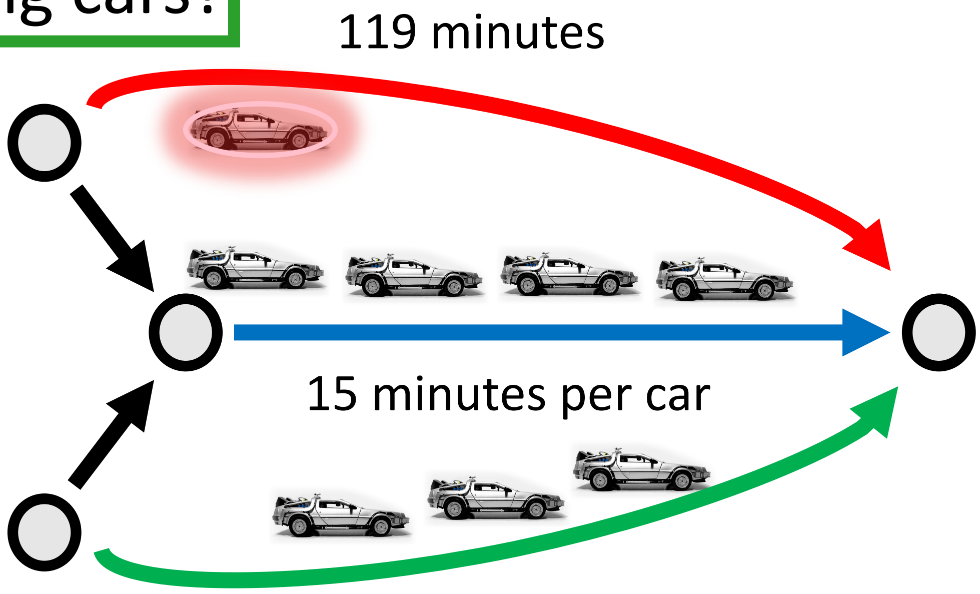
Selfish Traffic



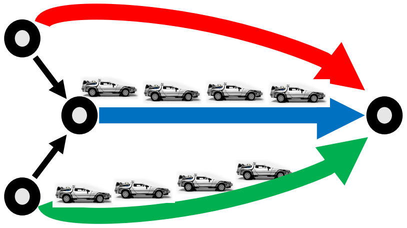
Total Time

8 hours

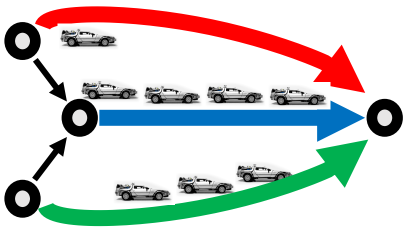
Altruistic self-driving cars?



Selfish Traffic



1 Altruistic Car

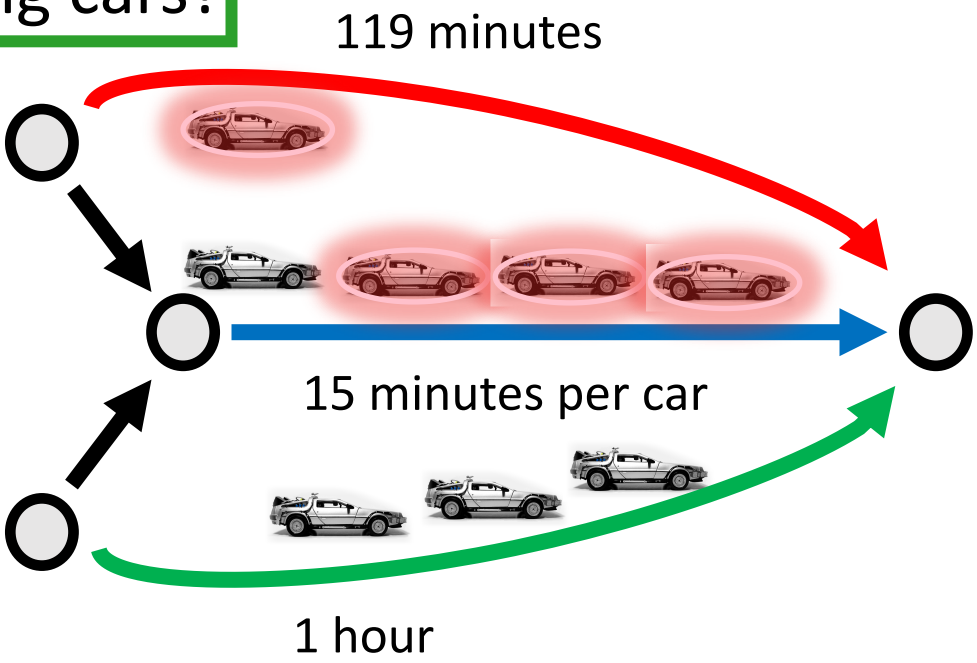


Total Time

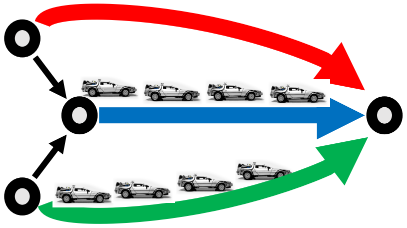
8 hours

9 hours

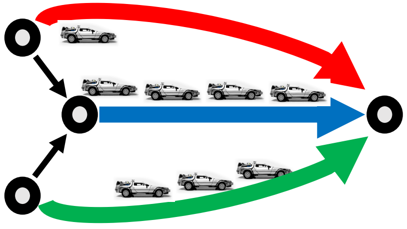
Altruistic self-driving cars?



