An Introduction to Intelligent Agents

Dominik Herrmann

Basel, Feb 7, 2014

Slides available at **http://dhgo.to/agents**

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Dominik Herrmann

Security in Distributed Systems Group (Prof. Hannes Federrath)

Interests: Application of Machine Learning to Privacy & Security

Motivation for research in Artificial Intelligence

Can we build intelligent machines?

How? What can they do for us?

Motivation for research in Artificial Intelligence

Can we build intelligent machines?
How? What can they do for us?

1958, H. A. Simon and Allen Newell: “within ten years a digital computer will be the **world's chess champion**” and “within ten years a digital computer will **discover and prove an important new mathematical theorem**.”

1965, H. A. Simon: “machines will be capable, within twenty years, of **doing any work a man can do**.”

1967, Marvin Minsky: “Within a generation ... the problem of creating 'artificial intelligence' will substantially be **solved**.”

1970, Marvin Minsky: “In from three to eight years we will have a machine with the **general intelligence of an average human being**.”

Artificial Intelligence research has resulted in numerous dumb specialists, but not in a truly intelligent machine.

http://www.pcmag.com/article2/0,2817,2404803,00.asp

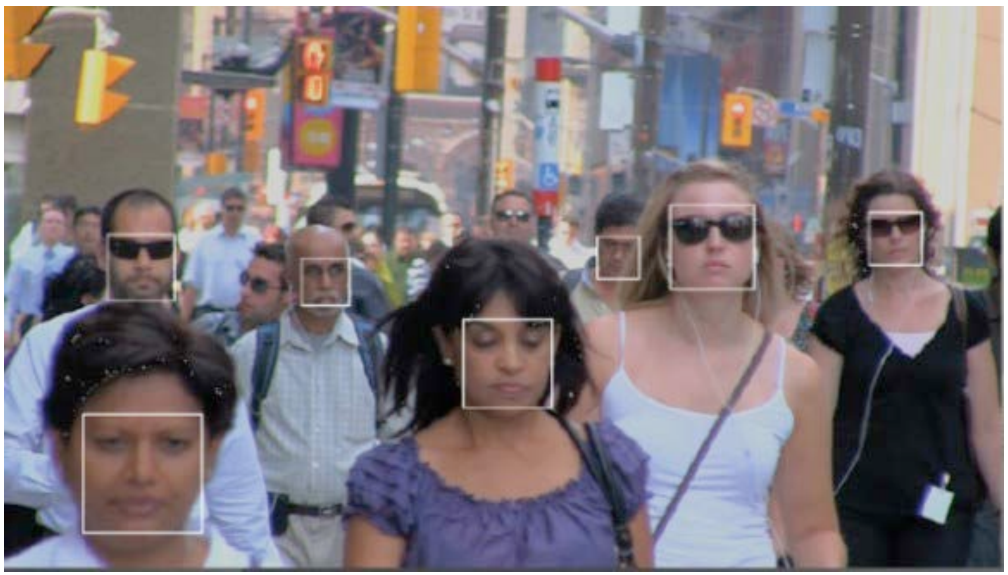
- **find the best route through cross-city traffic**
- suggest movies and recommend products
- recognize objects
- recognize faces in photos
- translate documents
- transcribe spoken text
- play checkers and chess
- play Jeopardy

Artificial Intelligence research has resulted in numerous dumb specialists, but not in a truly intelligent machine.

http://www.youtube.com/watch?v=_d0LfkIut2M

- find the best route through cross-city traffic
- suggest movies and recommend products
- recognize objects
- recognize faces in photos
- translate documents
- transcribe spoken text
- play checkers and chess
- play Jeopardy

Artificial Intelligence research has resulted in numerous dumb specialists, but not in a truly intelligent machine.



www.techradar.com/news/world-of-tech/10-bits-of-tech-to-scare-you-witless-1156960

- find the best route through cross-city traffic
- suggest movies and recommend products
- recognize objects
- **recognize faces in photos**
- translate documents
- transcribe spoken text
- play checkers and chess
- play Jeopardy

Artificial Intelligence research has resulted in numerous dumb specialists, but not in a truly intelligent machine.

http://www.mav-engineering.com/BLUEKINGDOM_PIC_7.jpg

- find the best route through cross-city traffic
- suggest movies and recommend products
- recognize objects
- recognize faces in photos
- translate documents
- transcribe spoken text
- play checkers and **chess**
- play Jeopardy

Artificial Intelligence research has resulted in numerous dumb specialists, but not in a truly intelligent machine.

http://mentalfloss.com/article/51543/what-ibm-watson-7-videos-jeopardy-era

- find the best route through cross-city traffic
- suggest movies and recommend products
- recognize objects
- recognize faces in photos
- translate documents
- transcribe spoken text
- play checkers and chess
- **play Jeopardy**

Intelligent agents seem to be the vehicle that makes results from Artificial Intelligence tangible for the masses – finally.

Initial Question
What are intelligent agents?

The most prominent example of intelligent agents are self-driving cars.

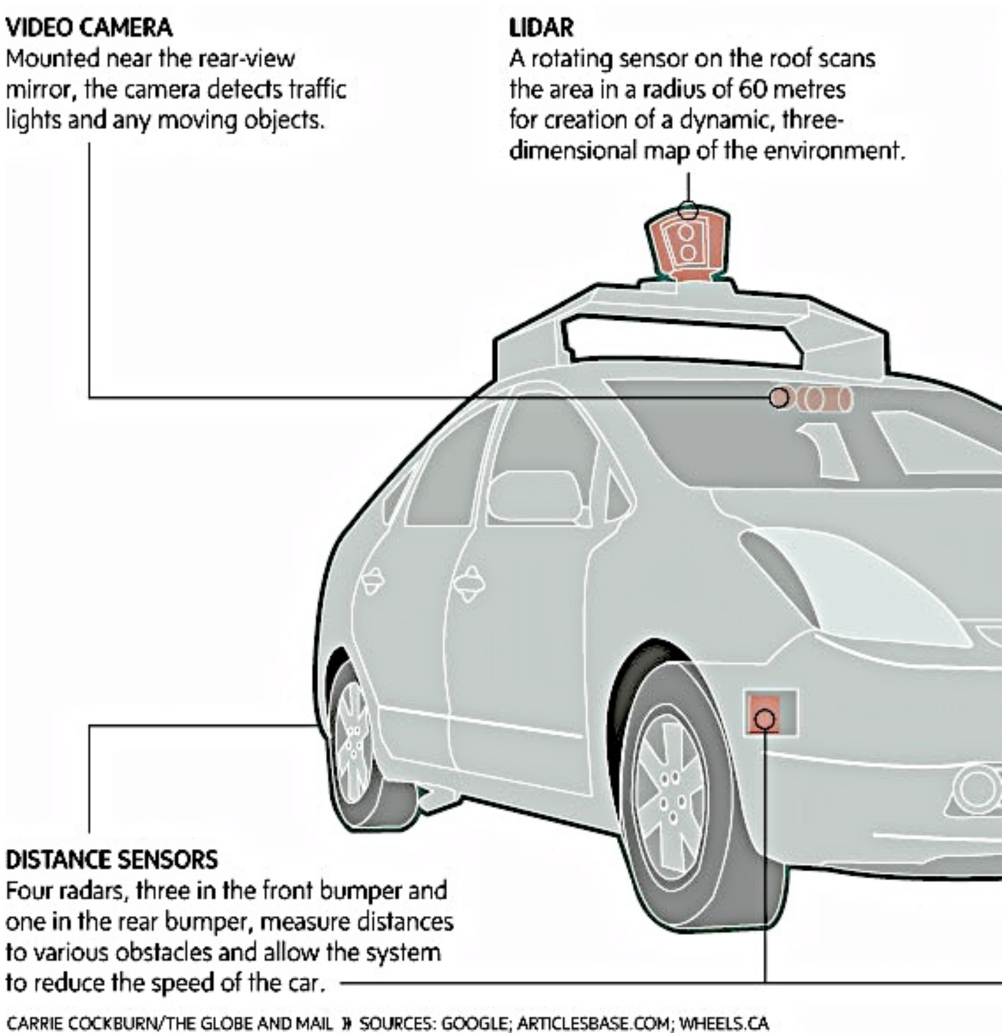
http://www.livescience.com/722-darpa-robot-challenge-urban.html

- Volkswagen Stanley won the 2005 DARPA Grand Challenge (offroad)
- uses machine learning for obstacle detection
- autonomous driving to be available in consumer cars by 2019 (KPMG, 2012)

Tests with self-driving cars on roads are promising.

http://www.youtube.com/watch?v=LHqB47F12vI

An autonomous car faces four challenges that rely on techniques from the field of “artificial intelligence”.

