Decision Making
in Robots and Autonomous Agents

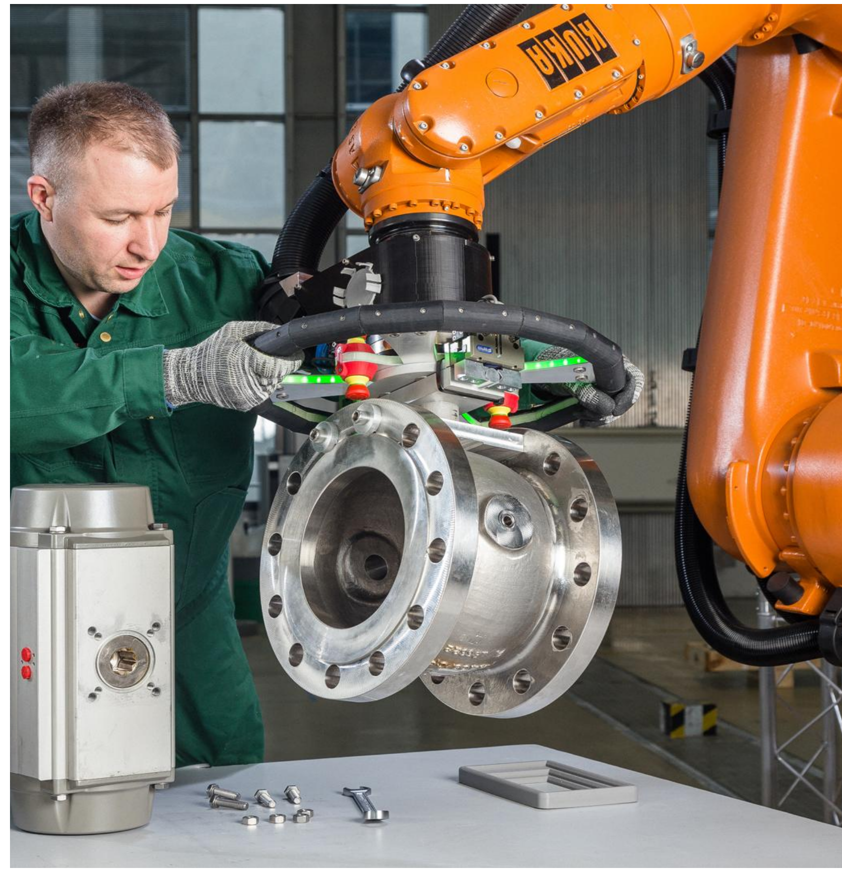
Safety:
Some Considerations and Methods

Subramanian Ramamoorthy
School of Informatics

9 March, 2018

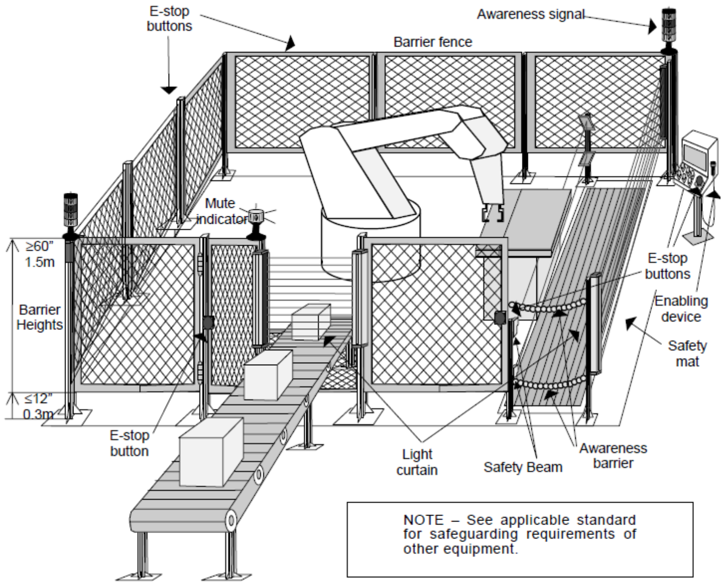
Why? Example 1

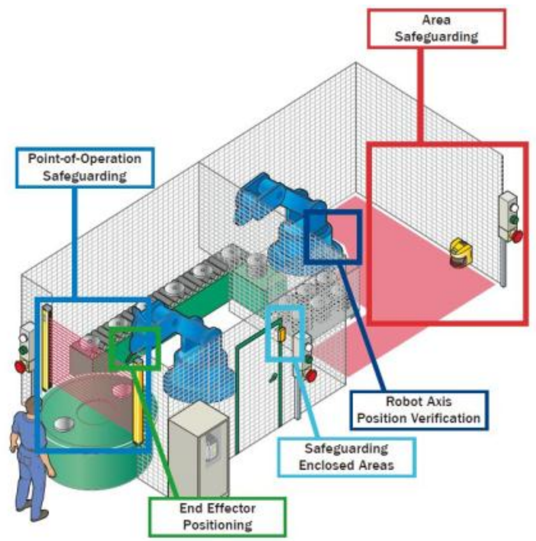
[Source: aslib.co.uk]



[Source: www.iff.fraunhofer.de]

Traditional Notions of Robot Safety





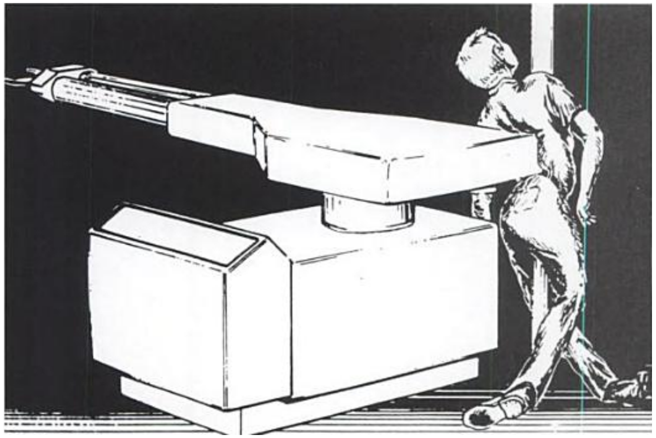
Deployed robot safety systems

Robots, depending on the task, may generate paint mist, welding fumes, plastic fumes, etc. In general, the robot, on occasion is used in environments or tasks too dangerous for workers, and as such creates hazards not specific to the robot but specific to the task.

