# Responsibility and AI

5ARC0

2025-09-16 Lecture

# Four main notions of responsibility

Passive (backward-looking) moral responsibility
Active (forward-looking) moral responsibility
Collective moral responsibility
Legal responsibility

Royakkers & van de Poel, Ethics, Technology, and Engineering: An Introduction, p. 13

# Four main notions of responsibility

Passive (backward-looking) moral responsibility
Active (forward-looking) moral responsibility
Collective moral responsibility
Legal responsibility

# Passive/ backward-looking moral responsibility (blameworthiness)

**Causal Contribution**: A causal connection is present between a person's action and certain consequences;

**Freedom of Action:** The person acted freely/ not under compulsion;

**Foreseeability:** The person could have reasonably known that these consequences would arise; and,

**Wrong-doing:** Bringing about these consequences is wrong.

Image: By Nazbrok - Own work, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=4140179

# Traditional focus of responsibility analysis: division of responsibility between management and engineer

Image: Wikipedia

# Combining responsibility and risk perspectives

People, institutions, and companies are responsible for risk (passive sense) when…

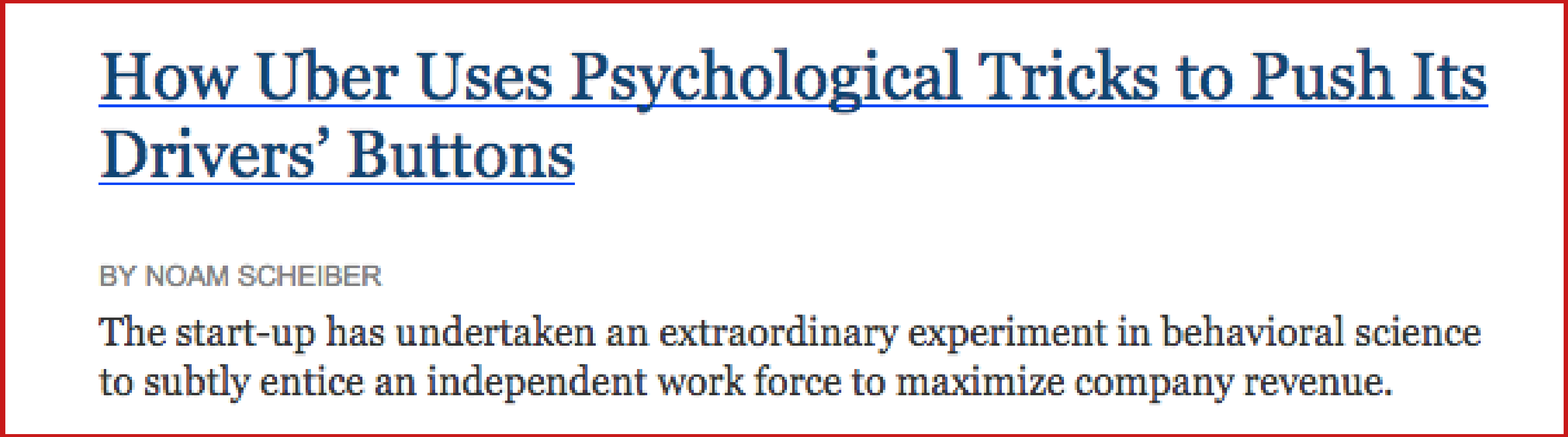
They causally contribute (increase the probability of the bad outcome);

They act freely in so doing;

They are or should be aware of the increased hazard; and,

The resulting risk is not acceptable.

# Problem case for discussion: technology-created incentives



How Uber Uses Psychological Tricks to Push Its Drivers' Buttons

BY NOAM SCHEIBER

The start-up has undertaken an extraordinary experiment in behavioral science to subtly entice an independent work force to maximize company revenue.

These incentives cause safety problems (because drivers travel through residential areas near airports and drivers do without sleep). Is the designer responsible for these problems?

# Definition of many hands problem

A situation in which a collective may reasonably be held responsible for an outcome, but no individual may reasonably be held responsible for that outcome.
The many hands problem results from applying the criteria for moral or legal responsibility to collectives and individuals.

https://commons.wikimedia.org/wiki/File:Pushing_van_together.jpg

# Responsibility gaps

The many hands problem is an example of a “responsibility gap”, where (a) it seems that we ought to hold somebody responsible for something, but (b) we cannot reasonably hold anybody responsible. Responsibility gaps can be accidental or intentional (purposeful avoidance of responsibility).  For example, a fashion brand might use many different suppliers with many different and changing subcontractors, in order to avoid being associated with exploitative labor conditions.

Syvden, Aug 16 2009, © Arnt Flatmo
The Blåfelden gap

# Responses to the many hands problem (and other responsibility gaps)

The many hands problem is sometimes used to motivate the importance of a clearer assignment of *active* ***responsibility***.

*Technology* can be used to help designate active responsibilities.
*The law* can be used to help designate active responsibilities.

# Active moral responsibility (role responsibility)

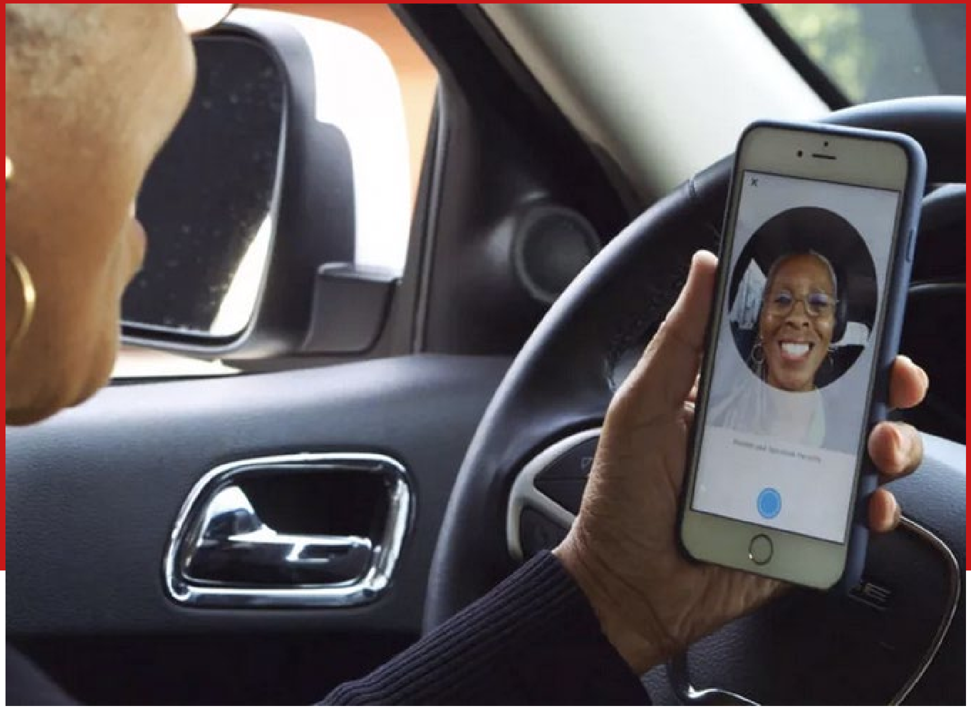
We can *designate* responsibility by giving a person a particular role that requires tasks. This is “active” in that it anticipates who is responsible *in the future*.