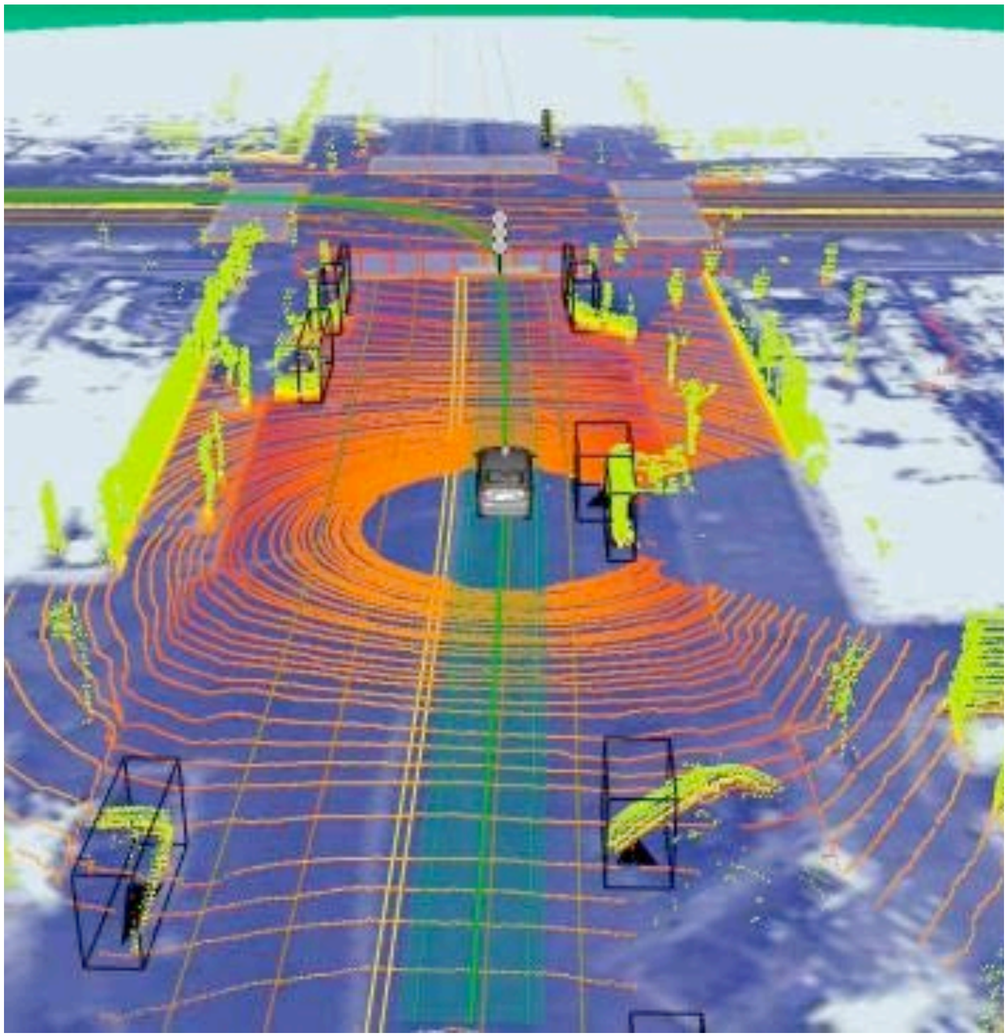


http://manilagawker.blogspot.de/2012/09/the-car-of-future-googles-self-driving.html

Challenges:

- perception
- planning
- decision making
- interacting with environment

An autonomous car faces four challenges that rely on techniques from the field of “artificial intelligence”.



http://asset1.cbsistatic.com/cnwk.1d/i/tim2/2013/08/30/Goog-self-driving-1_610x363.jpg

Challenges:

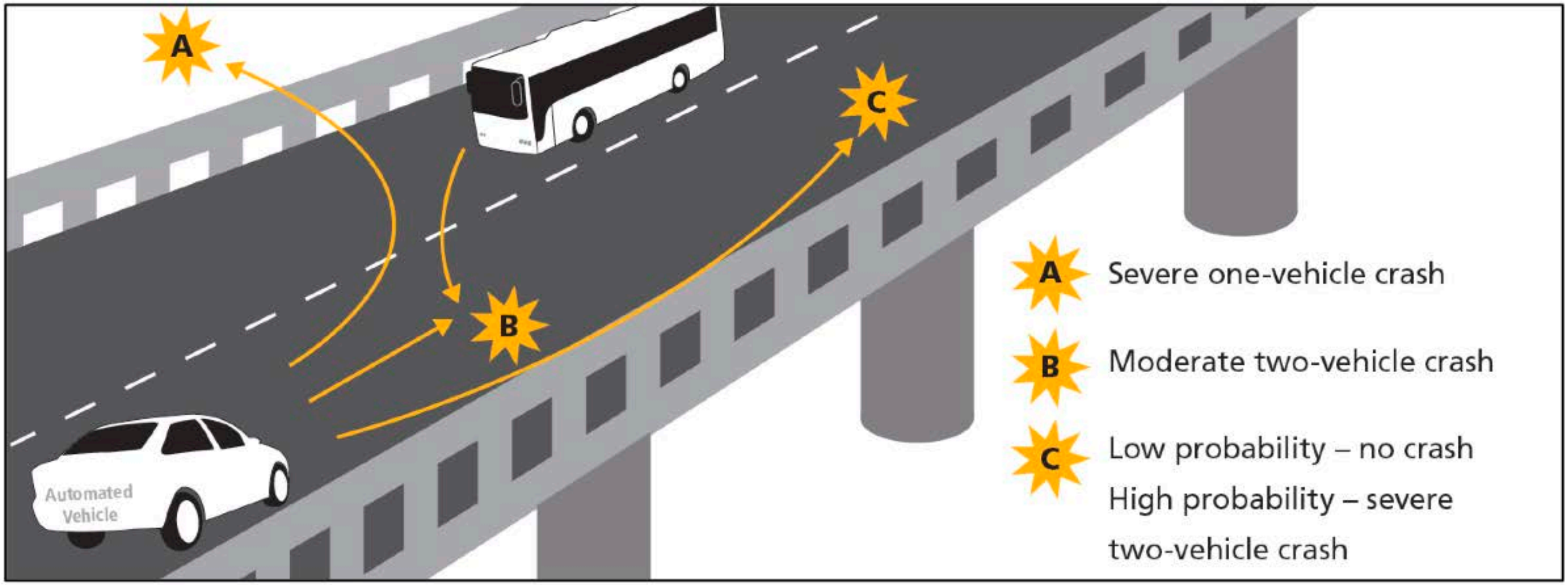
- perception
- planning
- decision making
- interacting with environment

Benefits:

- no human errors
- more efficient driving
- free time in the car

Risks?

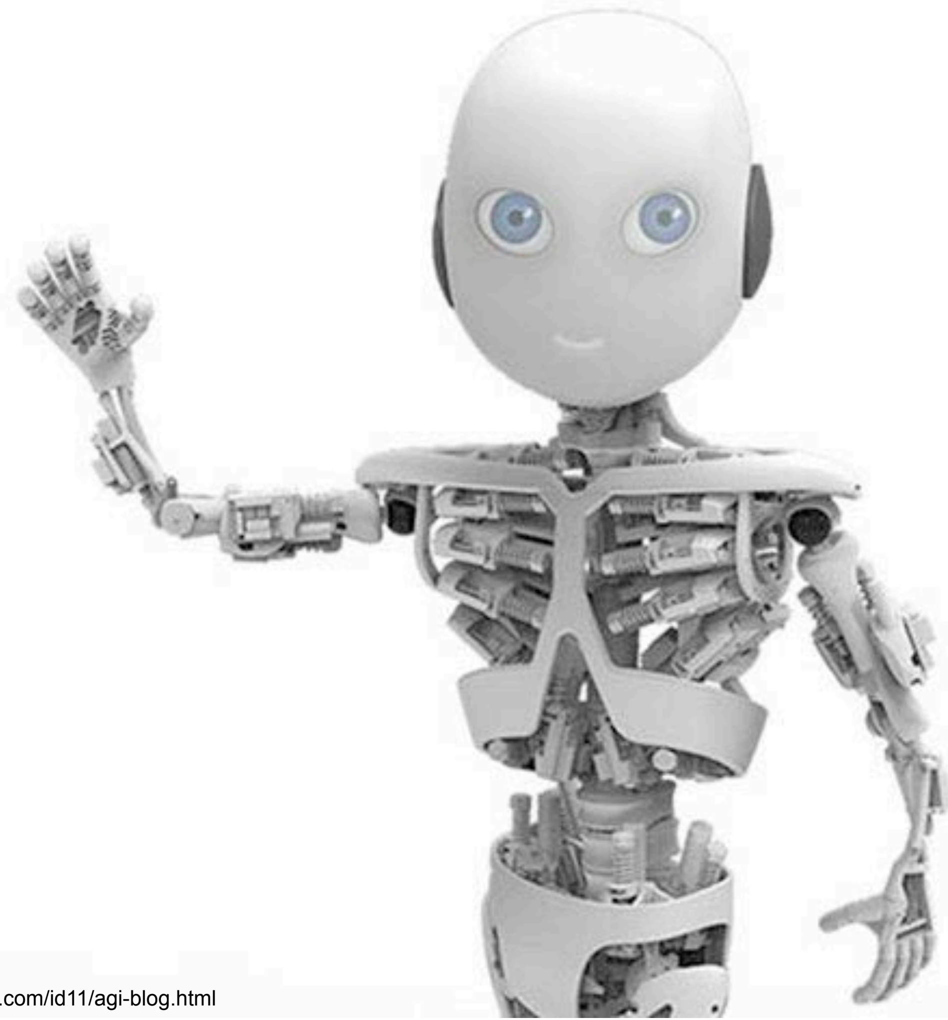
Advocates of self-driving cars argue that computers are better at making tough decisions under stress. Critics warn of ethical and moral issues.



http://people.virginia.edu/~njg2q/ethics.pdf

Cars are the most prominent example of intelligent agents.
But autonomous robots have already entered the consumer market.

Roboy (University of Zurich)

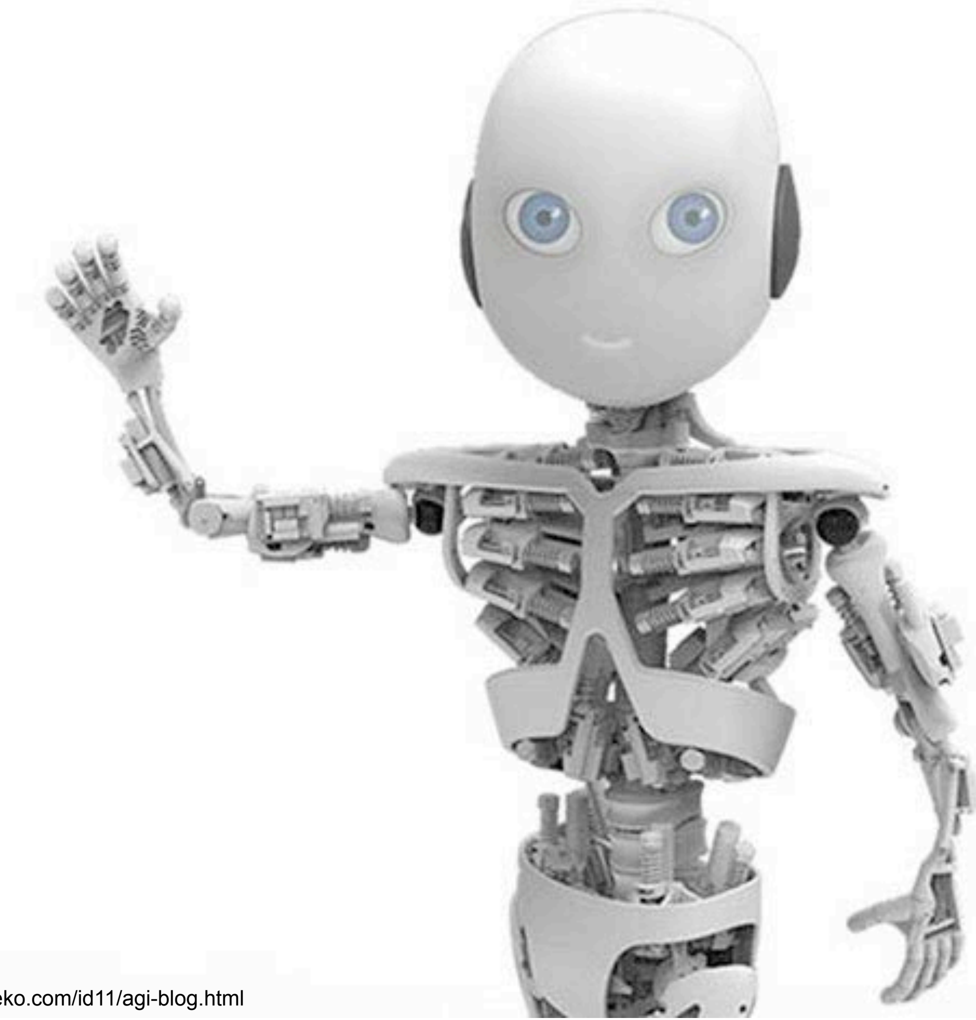


http://www.sony-aibo.com/aibo-models/sony-aibo-ers110/ and http://quoteko.com/id11/agi-blog.html

Autonomous robots have interesting applications for healthcare.