[Source: G. Cui et al., Ontario]

Examples of Accidents

Example 1: First fatal robot-related accident in the U.S.



On July 21, 1984, a die cast operator was working with an automated die cast system utilizing a Unimate Robot, which was programmed to extract the casting from the die-cast machine, dip it into a quench tank and insert it into an automatic trim press.

A neighboring employee discovered the victim pinned between the right rear of the robot and a safety pole in a slumped but upright position. The victim died five days later in the hospital.

[Source: G. Cui et al., Ontario]

Examples of Accidents

Example 2:

A material handling robot was operating in its automatic mode and a worker violated safety devices to enter the robot work cell. The worker became trapped between the robot and a post anchored to the floor, was injured and died a few days later.

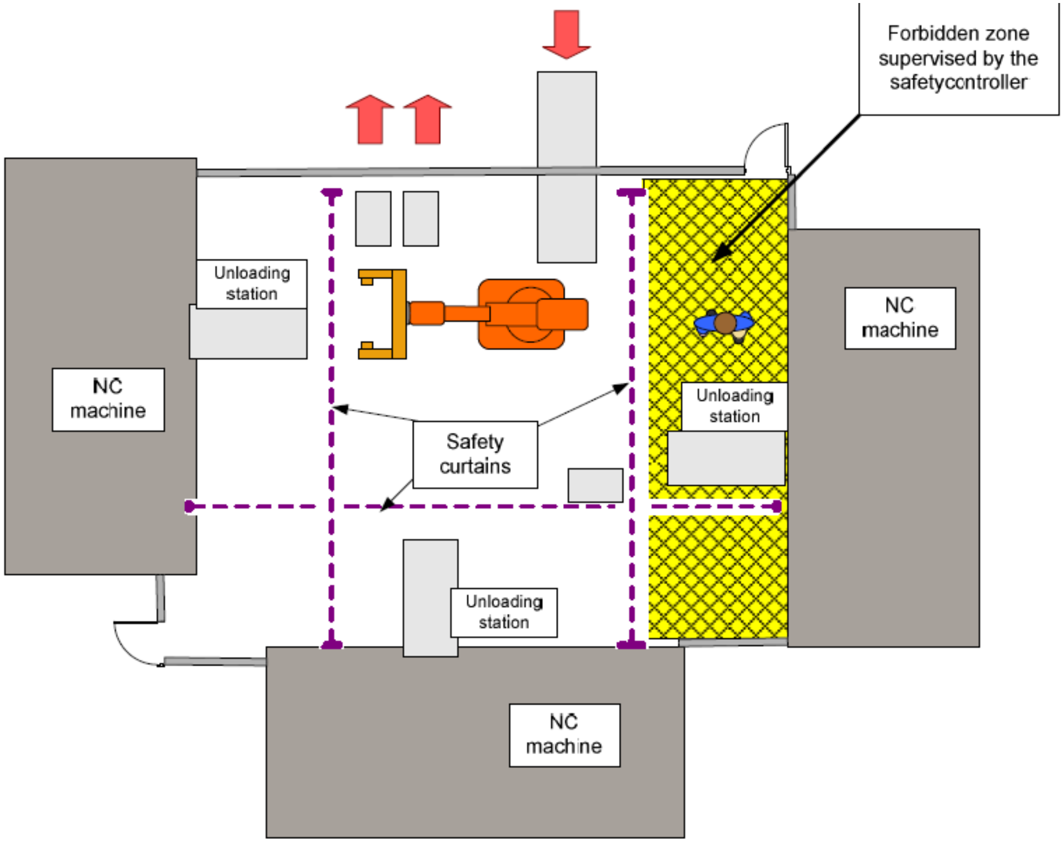
Example 3:

A maintenance person climbed over a safety fence without turning off power to a robot and performed tasks in the robot work zone while it was temporarily stopped. When the robot recommenced operation, it pushed the person into a grinding machine, killing the person.

[Source: G. Cui et al., Ontario]

How is ‘Safety’ Implemented?

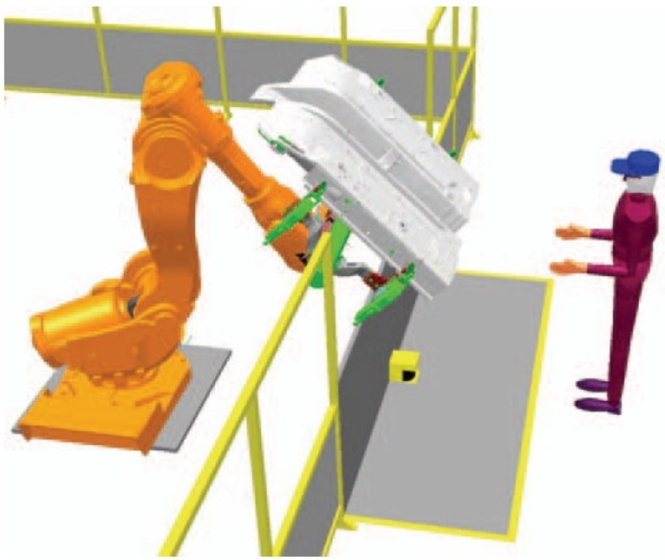
Example 1: Monitor and increase safety of tool zones