Image: Badger Herald

# Using technology to partially automate role responsibility

Andrew J. Hawkins, “Uber now requires drivers to take selfies for added security; critics say some drivers have never undergone background checks,” *The Verge*, 23 September 2016.

https://www.theverge.com/2016/9/23/13030682/uber-driver-selfie-facial-scan-fraud-security



clearly indicate whether your chosen actor should take responsibility, and under what conditions: when others (fail to) take action, and when different outcomes materialize

# Distributed role responsibilities regarding passenger safety

| Driver (Professional User) | Platform designer (Software engineer) | Management (Enterprise) | Government (Society) |
|---|---|---|---|
| Follow procedures<br>Report incidents | Provide tools for drivers<br>Anticipate technical and security gaps | Certify drivers and vehicles for competence and security<br>Maintain records of incidents | Require certification<br>Enforcement |

See https://eng.uber.com, Searching “safety”

# Values in the distribution of (active) responsibility

Feasibility (Is it realistically possible to live up to the responsibilities assigned?)
Transparency to outsiders (Can people outside the organization know when to hold somebody responsible?)
Fairness (Does the distribution of tasks, and praise/blame, give people what they deserve and is it distributed equitably?)

(Royakkers & van de Poel, Ethics, Technology, and Engineering: An Introduction, p. 13 p. 266)