Care-o-Bot 3 (Bayer)

Roboy (University of Zurich)



http://www.bayer.com.tr/ebbsc/cms/en/news/news0021.html and http://quoteko.com/id11/agi-blog.html

The use of autonomous robots for military applications is subject of a controversial debate.

Care-o-Bot 3 (Bayer)

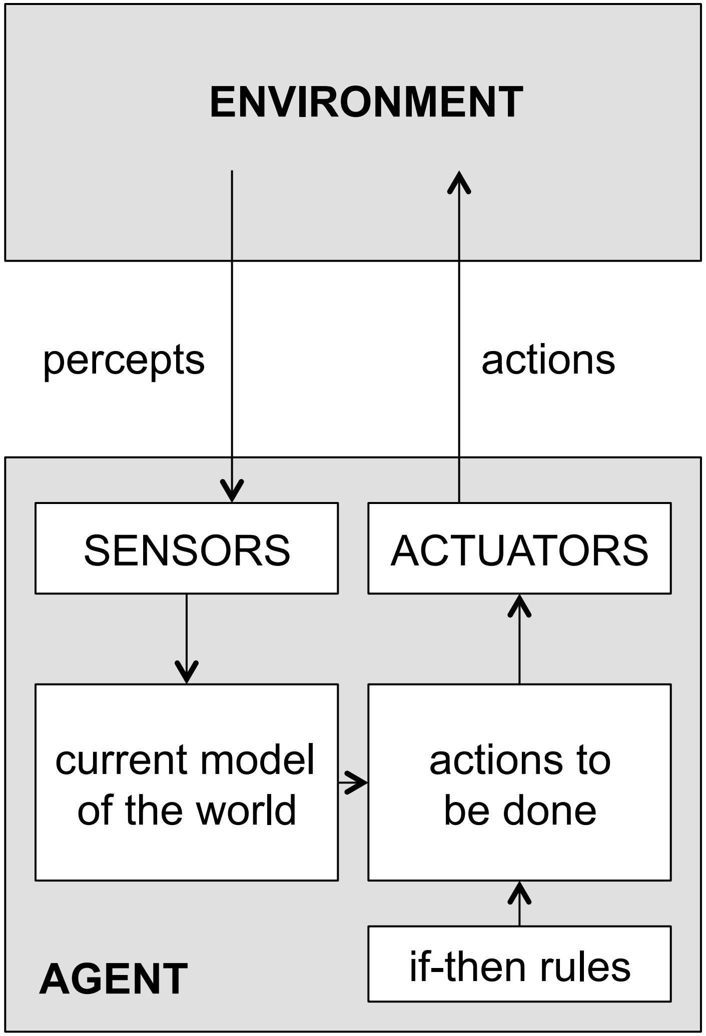
http://www.bayer.com.tr/ebbsc/cms/en/news/news0021.html and http://www.sott.net/image/image/s6/129832/full/Time_Mags_Drones.jpg

The objective of this talk is to provide a foundation for a discussion of legal implications and liability issues.

Questions addressed in this talk (and in the discussion)
What are intelligent agents?
How are they different from conventional software?
What are potential risks?
What techniques are used and what are their limitations?
Can engineers foresee the actions of agents?
How can engineers ensure controllability?

Intelligent agents are autonomous entities having distinct characteristics that distinguish them from regular computer programs.

- accommodate new **problem solving** rules incrementally
- **adapt** online and **in real time**
- are able to **analyze themselves** in terms of behavior, error and success
- **learn and improve** through interaction with the environment (embodiment)
- learn quickly from **large amounts of data**
- have memory-based exemplar **storage** and retrieval capacities
- have **parameters** to represent short and long term memory, age, forgetting, etc.



N. Kasabov, Introduction: Hybrid intelligent adaptive systems. International Journal of Intelligent Systems, Vol.6, (1998) 453–454 and Russell & Norvig, Artificial Intelligence: A Modern Approach, 2nd ed. (2003) as cited in http://en.wikipedia.org/wiki/Intelligent_agent

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Autonomy

Mobility

Indeterminacy

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Autonomy	Actions of conventional software are ultimately caused by its user.	Agents are in control of their own actions.
	Software responds synchronously and gives feedback to the user.	They operate asynchronously without constant interaction.
	Software is merely a tool. Damage is either the fault of the user or the developer.	If an agent causes damage, causality and liability are questionable.

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Autonomy	Actions of conventional software are ultimately caused by its user.	Agents are in control of their own actions.

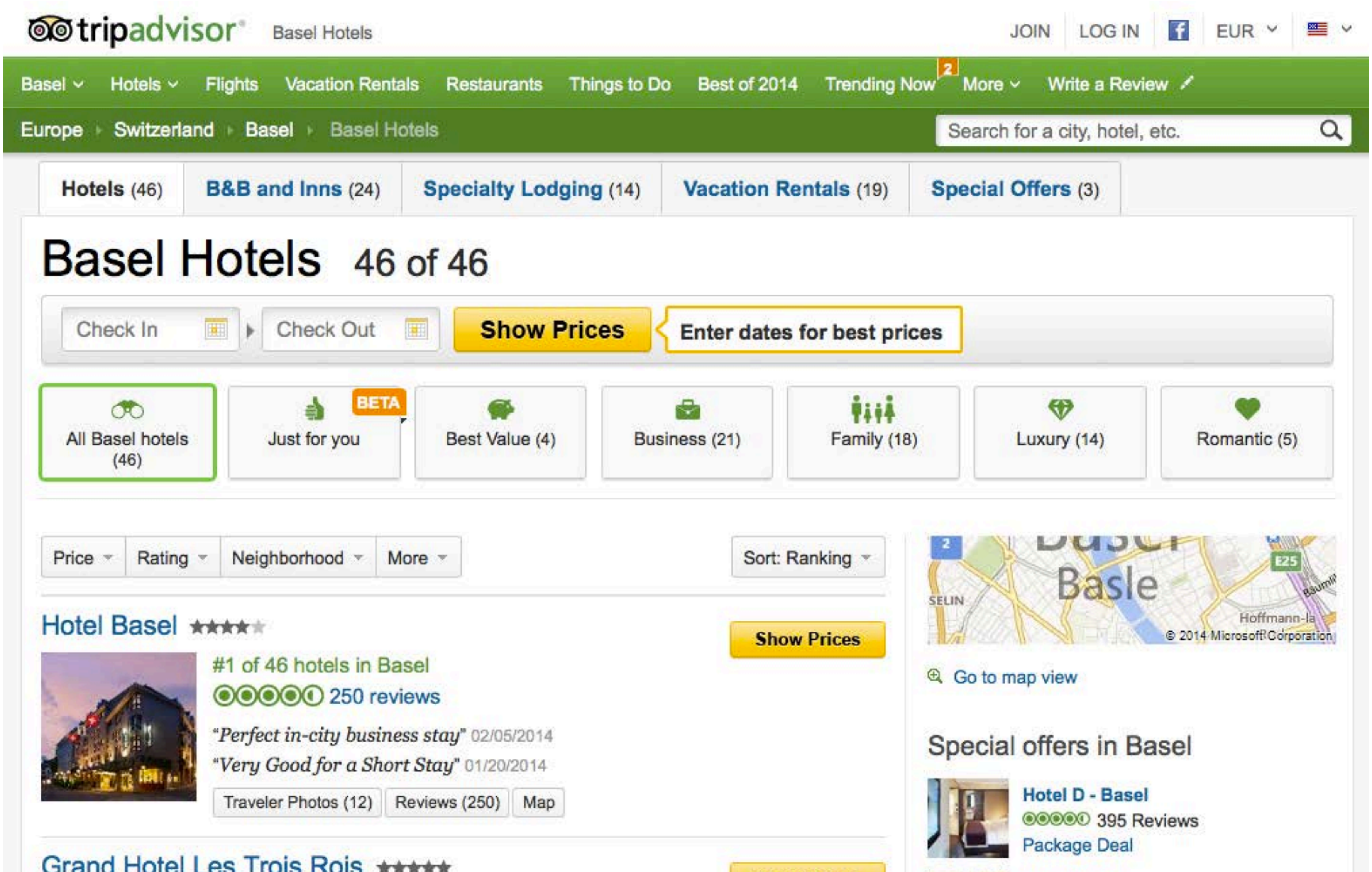
Users might argue “but **my agent** did it”; however, this sort of defense has not been accepted in court so far.

“There is no such thing as a ‘computer mistake’. Microchips are too dumb to do anything except follow instructions.”