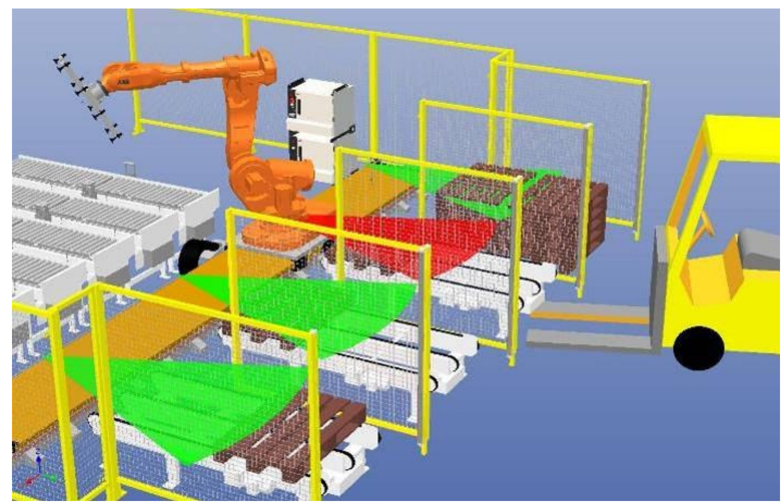
[Source: G. Cui et al., Ontario]

How is 'Safety' Implemented?

Example 2: Safe stand still/ direct loading of a robot



Example 3: Safe axis ranges with track motions



[Source: G. Cui et al., Ontario]

Characterizing an *Unsafe* Robot

Human Interaction

Control Errors

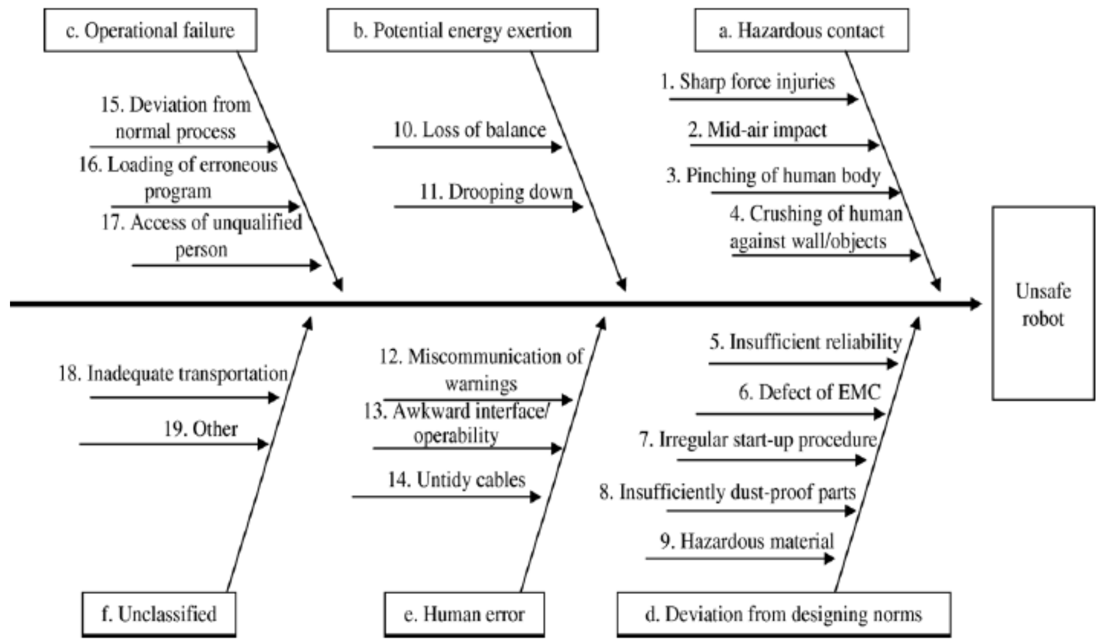
Unauthorized Access

Mechanical Failures

Environmental Sources

Power Systems

Improper Installation



Why? Example 2

HOME SEARCH The New York Times

TECHNOLOGY 855 CO

Google's Driverless Cars Run Into Problem: Cars With Drivers

By MATT RICHTEL and CONOR DOUGHERTY SEPT. 1, 2015

How will you characterise 'unsafe' in this context? Discuss!

Are these Issues Unique to Robotics?

NO!

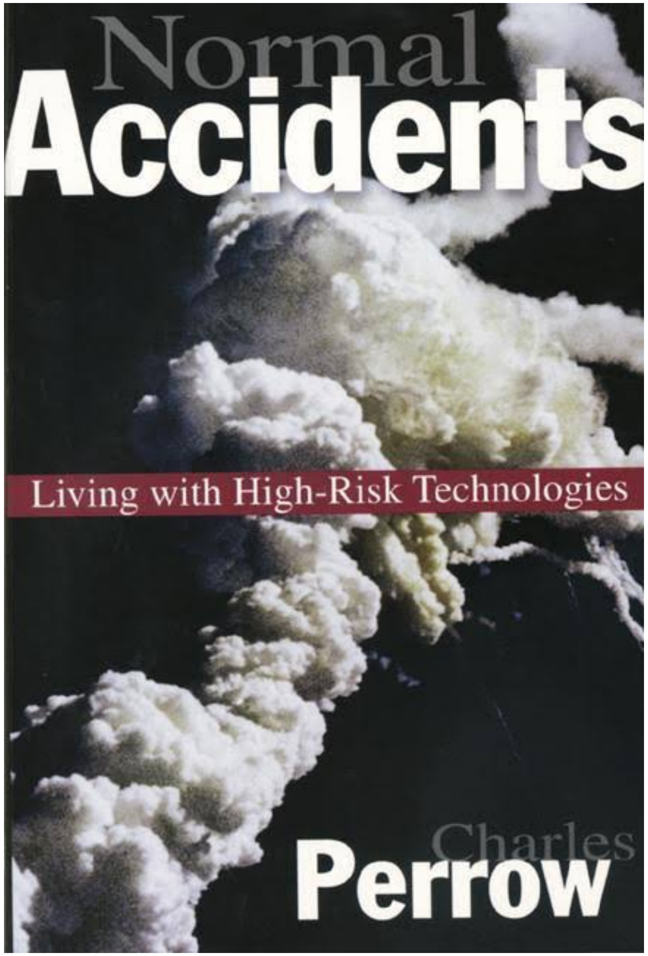
- Many other engineering systems have been through a similar path towards understanding safety
- Avionics, maritime systems, nuclear reactors, …
- ... office printers!