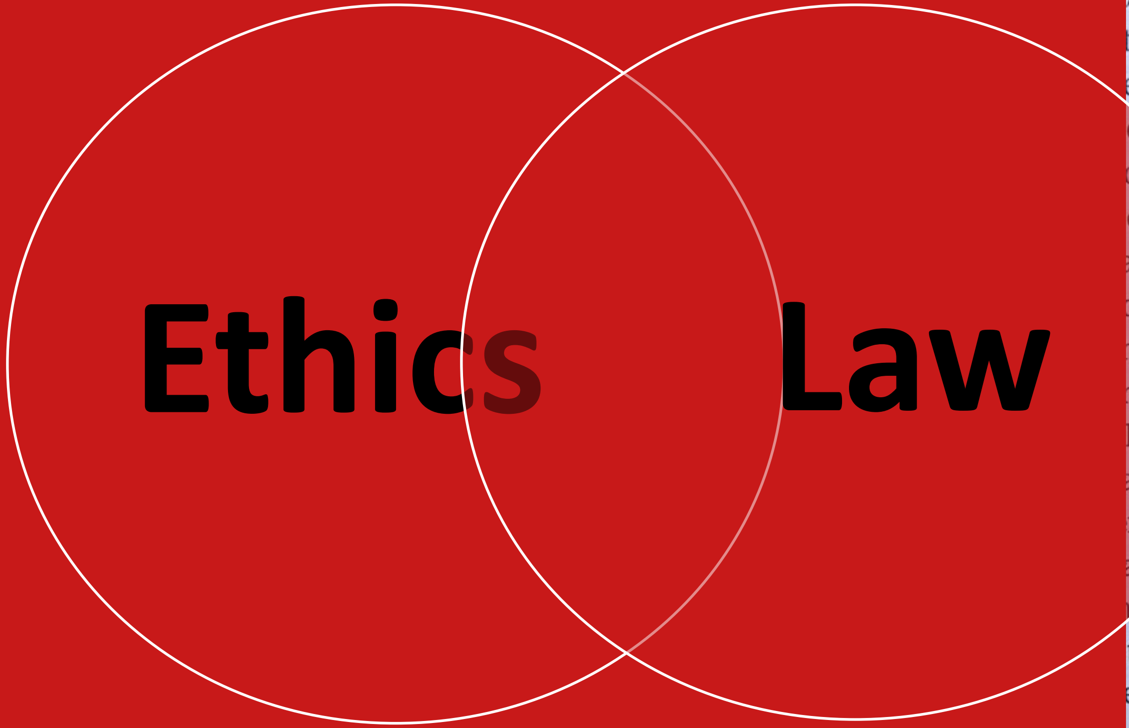
# Allocating active responsibility

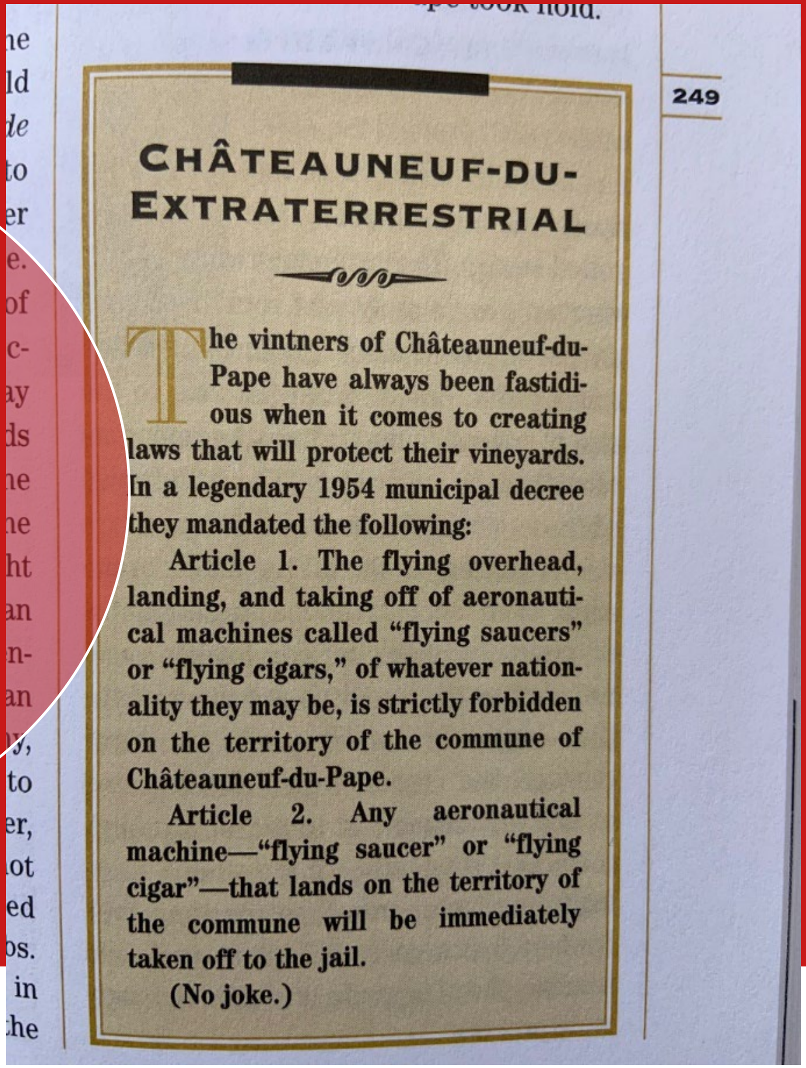
Active/ forward-looking responsibility is a matter of being assigned responsibility (or taking responsibility) for an area of activity, looking forward to the future.
This means that the responsible party puts herself in a position where she will:

- Have the ability to act
- Know what norms apply
- Adequately perceive violations of those norms
- Consider the consequences

(Royakkers & van de Poel, Ethics, Technology, and Engineering: An Introduction, p. 13)

# Moral responsibility vs legal





249

## CHÂTEAUNEUF-DU-EXTRATERRESTRIAL

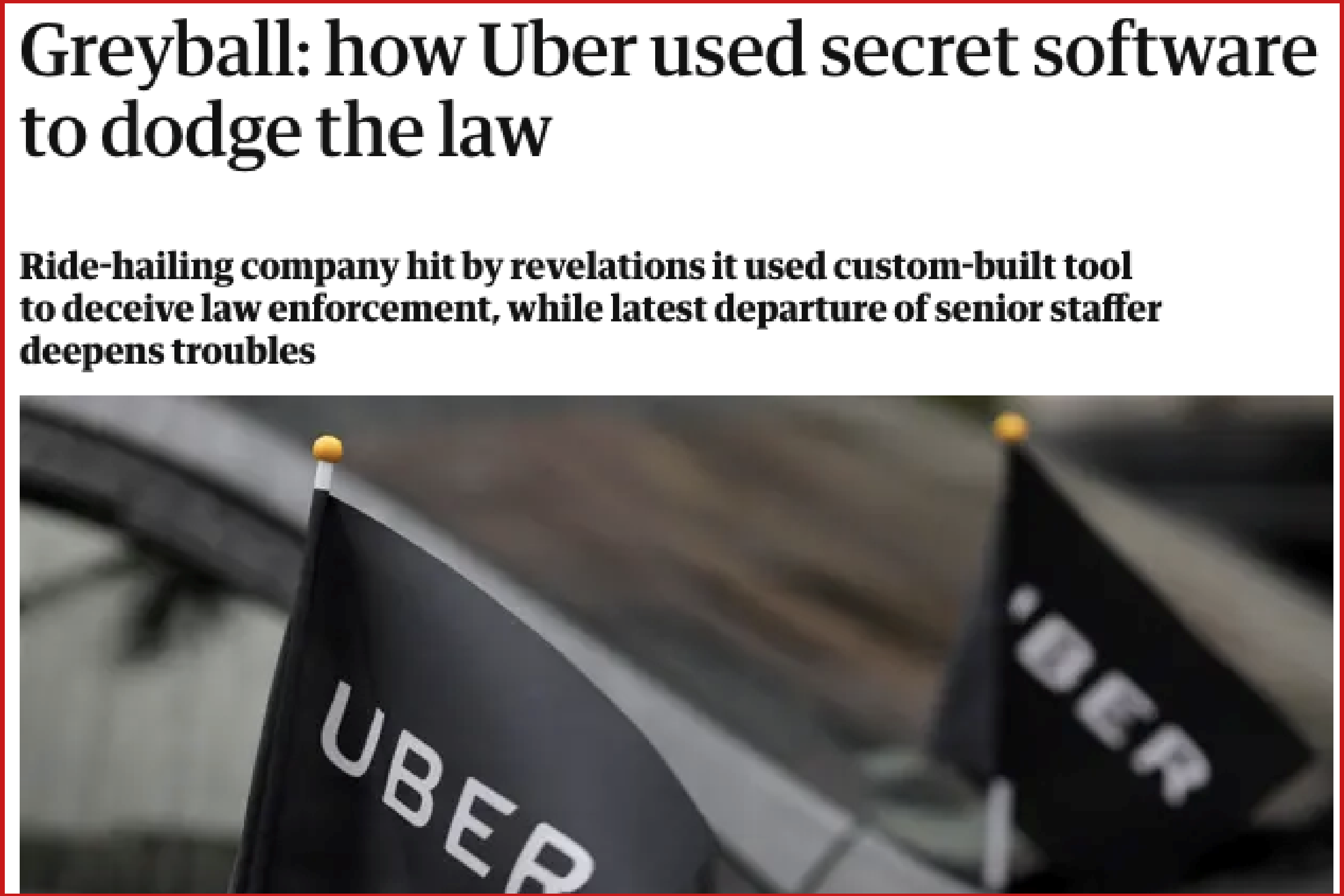
The vintners of Châteauneuf-du-Pape have always been fastidious when it comes to creating laws that will protect their vineyards. In a legendary 1954 municipal decree they mandated the following:

Article 1. The flying overhead, landing, and taking off of aeronautical machines called "flying saucers" or "flying cigars," of whatever nationality they may be, is strictly forbidden on the territory of the commune of Châteauneuf-du-Pape.