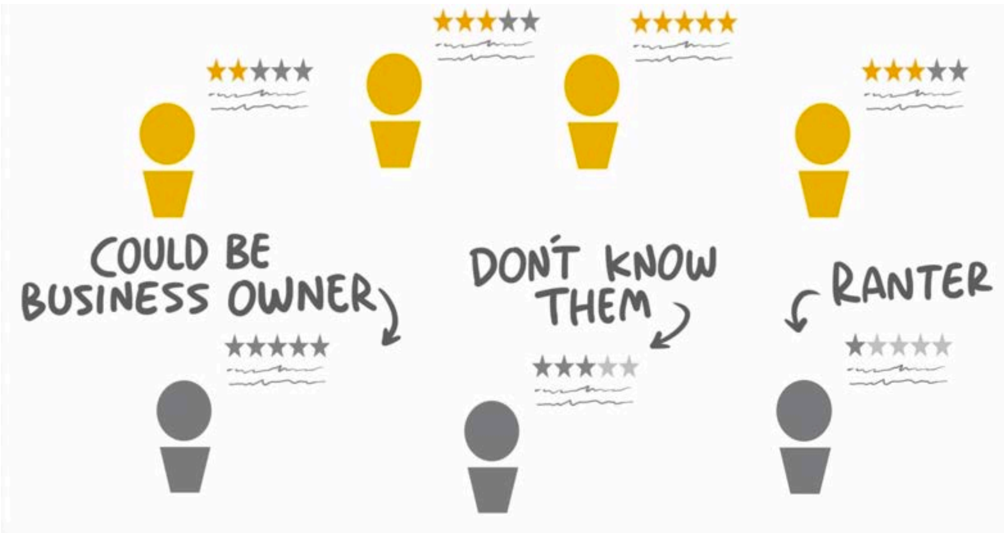
Users might blame engineers, who would argue that the design is adequate and professionally executed. But is it?

S. E. Miller, Civilizing Cyberspace. ACM Press, New York, NY, 19

Web portals that collect feedback for businesses want to detect fake reviews that try to increase/decrease the average rating (shilling).

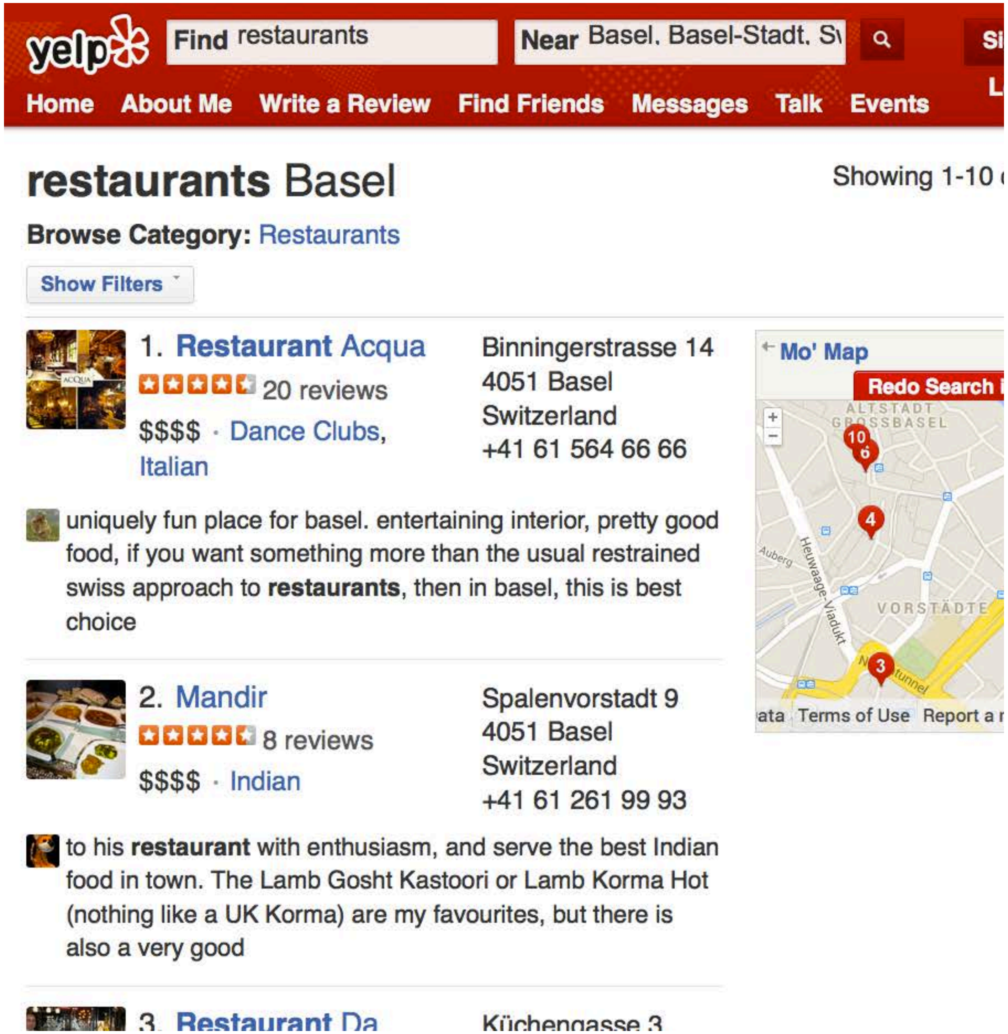


Machine learning techniques are used to prevent shilling.



- **yelp** analyzes all ratings daily and tags fishy reviews as “not recommended”
  - users with low activity
  - strong rants / overly praise
- avg. rating: “recommended” reviews only

http://www.youtube.com/watch?v=PniMEnM89iY



- **yelp** analyzes all ratings daily and tags fishy reviews as “not recommended”
  - users with low activity
  - strong rants / overly praise
- avg. rating: “recommended” reviews only
- high avg. rating => publicity

- **yelp** analyzes all ratings daily and tags fishy reviews as “not recommended”
  - users with low activity
  - strong rants / overly praise
- avg. rating: “recommended” reviews only
- high avg. rating => publicity

- the reasoning in the decision process is not disclosed
- a dentist from Hamburg **sued yelp and won**

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Autonomy	Actions of conventional software are ultimately caused by its user.	Agents are in control of their own actions.

Users might argue “but **my agent** did it”; however, this sort of defense has not been accepted in court so far.

“There is no such thing as a ‘computer mistake’. Microchips are too dumb to do anything except follow instructions.”

Users might blame engineers, who would argue that the design is adequate and professionally executed. But is it?

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Mobility	Conventional software is executed within the boundaries of the user's system.	Mobile agents may be designed to freely wander the network or the physical world.
	It remains under the control of the user at all times. The operator can supervise its activities.	They cannot be observed by the user (or the engineers) at all times.

We will look at the differences between agents and conventional software in regard to three properties in more detail.

Mobility	Conventional software is executed within the boundaries of the user's	Mobile agents may be designed to freely wander through the network or the

Mobile agents constitute **security** and **privacy** risks:

They might attack or interfere with the environment or other users inadvertently, damaging foreign property or causing outages.

Mobile agents acting on behalf of a user may involuntarily disclose private information about him to other systems, e.g., his location, interests or income.

The agent might be attacked in order to force disclosure.

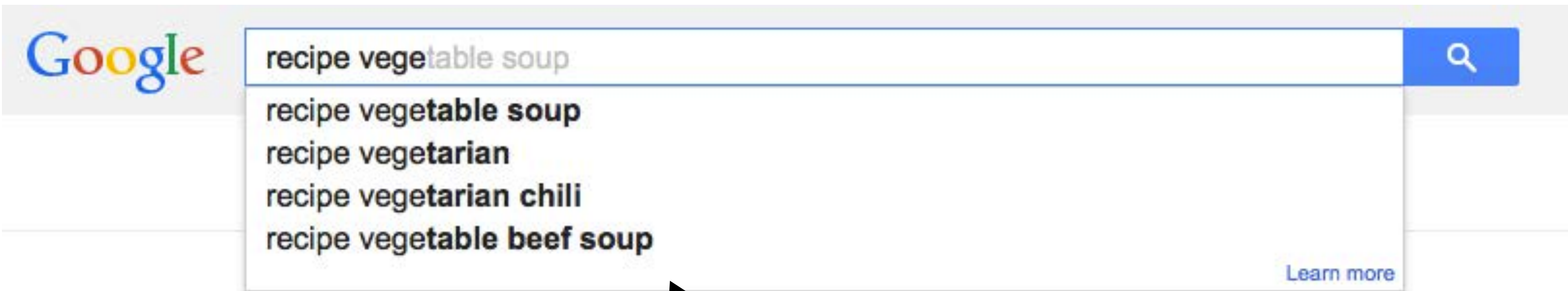
We will look at the differences between agents and conventional software in regard to three properties in more detail.

Indeterminate **environment**	Conventional software operates on well-known inputs and produces well-known outputs.	Agents are meant to **sense and act** in more complex environments, e.g., the physical world.
	It’s behavior is predictable, because it is exectued in a constrained environment, e.g., a desktop machine.	The environment of an agent is more complex and may change over time.
	Exhausive tests can be run during development to ensure correct execution.	Engineers cannot test (and cannot know) all possible environments the agent might face, once it is deployed.

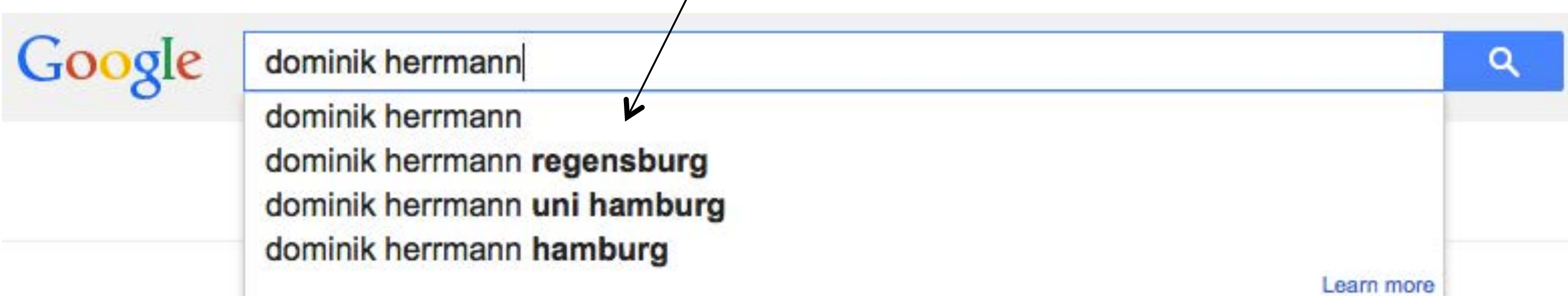
We will look at the differences between agents and conventional software in regard to three properties in more detail.