- Many famous examples of failures which are systemic rather than individual component driven

Perrow's Notion of *Normal Accidents*

- While many initial accident analyses have blamed the human operators, the real fault lies in system design
- Certain high-risk systems, because of the way they configure sequences of subsystems, are *naturally* prone to eventually resulting in an accident.
- So, Three Mile Island was a Normal Accident

(cars and humanoids too?!)



Normal Accidents: Core Argument

- Interactive Complexity
  - Failures of two components interact in an unexpected way
- Tightly Coupled
  - Processes that are parts of a system that happen quickly and cannot be turned off or isolated

- Perrow's Thesis: Tightly coupled systems with high interactive complexity will have Normal Accidents

Example: Three Mile Island

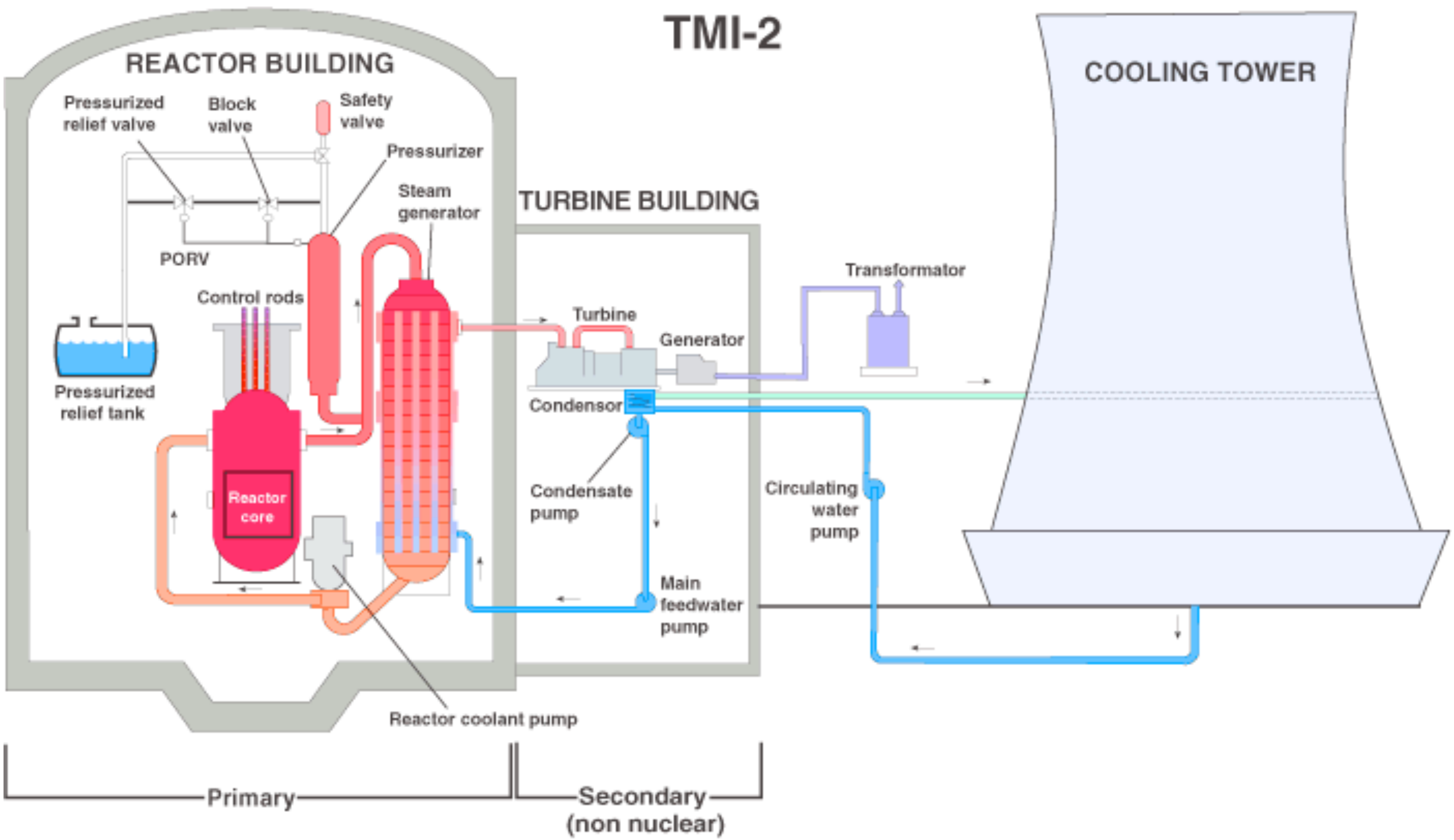
- Perhaps the most famous nuclear accident in the US
- On March 16, 1979, the movie China Syndrome (addresses social issues around nuclear accidents)
- 12 days later, March 28, 1979, the worst civilian nuclear accident in the US occurred at the Three Mile Island Nuclear Power Plant on the Susquehanna River, south of Harrisburg, PA.