Article 2. Any aeronautical machine—"flying saucer" or "flying cigar"—that lands on the territory of the commune will be immediately taken off to the jail.

(No joke.)

# Problem case: Greyball



Greyball: how Uber used secret software to dodge the law

**Ride-hailing company hit by revelations it used custom-built tool to deceive law enforcement, while latest departure of senior staffer deepens troubles**

Julia Carrie Wong, "Greyball: How Uber used secret software to dodge the law," *The Guardian*, 4 March 2017

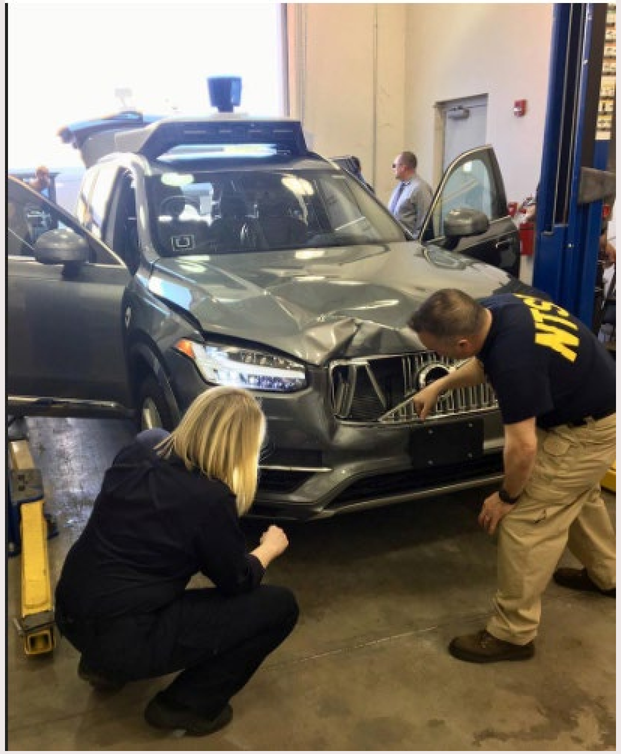
# Legal responsibility

Type 1. **Criminal responsibility**
Type 2. **Liability** is backward-looking legal responsibility in which the responsible party must pay a fine, arrange repair, or pay damages.
There are different standards possible for liability: negligence, and strict product liability

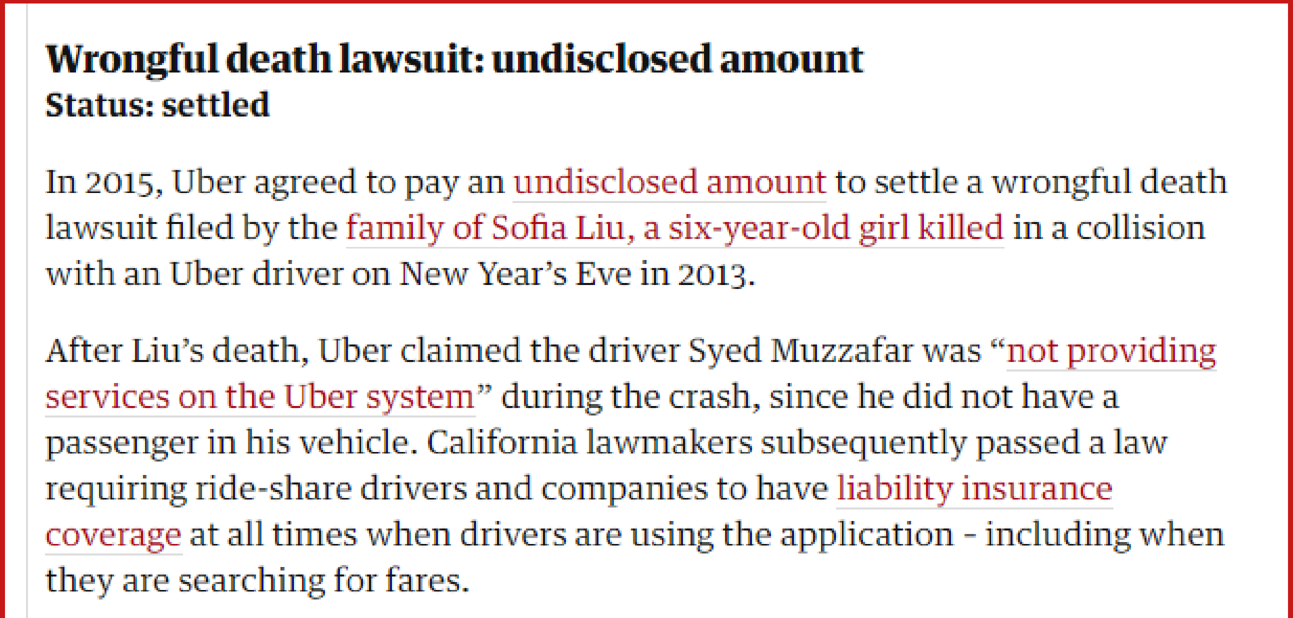
By Patrick Denker - originally posted to Flickr as IMG_5078, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=5580253

# Uber self-driving car crash (bicyclist killed)

Uber Technologies Inc was held not to be *criminally* responsible for the crash. The back-up driver, Rafaela Vasquez, could have faced charges of vehicular manslaughter. (Vasquez was watching streaming TV in the car at the time of the crash, and did not react in time to take over and avoid hitting the bicyclist.) Conceivably, Uber could have been sued for liability (wrongful death). But it was unlikely to be convicted of a crime in this case.



# “Wrongful death lawsuit”: liability



**Wrongful death lawsuit: undisclosed amount**
**Status: settled**

In 2015, Uber agreed to pay an undisclosed amount to settle a wrongful death lawsuit filed by the family of Sofia Liu, a six-year-old girl killed in a collision with an Uber driver on New Year’s Eve in 2013.

After Liu’s death, Uber claimed the driver Syed Muzzafar was “not providing services on the Uber system” during the crash, since he did not have a passenger in his vehicle. California lawmakers subsequently passed a law requiring ride-share drivers and companies to have liability insurance coverage at all times when drivers are using the application - including when they are searching for fares.