Indeterminate **users**	Users are trained to adapt to the software.	Agents are meant to adapt to the habits and different behavior of humans.
	Configuration options that allow for personalization are limited and do not change the functionality of the software.	Agents may be designed to be adaptive by **online learning**, after having been deployed. Online learning leads to indeterministic behavior, which is difficult to foresee during development.

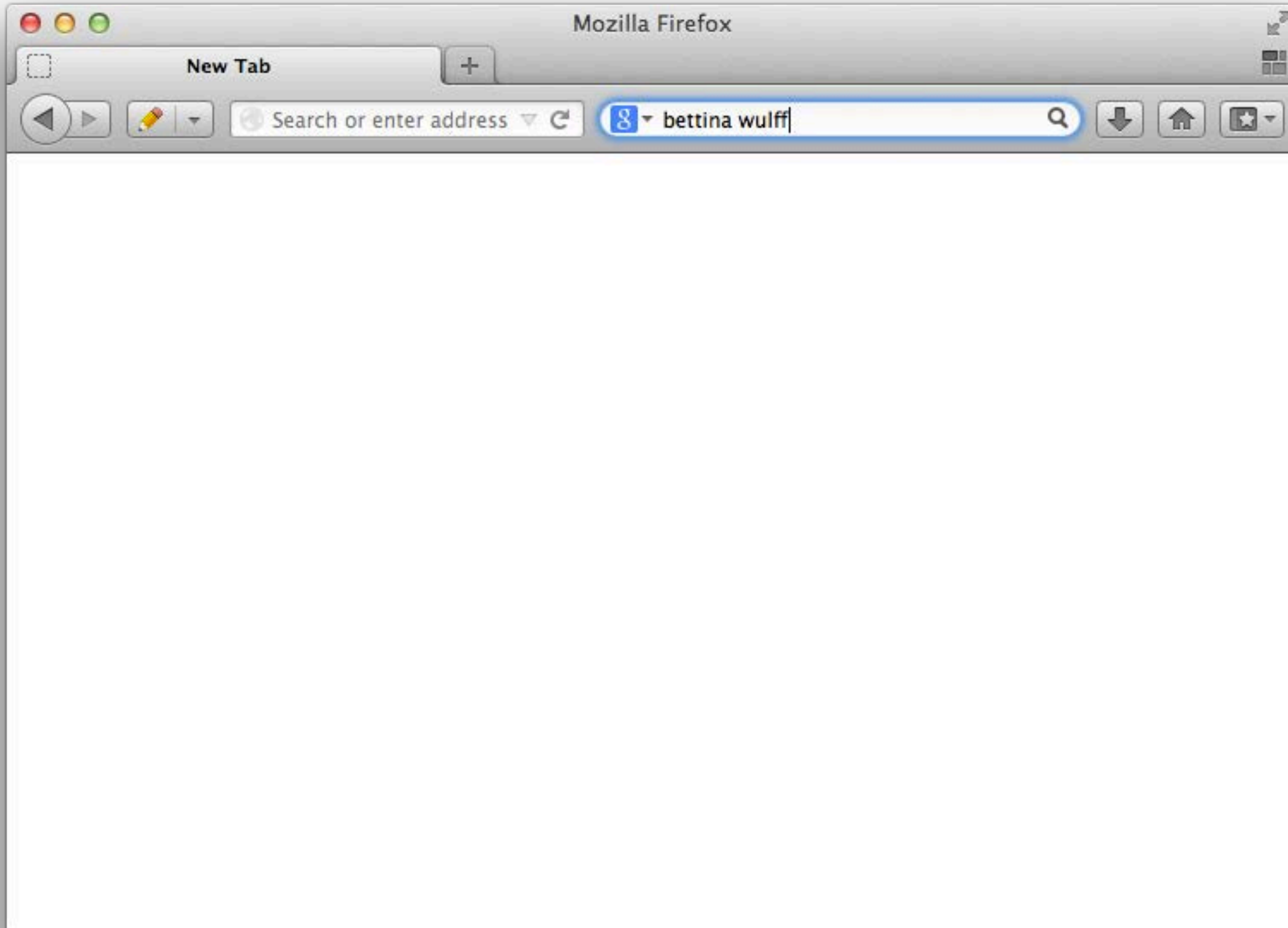
Google Suggest provides useful search recommendations to users while they are typing.



based on recent search queries of other people in the same geographic area

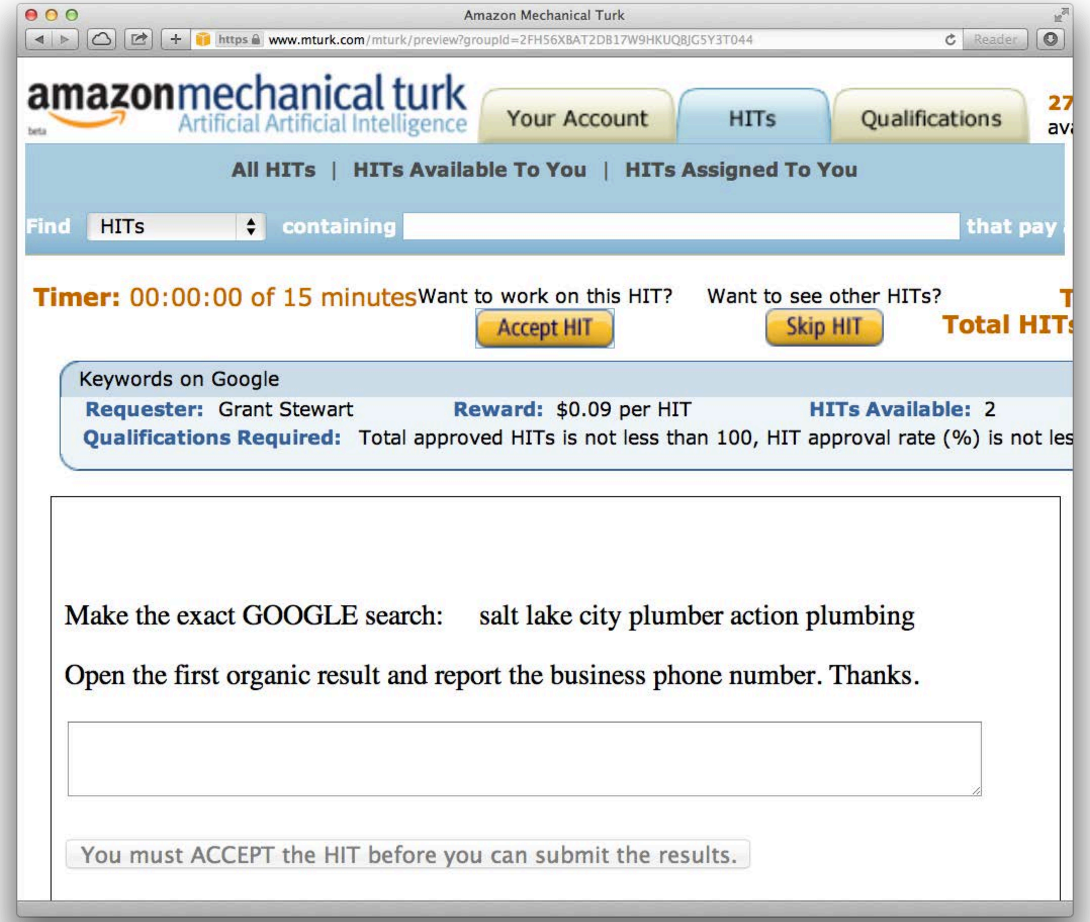


Google Suggest provides useful search recommendations to users while they are typing. This has caused problems.



Google Suggest provides useful search recommendations to users while they are typing. This has caused problems – and been brought to court.

Strategic users are exploiting the way Google Suggest works in order to increase their publicity. However, this requires much manpower.



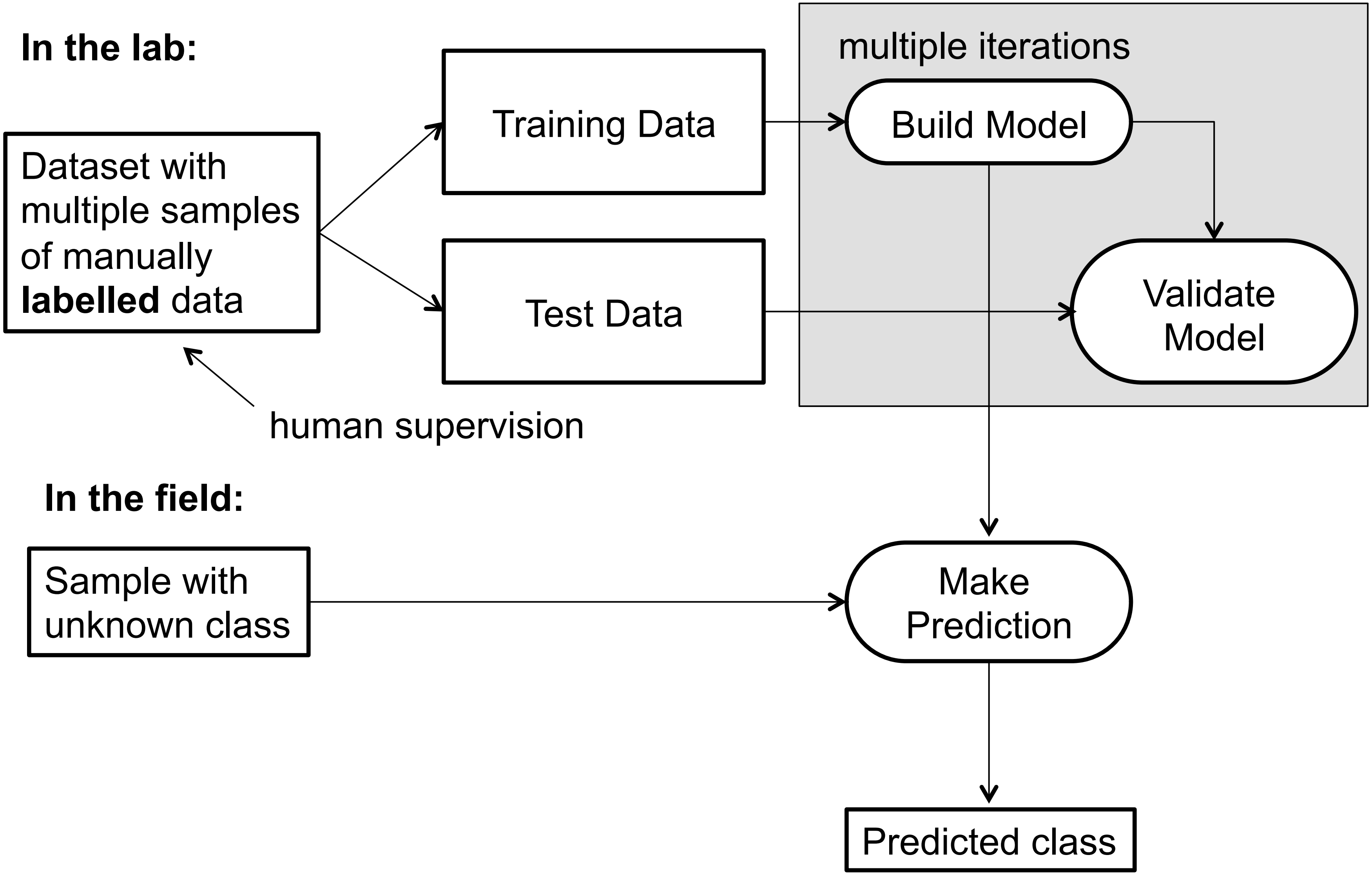
Open Questions

Why is it difficult to foresee the actions of agents?