[Source: Michael Carini, astro.wku.edu]

Three Mile Island: Location

[Source: Michael Carini, astro.wku.edu]

Example: Three Mile Island



Cooling System Setup

- Primary Cooling System
  – High pressure, radioactive, water circulating through the reactor.
  – Heat Exchanger transfers heat to the secondary system

- Secondary Cooling System
  – Cools the primary cooling system
  – Creates steam to run the turbines to generate electricity
  – Due to thin tubes in the turbine it must be very pure Continuously cleaned by a "polisher system"

A Sequence of Events

- The polisher leaked about a cup a day of water through a seal
- Water vapor got into a pneumatic system that drives some instruments
- This water vapor interrupted pressure to two valves in the feedwater system, which caused two feedwater pumps to shut down
- Lack of flow in the secondary system triggered a safety system that shut down the turbines
- This was the first indication of trouble to the operators
- At this point the reactor still needs to be cooled – or else

Sequence of Events: Emergency System

- An emergency feedwater system starts up to pump stored cold water through the secondary system to remove the accumulating heat
- The pumps were running, but valves on the pipes were incorrectly left closed from prior maintenance
- The operators insist they were left open; checklist says so
- A Repair Tag on a broken indicator hung over the indicator on the control panel that indicated that the valves were closed
- Redundant pipes, redundant pumps, and redundant valves, all thwarted by having the two valves physically at the same place and mis-set
- Eight minutes later they noticed they were shut by then the damage was done

No Cooling = Reactor Heats Up

- Due to overheating the reactor "scrammed" automatically
- This shuts down the reaction
- Enough heat remains in the reactor to require a normal working cooling several days to cool off
- Without cooling the pressure goes up
- An ASU Automatic Safety Device takes over to temporarily relieve the pressure: the Pilot Operated Relief Valve (PORV)

PORV (Pilot Operated Relief Valve)

- The PORV is supposed to vent pressure briefly, and then reclose
- If it stays open too long liquid escapes, pressure in the reactor drops, steam forms causing voids in the water, cooling is impaired and some places get yet hotter
- Thirty-two thousand gallons of water eventually went out this unclosed valve
- There was an indication on the control panel that the message to reseat had been sent to the valve
- However, no indication was available that it had reseated
- We are now thirteen seconds into the "transient"
- An indicator shows that there is extra water from an unknown source

Automatic Cooling Pump Starts

- This is another automatic safety system that pumps water to cool the reactor automatically starts at 13 seconds. The second was manually started by the operator
- For three minutes it looked like the core was being cooled successfully
- However, apparently due to the steam voids, the cooling was not happening
- The secondary steam generators were not getting water and boiled dry - at the same time water was flowing out of the primary cooling system through the stuck pressure relief valve

High Pressure Injection Starts

- This is an automatic emergency device that forces cold water into the reactor to cool it down.
- The reactor was flooded for two minutes, and then the operators drastically cut back the flow. This was regarded as the key operator error; what they did not realize was that the water was flowing out the PORV and the core would become uncovered
- Two dials confused the operators:
  - one said the pressure in the reactor was rising
  - the other said it was falling
- The Kemeny commission thought the operators should have realized this meant LOCA (Loss of Coolant Accident)

What is it Like in Control Room?

- Three audible alarms are making a din
- Many of the 1,600 indicator lights are blinking

- The computer is way behind in printing out error messages
- It turns out they can only be printed, not spooled to disk, to see the current condition they would have to purge the printer and loose potentially valuable information

- The reactor coolant pumps begin the bang and shake, due to cavitation from lack of water to pump-they are shut off

Stuck Open PORV Valve Discovered

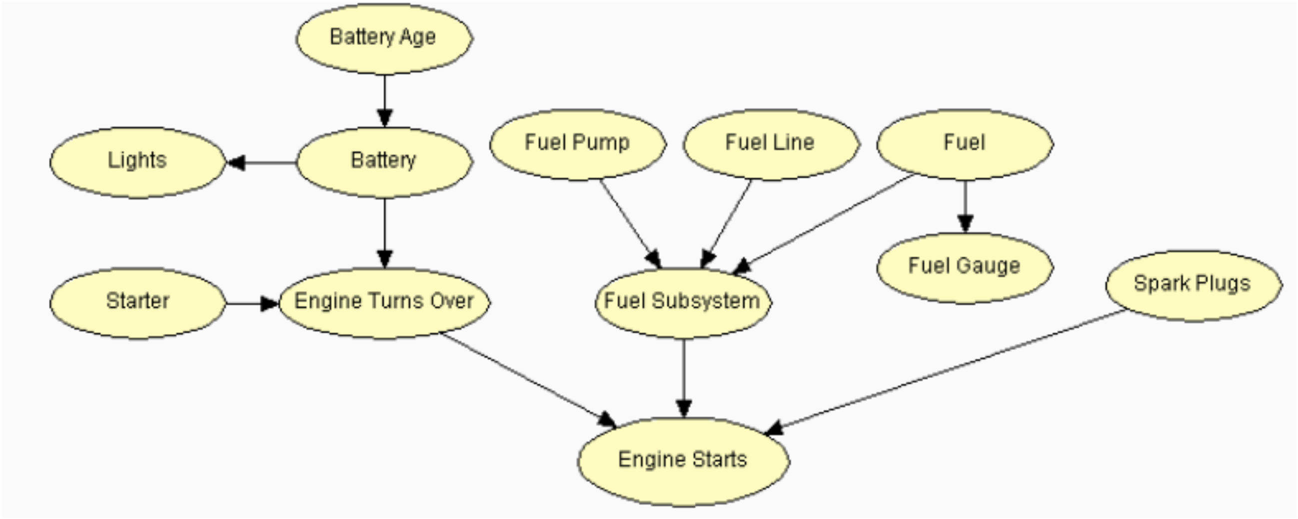
- The operators checked the valve and found it open
- They closed it
- With some trepidation since they were messing with a safety system
- The reactor core had been uncovered at this point and had partially melted
- Another 30 minutes without coolant and it would probably have been a total melt down (e.g., Chernobyl)

Tell tale Signs

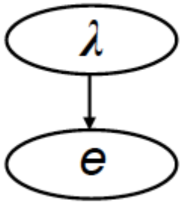
- The whole system is never all up and working as designed thus it is hard to understand
- When things start to fail the system is even harder to understand
- Safety systems are not always working
  - some are down, and known to be
  - some are accidentally turned off
  - some are not set properly
  - others fail to work when needed
- There are often not direct indicators of what is happening operators figure it out indirectly

Can this happen elsewhere? Robots?