Source: Sam Levin, “Uber Lawsuits Timeline,” 13 April 2016. https://www.theguardian.com/technology/2016/apr/13/uber-lawsuits-619-million-ride-hailing-app

# Legal Responsibility

Liability:

- Backward-looking responsibility according to the law
- Relates to obligation to pay a fine or repair or repay damages.

Legal and moral backward-looking responsibility overlap but can come apart in two directions.

| Moral responsibility | Legal liability |
|---|---|
| Based on the four criteria | Based on legal conditions |
| Informal | Formal (court of law with proof standards) |
| Connected to blame, forgiveness, moral emotions | Connected to fines and damages |

# Negligence: the usual condition for legal liability

Proof consists of showing:

- A duty of care
- A breach of the duty
- Injury or damage
- A causal connection between the breach and the injury or damage

Foreseeability (a part of the breach of duty) is understood in terms of the reasonable person standard: a reasonable person in the position of the defendant could have foreseen the damage.

# Other standards of liability

Strict liability:

- The defendant engaged in a risky activity
- There was injury or damage
- This activity caused the injury or damage

Product liability:

- A kind of strict liability in which manufacturers are liable for defects in a product, without the need to show negligence.

# EU Standard for Product Liability

Development risks:
“the producer shall not be liable as a result of this Directive if he proves … that the state of scientific and technical knowledge at the time when he put the product into circulation was not such as to enable the existence of the defect to be discovered”
But:
“Each Member State may [make legislation such that] the producer shall be liable even if he proves that that the state of scientific and technical knowledge at the time when he put the product into circulation was not such as to enable the existence of the defect to be discovered”

# EU justification for product liability

“liability without fault on the part of the producer is the sole means of adequately solving the problem, peculiar to our age of increasing technicality, of a fair apportionment of the risks inherent in modern technological production”

Council Directive 85/374/EEC

# Other notions of liability

Corporate liability: Liability of a company

Limited liability: Shareholders may only be liable up to the value of their shares

# Relationship between notions of responsibility

**Backward looking/ passive** moral responsibility is basic, and has to do with linking an agent with a past event in a certain way.
**Forward looking/ active** moral responsibility is about *taking or being given* (backward-looking) moral responsibility in regard to future events.
**Legal** responsibility concerns punishment and compensation as part of a system meant to ensure justice. It has additional requirements such as provable harm, efficacy, and feasibility.
**Collective** responsibility refers to ways of assigning responsibility when there are multiple actors and automated decision making, responding to the so-called *many hands problem*.

# Cases for discussion

Robinhood/ GameStop short squeeze

Cruise/ 2023 accident in SF severely injuring woman

ChatGPT/ 2023 hallucination of court cases

# Complications: Responsibility for actions vs omissions

Actions: Intentional behaviors with a causal impact originating with (the motion of) an agent and attributable to them.

Omissions: Non-actions or failures to act (i.e., no motion), attributable to an agent.

People can be responsible for (the foreseeable effects of) both actions and omissions.

# Complications: Doubled or shared responsibility

Suppose that two people each apparently contribute to some event, and there is a question of responsibility.

For example, I give a speech that incites a crowd to cause property damage. The crowd is responsible, but I am also responsible.

The following sort of premise is not true and should be avoided:

***Either I am responsible, or the crowd is responsible.***

# Complications: Responsibility for risks and failure conditions

Risks: Scenarios in which a known, possible, bad outcome could materialize (with a certain probability).

Failure conditions: Scenarios in which there some element or person (whether inside or outside the system) does not behave according to the script (i.e., the desired operating conditions), with a potentially bad outcome as a result.