How can engineers ensure controllability?

What techniques are used?

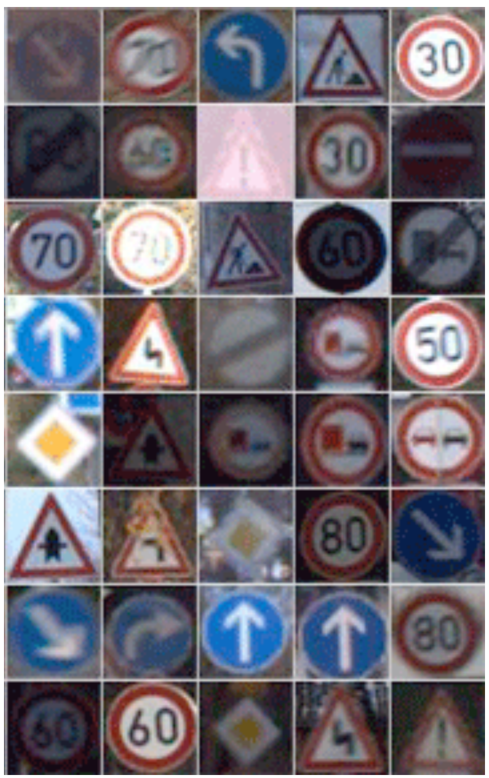
Agent perception and decision making relies on supervised machine learning techniques.

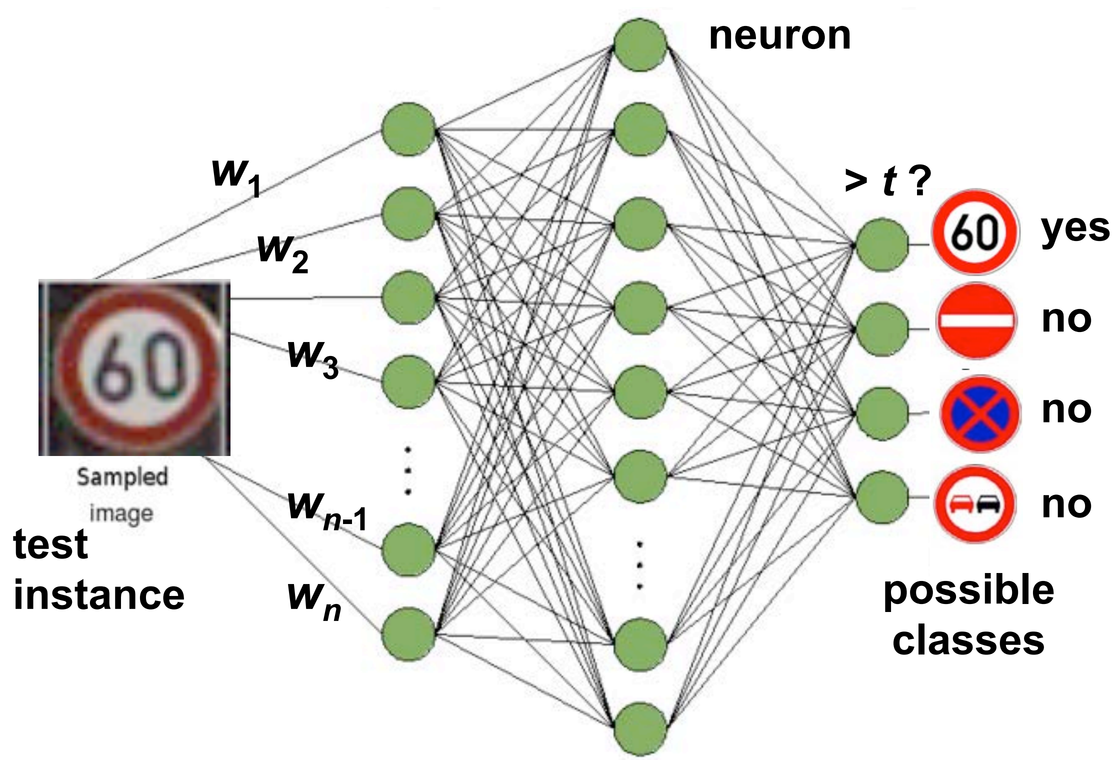


Example of supervised learning: Traffic sign detection, a computer vision task that can be solved with artificial neural networks (ANNs).

The ANN learns to differentiate by training it with real-world images.

Neurons get activated when their input value reaches a certain **threshold value**. The output of a neuron is adjusted with a **weight value**. These values are learned during training.





http://www.idsia.ch/~juergen/trafficsigns288a.gif and http://en.wikibooks.org/wiki/Cyberbotics%27_Robot_Curriculum/Advanced_Programming_Exercises

Open Questions

Why is it difficult to foresee the actions of intelligent agents?
How can engineers ensure controllability?

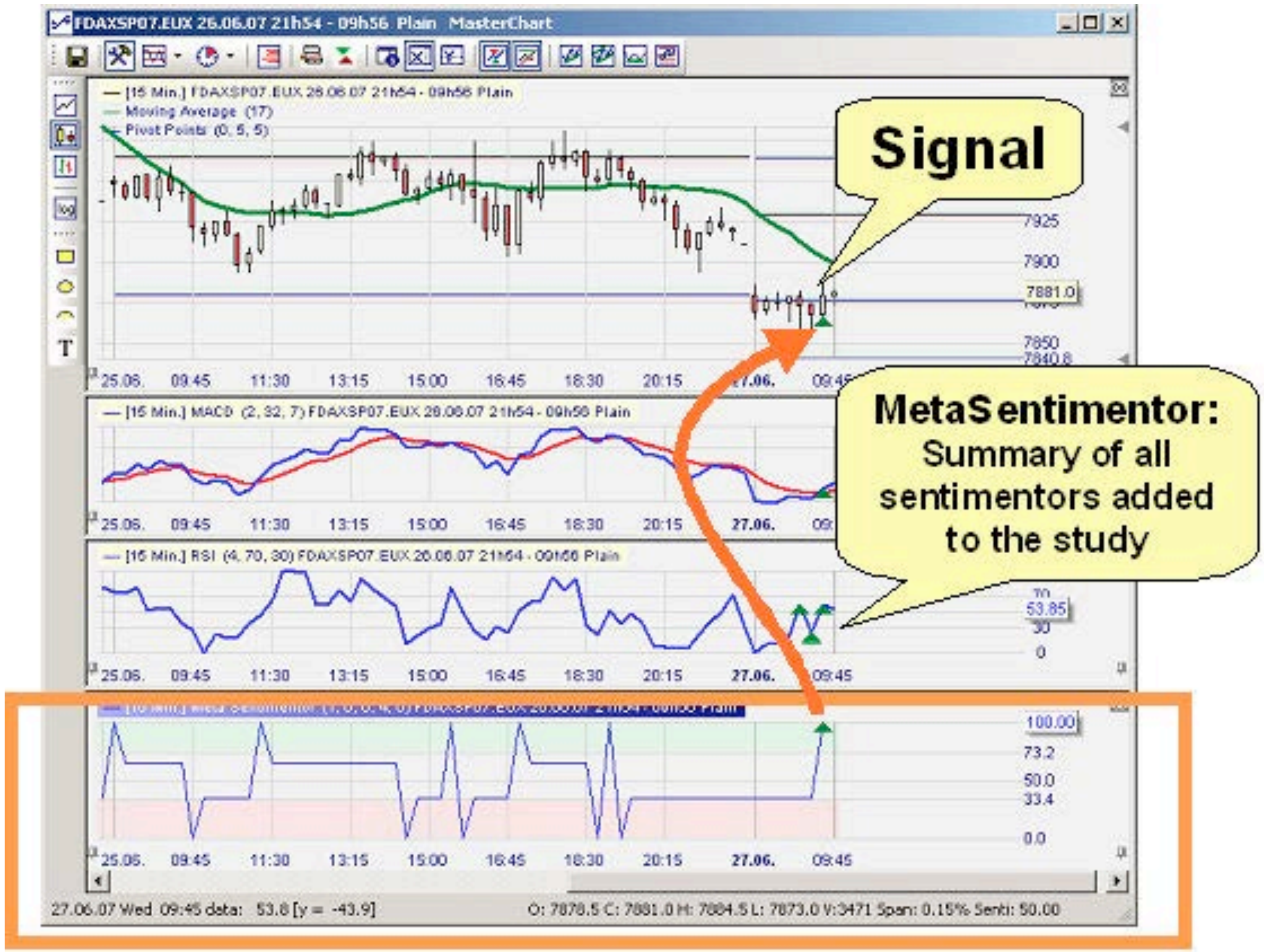
Reasons:

- autonomy
- machine learning
- mobility

Regarding the first, have a a look at the stock market, where we can see what happens when “intelligent” agents are allowed to act **autonomously**.