Techniques for Modelling Systems: Bayes Nets



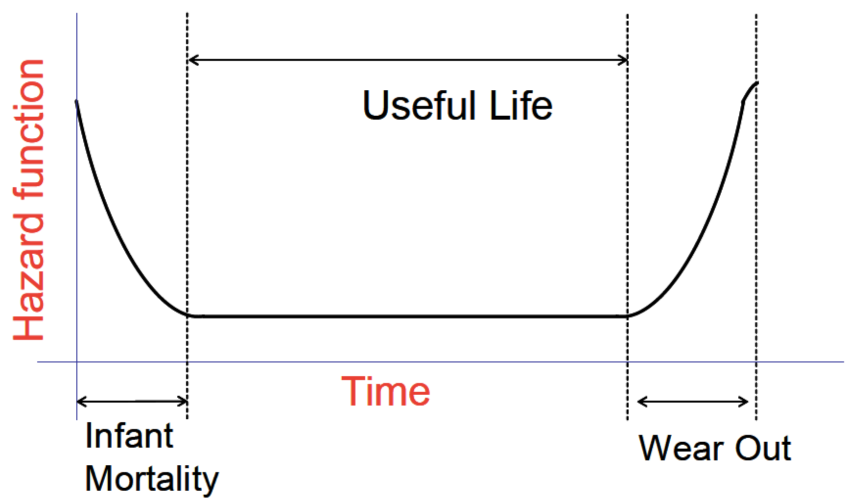
$$P(\lambda/e) = \frac{P(e\lambda)}{P(e)} = \frac{P(e/\lambda)P(\lambda)}{P(e)}$$



Big Issue: How to get probability tables?

Models of Reliability

- How to model the reliability performance of components
- Reliability performance of a large sample of homogenous items: if we observe the items over their lifetime without replacement then we can observe three distinct periods, the so-called bathtub curve



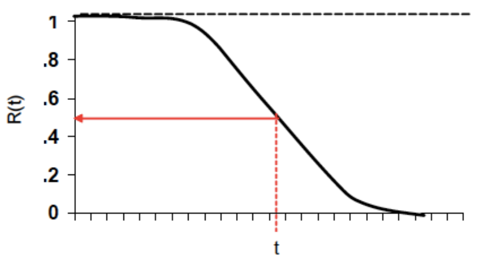
A Reliability Model

- Recall that density function and cumulative density functions can be represented as f and F, such that,

$$p(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx \qquad F(t) = \int_{-\infty}^{t} f(t)dt$$

- A survival function is defined as,

$$R(t) = 1 - F(t) = 1 - \int_{-\infty}^{t} f(t)dt$$



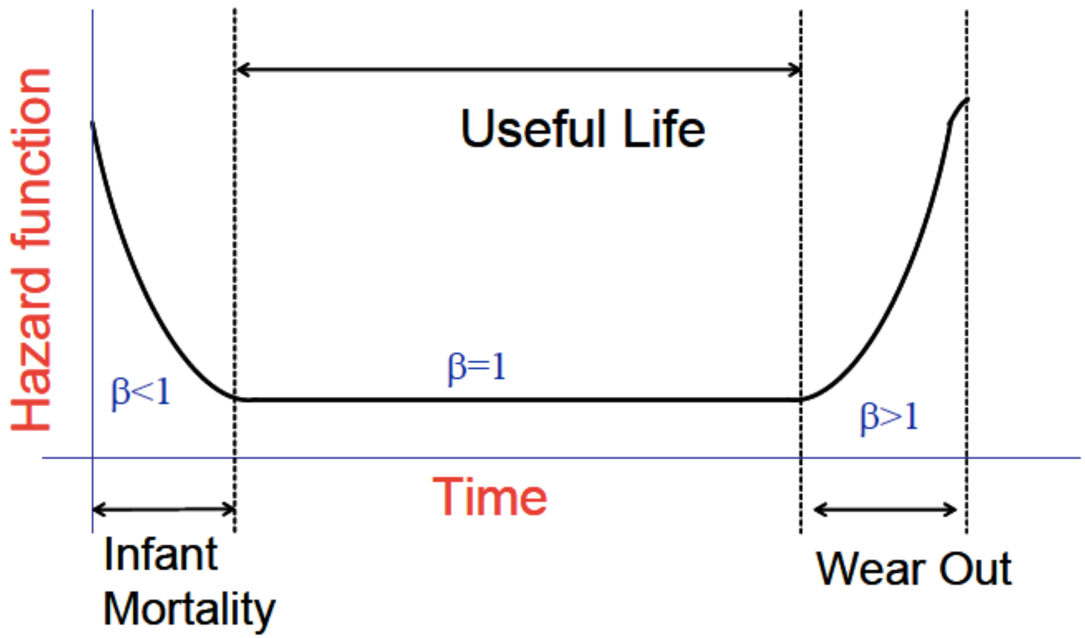
- Based on this, define hazard as,

$$h(t) = \frac{f(t)}{R(t)} = 1 - \frac{f(t)}{F(t)}$$

Conditional probability of failure in the interval t to ($t+dt$), given that no failure has occurred by t.