# Automation and responsibility

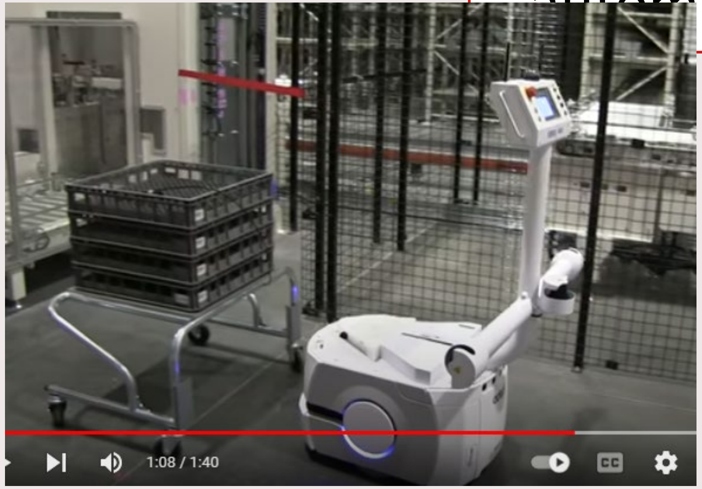
automation

full automation

human in the loop

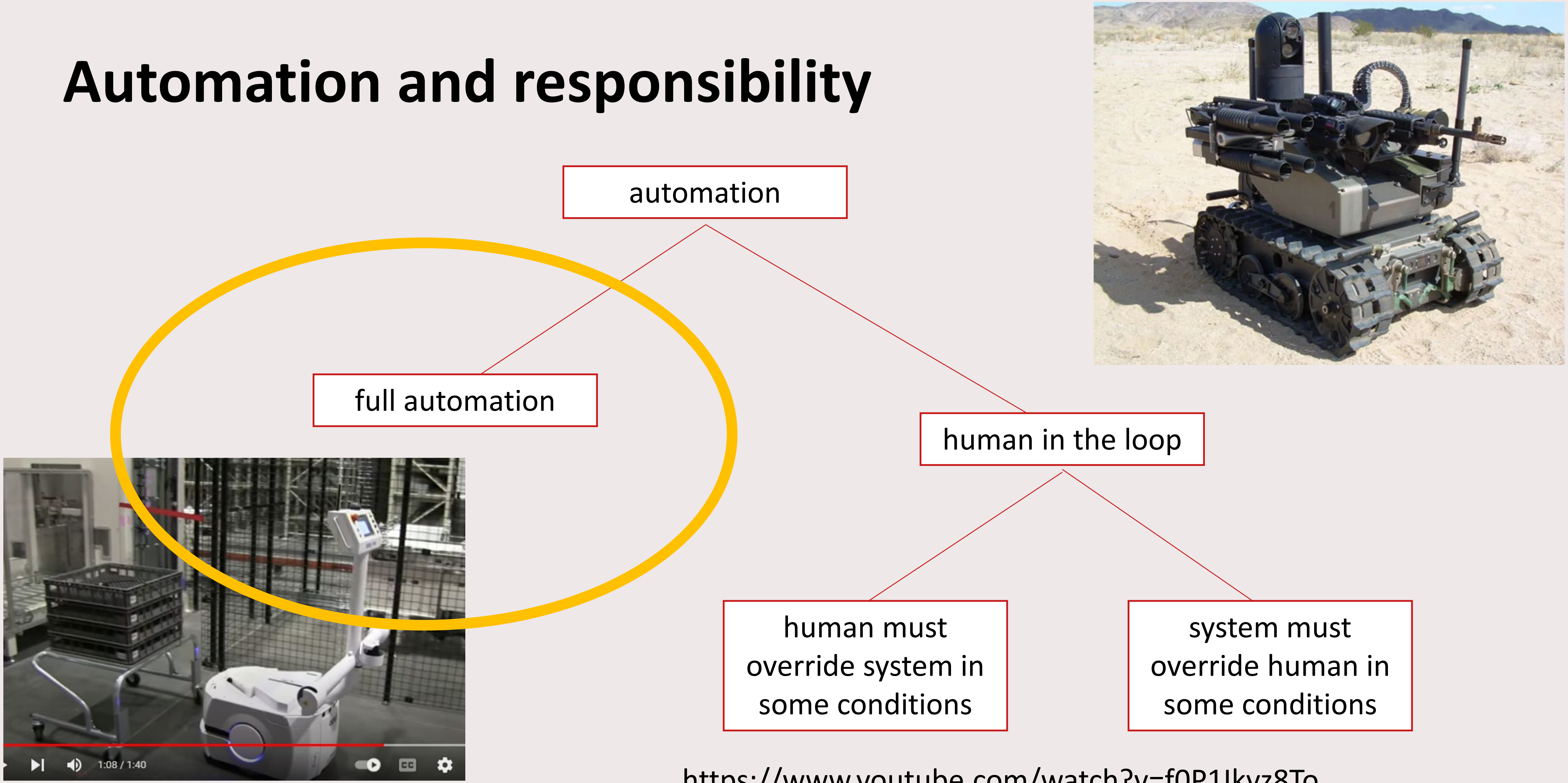
human must override system in some conditions

system must override human in some conditions



https://www.youtube.com/watch?v=f0P1Ikyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

# Automation and responsibility


automation
full automation
human in the loop
human must override system in some conditions
system must override human in some conditions
1:08 / 1:40

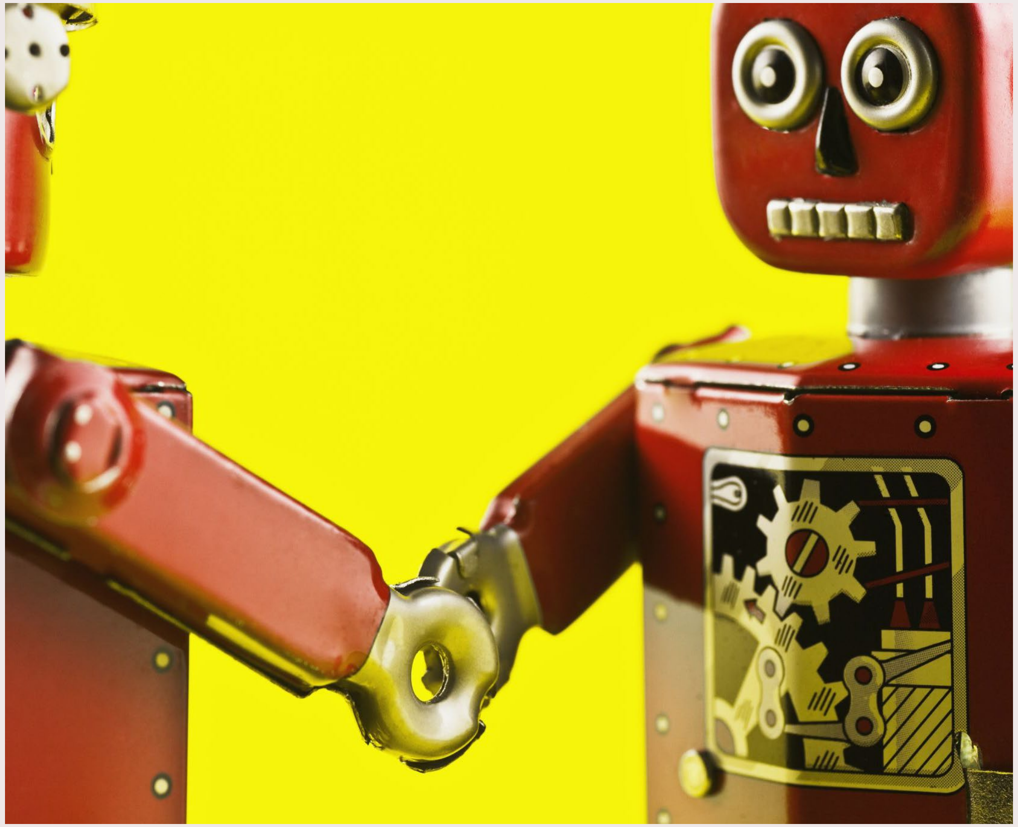
https://www.youtube.com/watch?v=f0P1Ikyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