Selfish Traffic



1 Altruistic Car

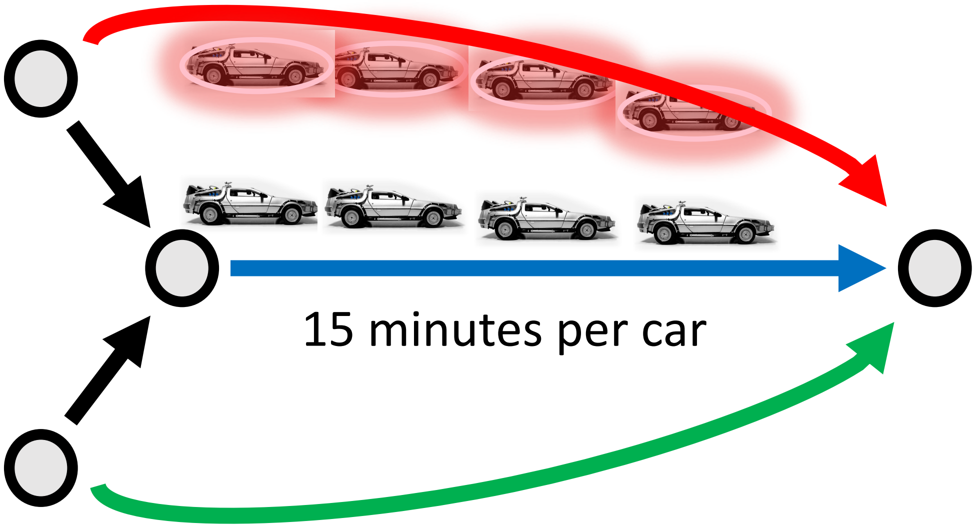


Total Time

8 hours

9 hours

Altruistic self-driving cars?

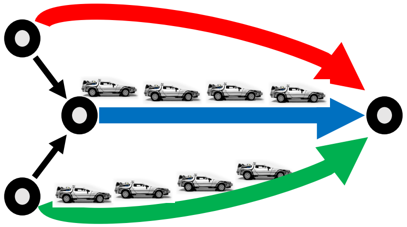


119 minutes

15 minutes per car

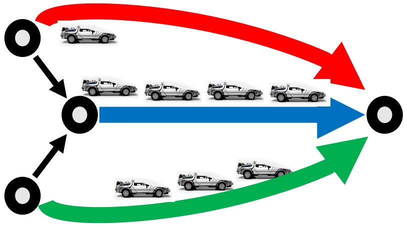
1 hour

Selfish Traffic



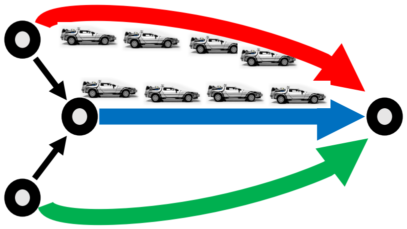
8 hours

1 Altruistic Car



9 hours

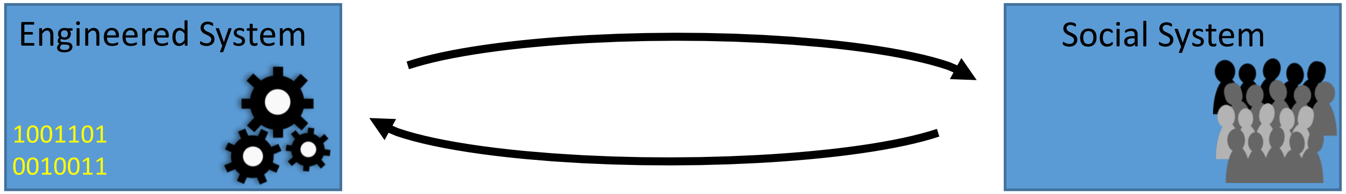
4 Altruistic Cars



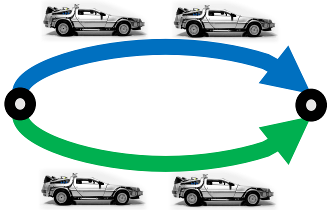
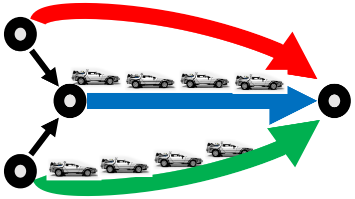
12 hours