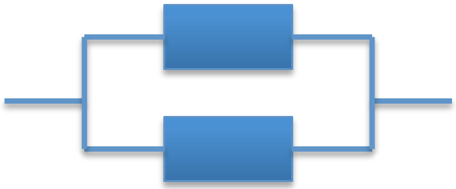
Weibull Distribution for Reliability

$$R(t) = e^{-(\frac{t}{\eta})\beta}$$



Reliability after Composition (Redundancy)

- Often engineering systems have redundant components
- So, failure of the composite is of interest, e.g.,



$$P(X_1, X_2|H)$$

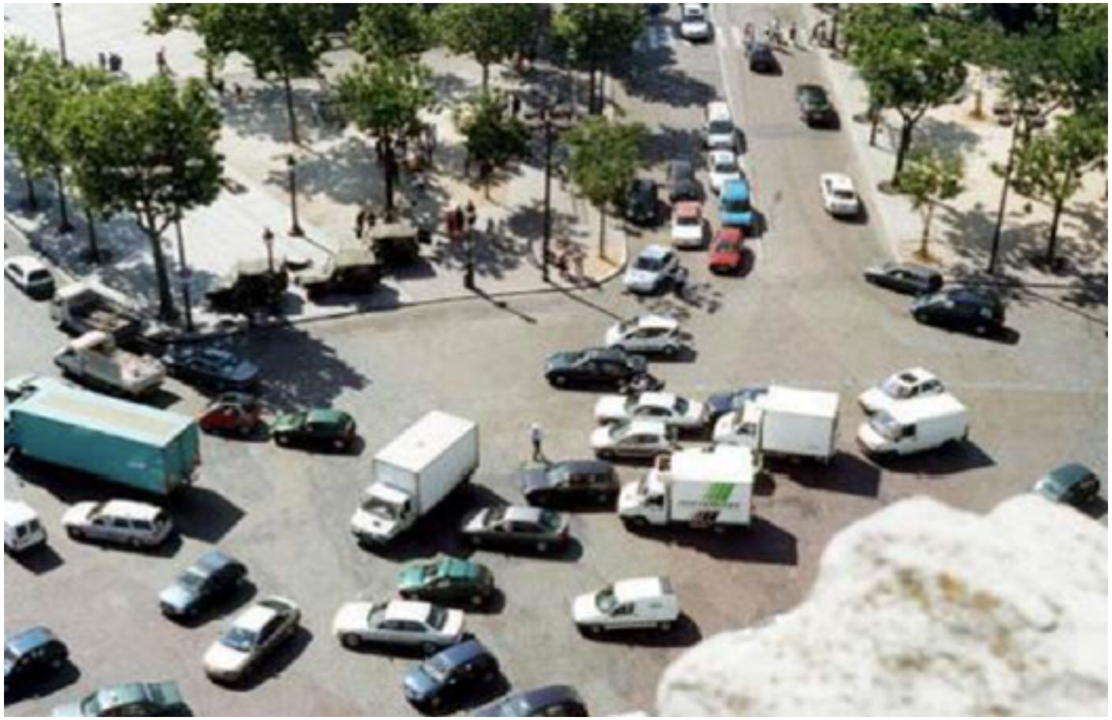
$$X_i \implies X_i \leq x_i$$

$$P(X_1, X_2|H) = \sum_{\theta_1,\theta_2} \{P(X_1, X_2|\theta_1, \theta_2)P(\theta_1, \theta_2|H)\}$$

- In a few cases, e.g., exponential distribution based failure models, this may yield expressions wherein survival function R can be additively written based on model structure, etc.

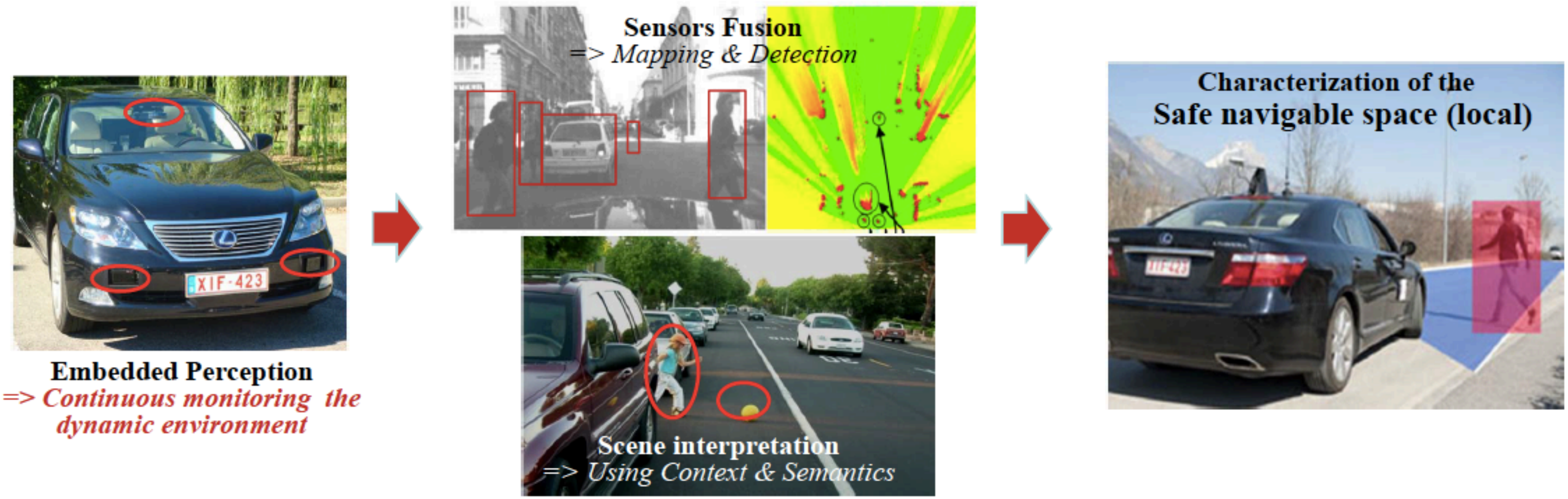
Concrete Problem: Driver Assist & Autonomous Vehicles

Why Difficult? Typical Operating Scenarios



Issues:
- Dynamic & Open Environments
- Incompleteness & Uncertainty (Model & Perception)
- Human in the loop (Social & Interaction Constraints)