# Can a machine be responsible?

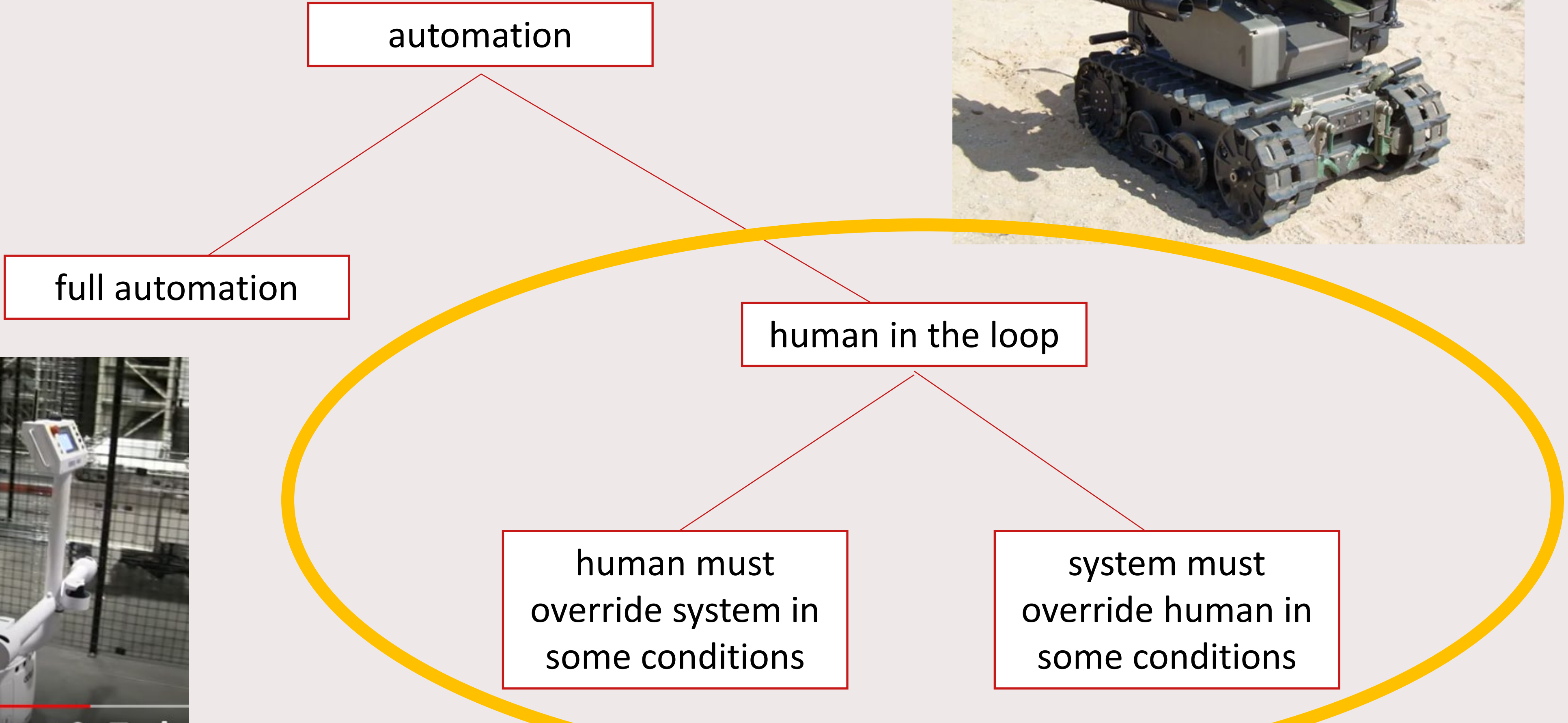
Apparently not: responsibility contains a freedom condition that machines do not satisfy.

In addition, there is a “retribution gap” (Danaher 2016): it does not make sense to blame or punish a machine.

One reason this might matter is that a machine cannot feel guilt, suffer, or correct its own behavior in response to blame.



# Automation and responsibility



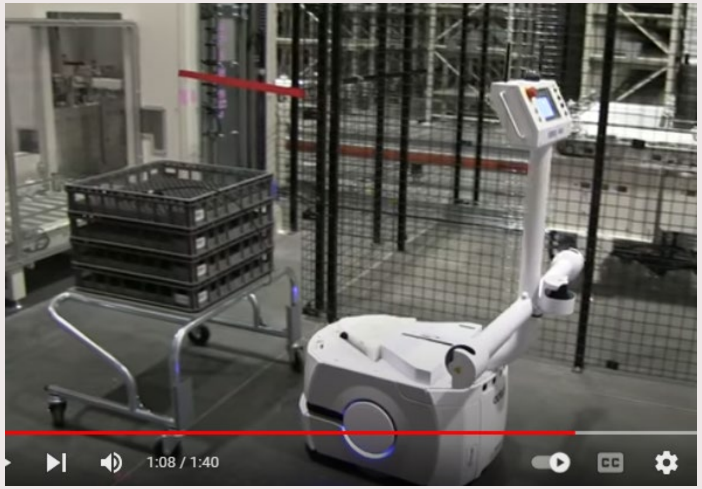
automation

full automation

human in the loop

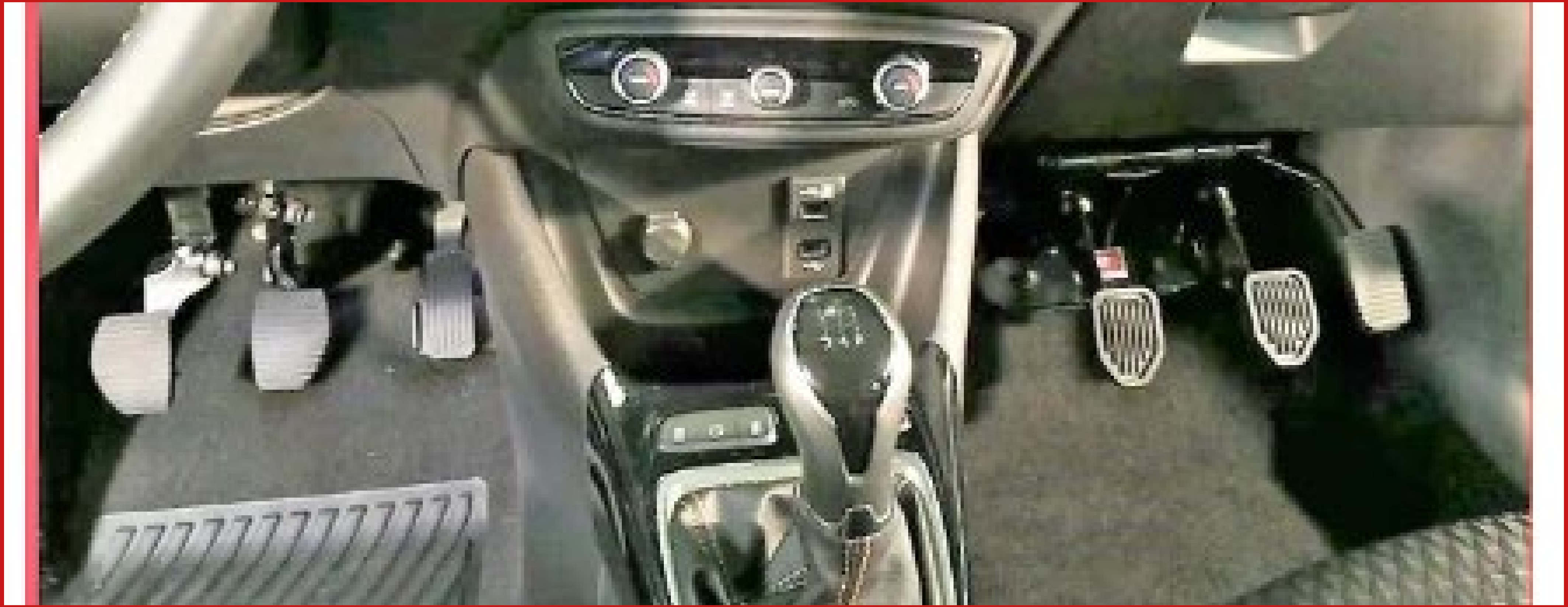
human must override system in some conditions