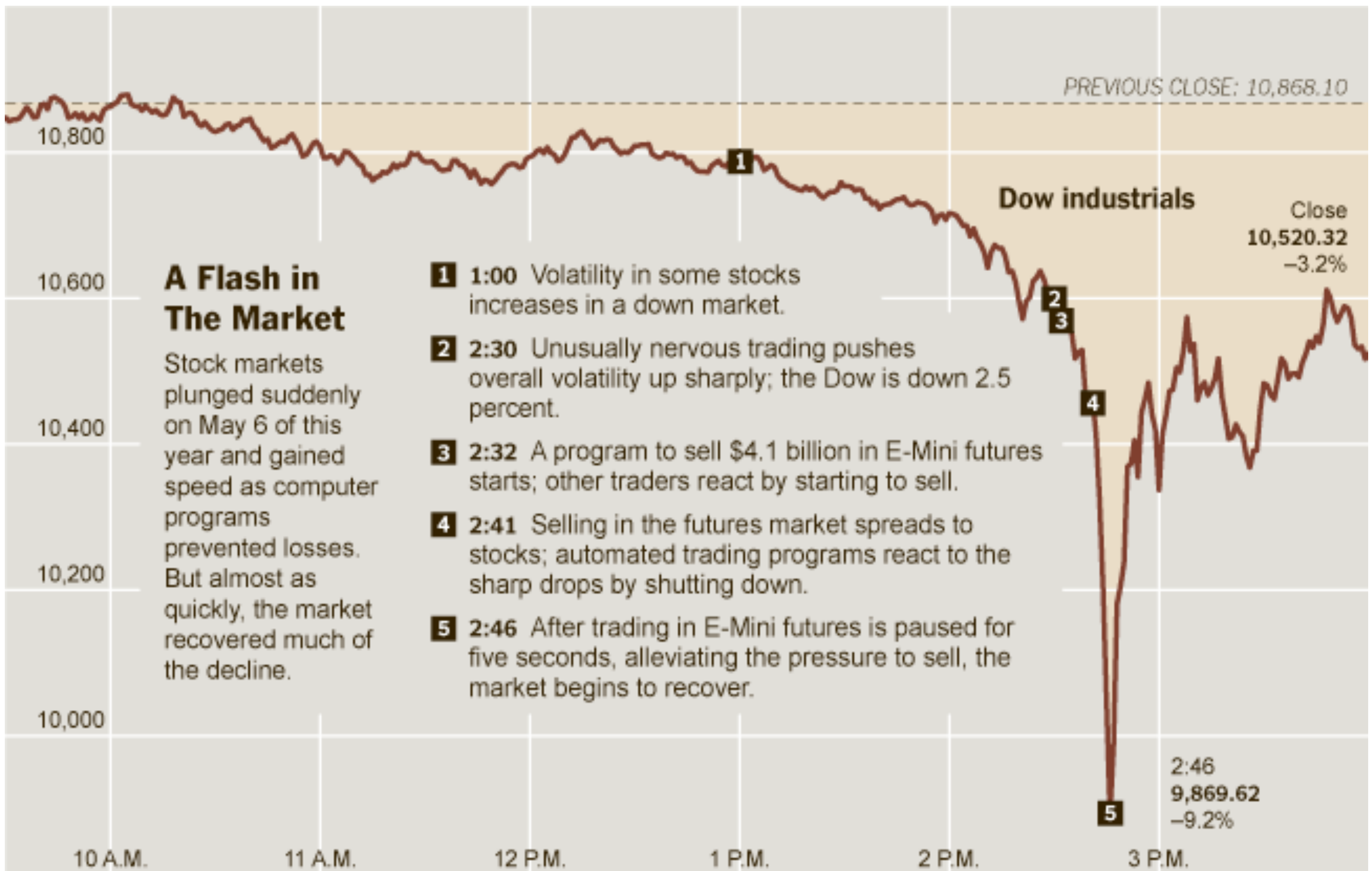
http://access.ncsa.illinois.edu/Stories/97Stories/KUFRINfinalgraph.gif

Algorithmic trading tools use built-in models to autonomously buy and sell stocks based on signals within milliseconds (high-frequency trading).



http://www.fipertec.com/TradingSystems.html

Financial regulators were surprised when algorithmic trading programs sold large volumes within minutes, leading to the 2010 Flash Crash.



Sources: Bloomberg (Dow industrials); Securities and Exchange Commission
THE NEW YORK TIMES

http://www.ritholtz.com/blog/wp-content/uploads/2010/10/flash-crash-dow-popup.png and http://www.zerohedge.com/news/2014-01-06/gold-flash-crashes-halts-trading

Regarding the second, using **machine learning** for perception and decision making will inevitably create unforseeable behavior.

- **Closed-world setting:**
  Offline training **in the lab** includes only a subset of all possible data, thus the system will eventually encounter unknown samples, with which it has never been tested. This will cause some degree of indeterministic behavior.

- **Open-world setting:**
  Additional online training **in the field** will increase the unforseeability of actions, because there is no human supervision.

More problematic:
It may be difficult to determine what a machine learning model "knows" and what it does not know.

Therefore, it may become impossible to understand and explain the decisions of an agent.

http://access.ncsa.illinois.edu/Stories/97Stories/KUFRINfinalgraph.gif

The challenge at border crossings consists in determining which passengers should be screened. Expert systems can make this decision.

http://www.bellinghamherald.com/2014/01/22/3432627/border-traffic-into-whatcom-county.html

Features

- car type and make
- number, age and gender of passengers
- license plate
- weekday and time of day

An expert system makes rule-based decisions based on experience of human experts.

(such screening is also done for aircraft passengers)

In case of a wrong decision the engineers might be held liable. With the model they can prove that the system was implemented correctly.

historic data (training instances) **class**

Vehicle Type	Driver Lic Type	Driver Age	Time of Day	High Risk
Passenger	Regular	> = 35	Night	N
Passenger	Regular	> = 35	Day	N
Workvan	Regular	> = 35	Night	Y
Truck	Disabled	> = 35	Night	Y
Truck	Commercial	<35	Night	Y
Truck	Commercial	<35	Day	N
Workvan	Commercial	<35	Day	Y
Passenger	Disabled	> = 35	Night	N
Passenger	Commercial	<35	Night	Y
Truck	Disabled	<35	Night	Y
Passenger	Disabled	<35	Day	Y
Workvan	Disabled	> = 35	Day	Y
Workvan	Regular	<35	Night	Y
Truck	Disabled	> = 35	Day	N

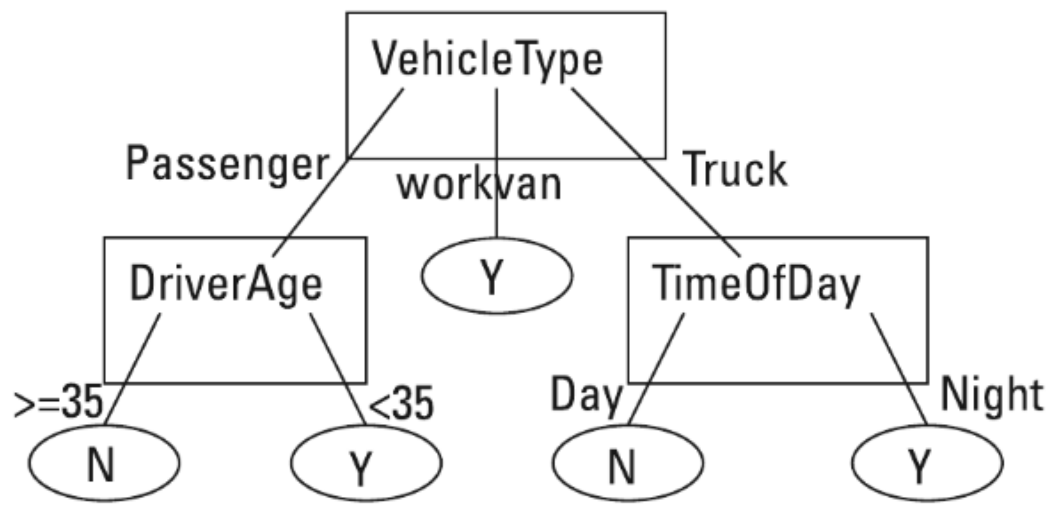
M. Grossi: Homeland Security Technology Challenges: From Sensing and Encrypting to Mining and Modeling, 2008

In case of a wrong decision the engineers might be held liable. With the model they can prove that the system was implemented correctly.

historic data (training instances) **class**

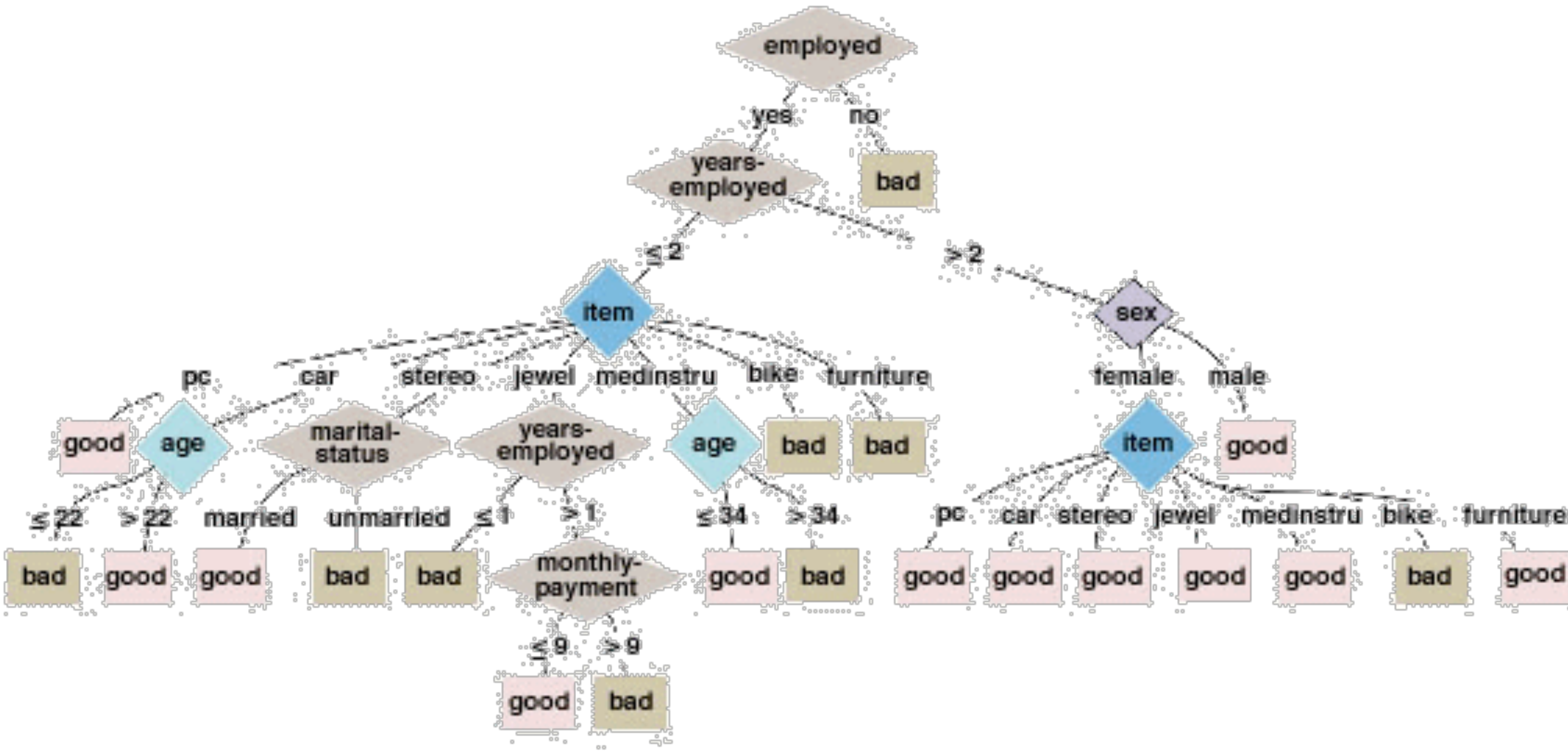
Vehicle Type	Driver Lic Type	Driver Age	Time of Day	High Risk
Passenger	Regular	> = 35	Night	N
Passenger	Regular	> = 35	Day	N
Workvan	Regular	> = 35	Night	Y
Truck	Disabled	> = 35	Night	Y
Truck	Commercial			
Truck	Commercial			
Workvan	Commercial			
Passenger	Disabled			
Passenger	Commercial			
Truck	Disabled			
Passenger	Disabled			
Workvan	Disabled			
Workvan	Regular			
Truck	Disabled			