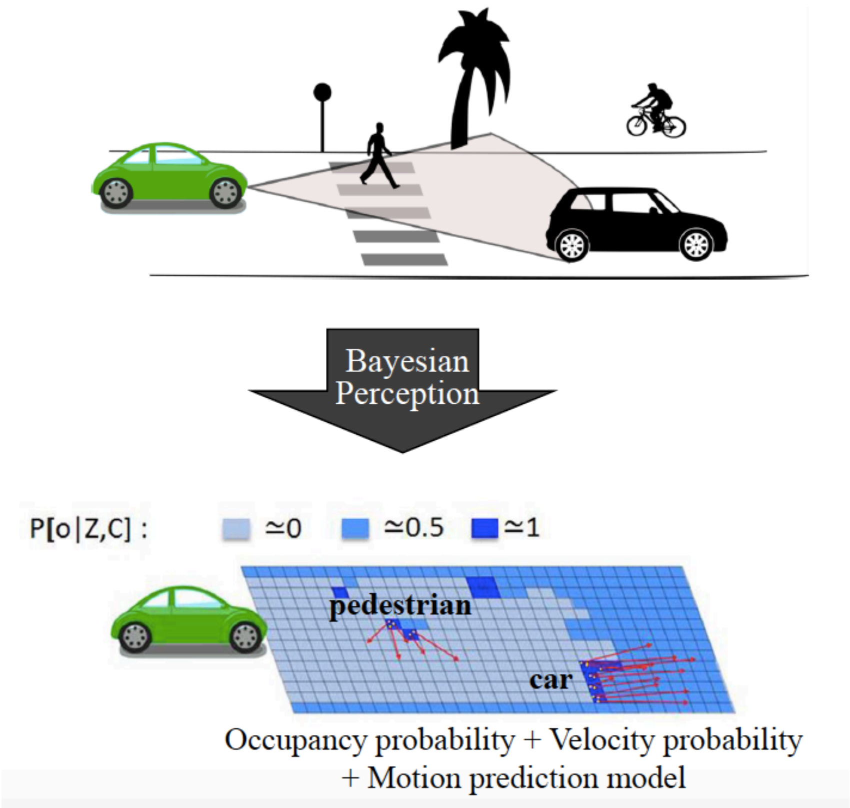
Bayesian Perception Approach



[Source: C. Laugier et al.]

Bayesian Perception for ADAS/AV

- Estimate Spatial occupancy
- Analyze Motion Field (using Bayesian filtering)
- Initially, reason at the Grid level (no object segmentation, just prob. of occupancy, o, given observation z and state c)
- Then, register other objects on top of this data structure



[Source: C. Laugier et al.]

Can We Ensure Safety, Always?

- Active topic of discussion
- Consider some examples due to A. Shahua*

Q: Can the central car avoid all collisions?

* [S. Shalev-Shwartz, S. Shammah, A. Shashua, On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374, 2017]

Shashua's Approach to Safety

In practice, the AV needs to know two things:

- Safe State: This is a state where there is no risk that the AV will cause an accident, even if other vehicles take unpredictable or reckless actions.
- Default Emergency Policy: This is a concept that defines the most aggressive evasive action that an AV can take to maintain or return to a Safe State.

They coin the term Cautious Command to represent the complete set of commands that maintains a Safe State. Set a hard rule that the AV will never make a command outside of the set of Cautious Commands. This ensures that the planning module itself will never cause an accident.

Example: Safe Longitudinal Distance

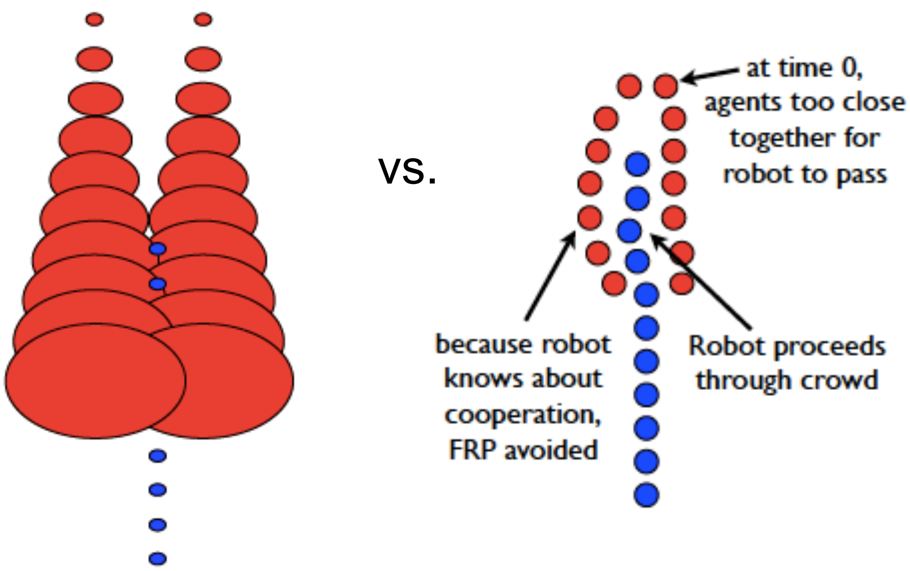
Safe Distance Formula

$$d_{\min} = L + T_f\left[v_r - v_f + \rho\left(a_a + a_b\right)\right] - \frac{\rho^2 a_b}{2} + \frac{(T_r - T_f)(v_r + \rho\, a_a - (T_f - \rho)a_b)}{2}$$

- L is the average length of the vehicles
- ρ is the response time of the rear vehicle
- v_r, v_f are the velocities of the rear/front vehicles
- a_a, a_b are the maximal acceleration/braking of the vehicles
- T_f is the time for the front car to reach a full stop if it would apply maximal braking
- T_r is the time for the rear car to reach a full stop if it would apply maximal acceleration during the response time, and from there on maximal braking

Prospects for the Future

- Can more generic policy learning, e.g., reinforcement learning to navigate past crowded intersection, respect these?
  - Formal methods often use concept of “guard condition”
  - How best to define in open-ended environment?
- Can we avoid “frozen robots”?



[Trautman + Krause, IROS 2010]