Total Time

Conclusions



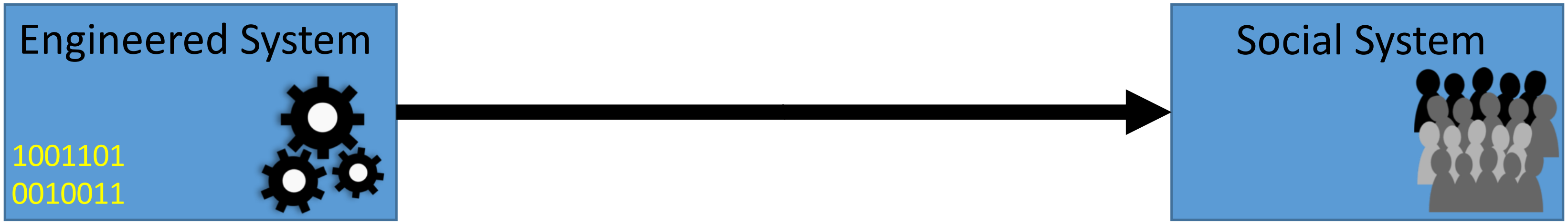
Decision design for socio-technical systems

Machine Policy			
Selfish	Not Optimal	Not Optimal	
Altruistic	Inconsiderate?	Also potentially inefficient!	

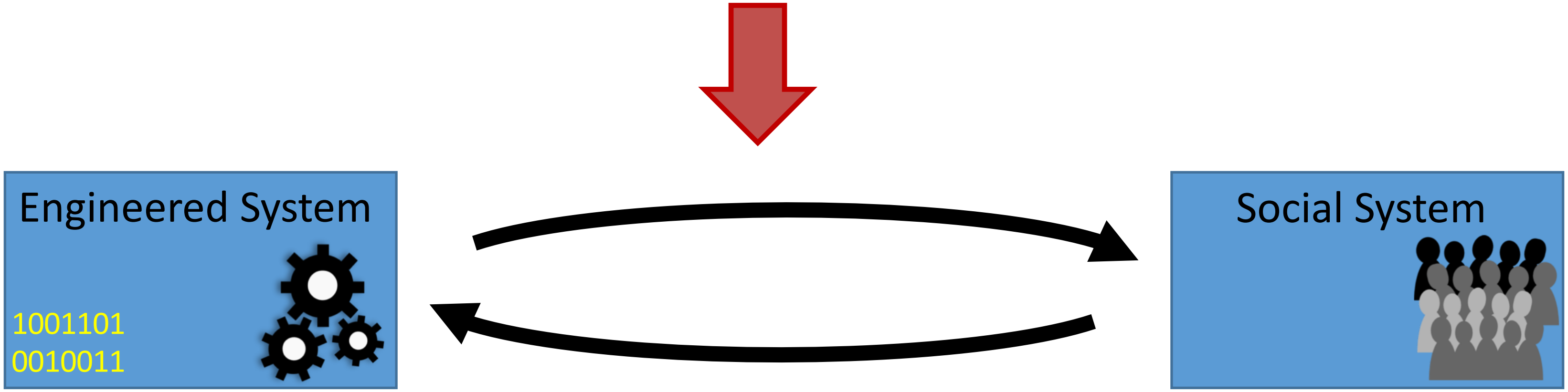
If **Ethical** means **utilitarian**, then **sometimes altruism is good!**

If **Ethical** means **risk-averse**, then **machines should be selfish!**

Broader Picture

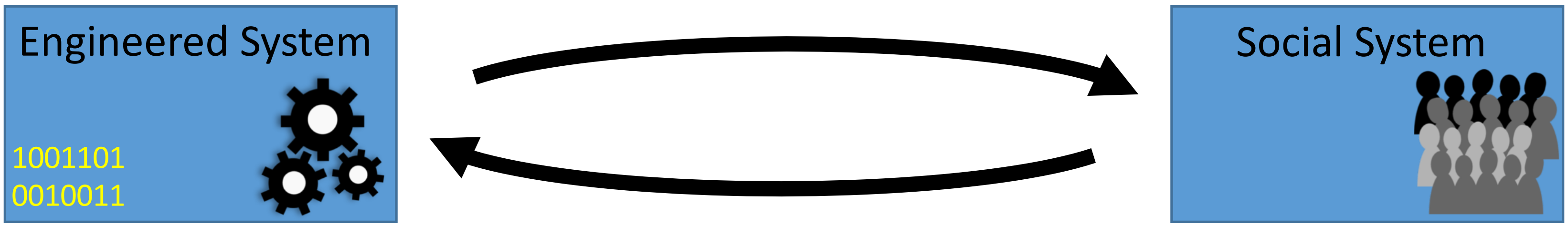


Traditional Ethics Message: Engineer things that don’t break



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

Broader Picture



Engineered System

1001101
0010011

Social System

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 → counterintuitive outcomes