system must override human in some conditions



https://www.youtube.com/watch?v=f0P1Jkyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

# Responsibility of primary and secondary operators



https://www.drivingschoolsupplies.ie/store/p1/dual-controls-for-driving-schools.html

# Responsibility and design

‘‘many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors.”

Akrich 1992, P. 216

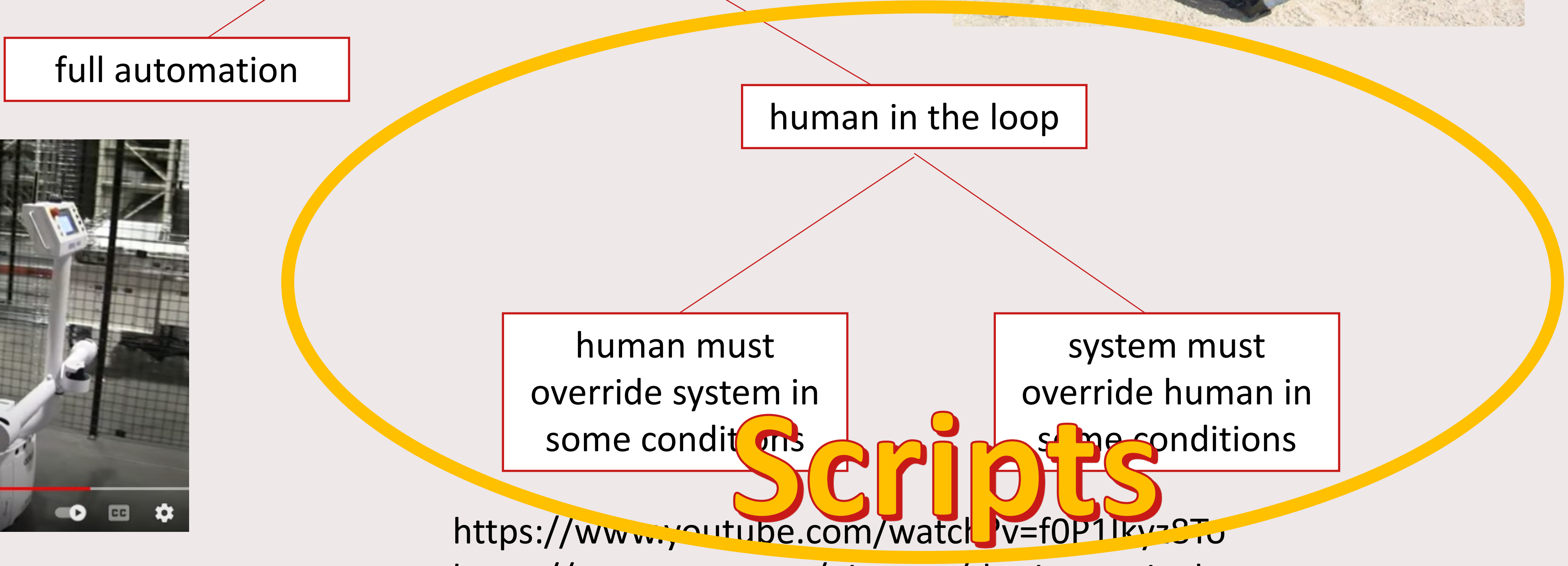
# Responsibility and design

“many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors.”

Akrich 1992, P. 216

Image: https://industrialscripts.com/stage-directions/

# Automation and responsibility



automation

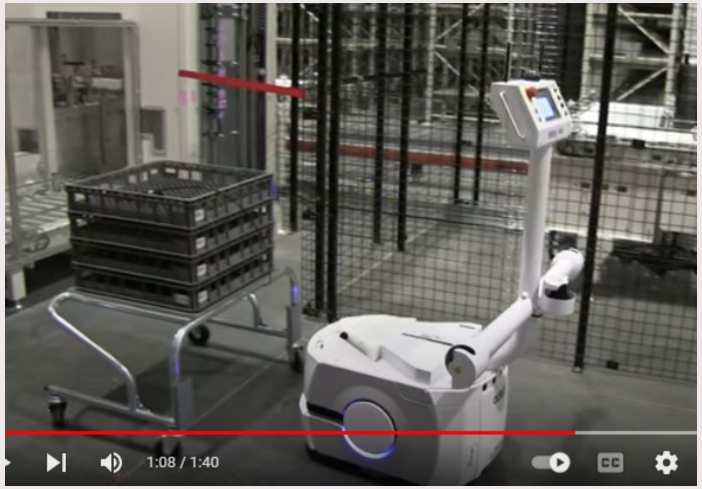
full automation

human in the loop

human must override system in some conditions

system must override human in some conditions

Scripts



https://www.youtube.com/watch?v=f0P1Jkyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

## Solutions to the problem of responsibility and failure conditions: assigning forward-looking responsibility more strictly

Can we write scripts more clearly and broadly to anticipate more failure conditions?

Can we insulate the automated system from potential failure conditions?

Can we enforce the scripts?