Learned Decision Tree



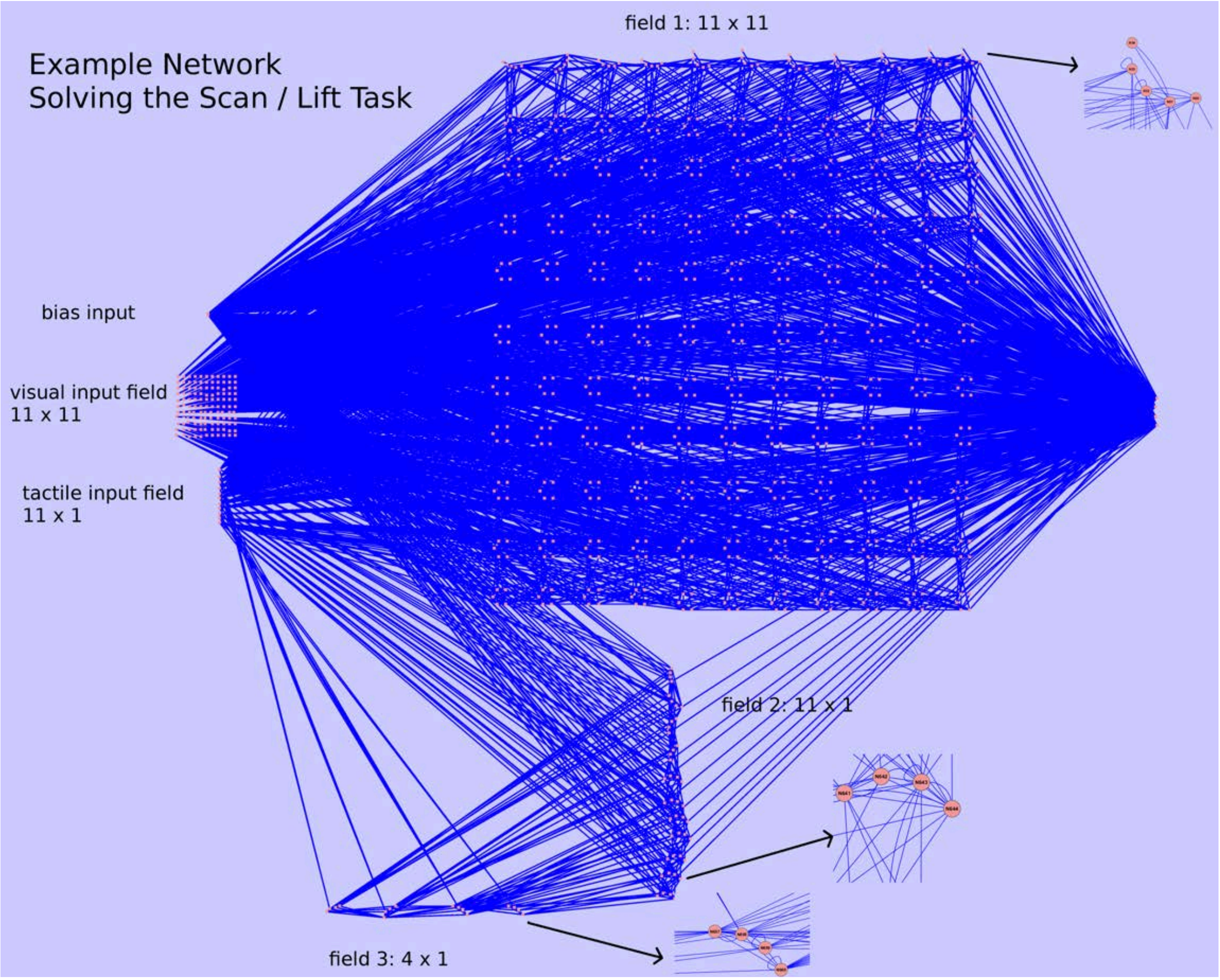
M. Grossi: Homeland Security Technology Challenges: From Sensing and Encrypting to Mining and Modeling, 2008

However, for more complex decision trees the validity of the model may be difficult to assess, because it is not explicitly obvious.



http://access.ncsa.illinois.edu/Stories/97Stories/KUFRINfinalgraph.gif

And with large neural networks it may become impossible to deduce the knowledge and reasoning by looking at the model.



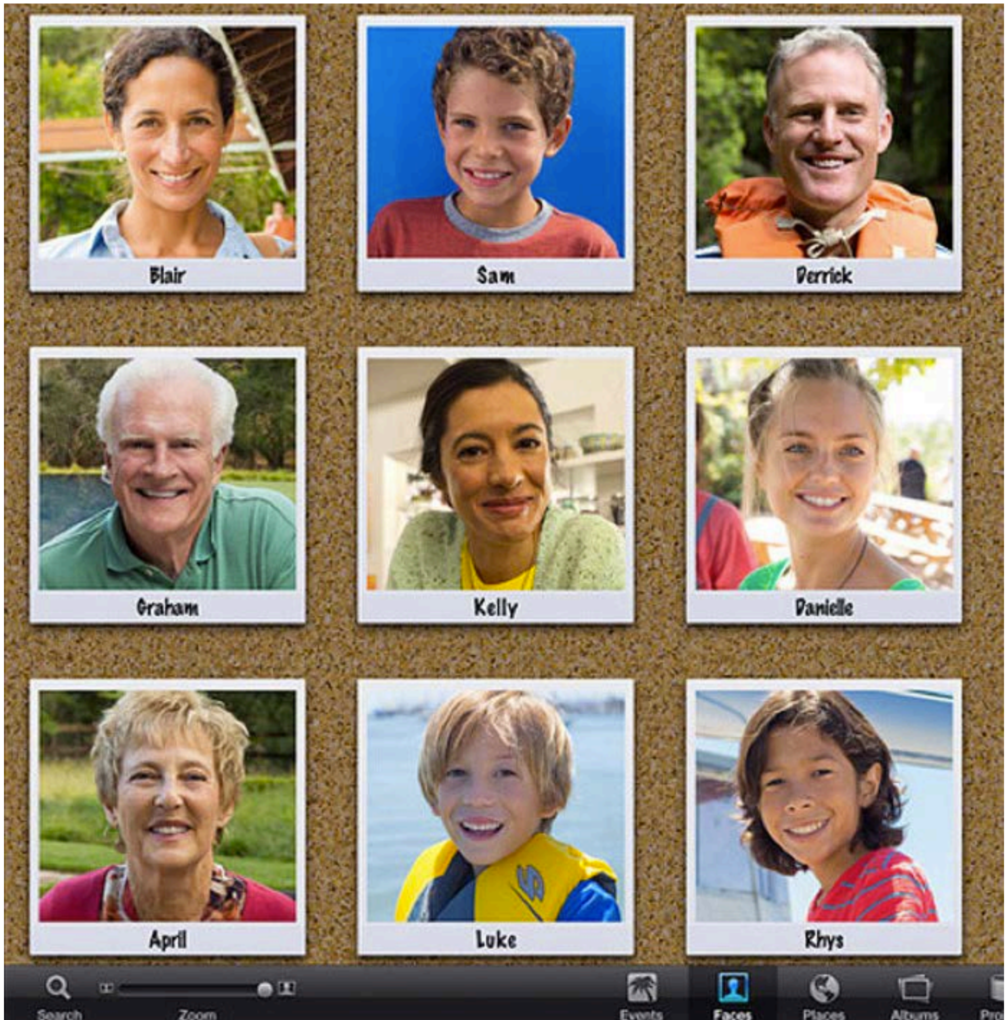
http://ni.www.techfak.uni-bielefeld.de/files/MSI1edited.png

In contrast, conventional software is programmed very explicitly.
The behavior can be deduced by looking at the code.

```
import csv

def process_row(row):
    first = row[0].strip()
    last = row[1].strip()
    address = row[2].strip

    try:
        (first, middle) = first.split(' ')
        (address, middle) = first.split(' ')
    except ValueError:
        pass

    #print first, last
    return (first, last, address)

def process_file(f):
    voters = []
    for row in f:
        try:
            if (len(row) > 0):
                (first, last, address) = process_row(row)
                first = first.upper()
                last = last.upper()
                voters.append((first, last, address))
        except IndexError:
            print "Error in row"
            print "[", row, "]"
            raise
    print voters
    return voters
```


http://www.netnewscheck.com/article/22622/digital-tools-can-enhance-news-storytelling

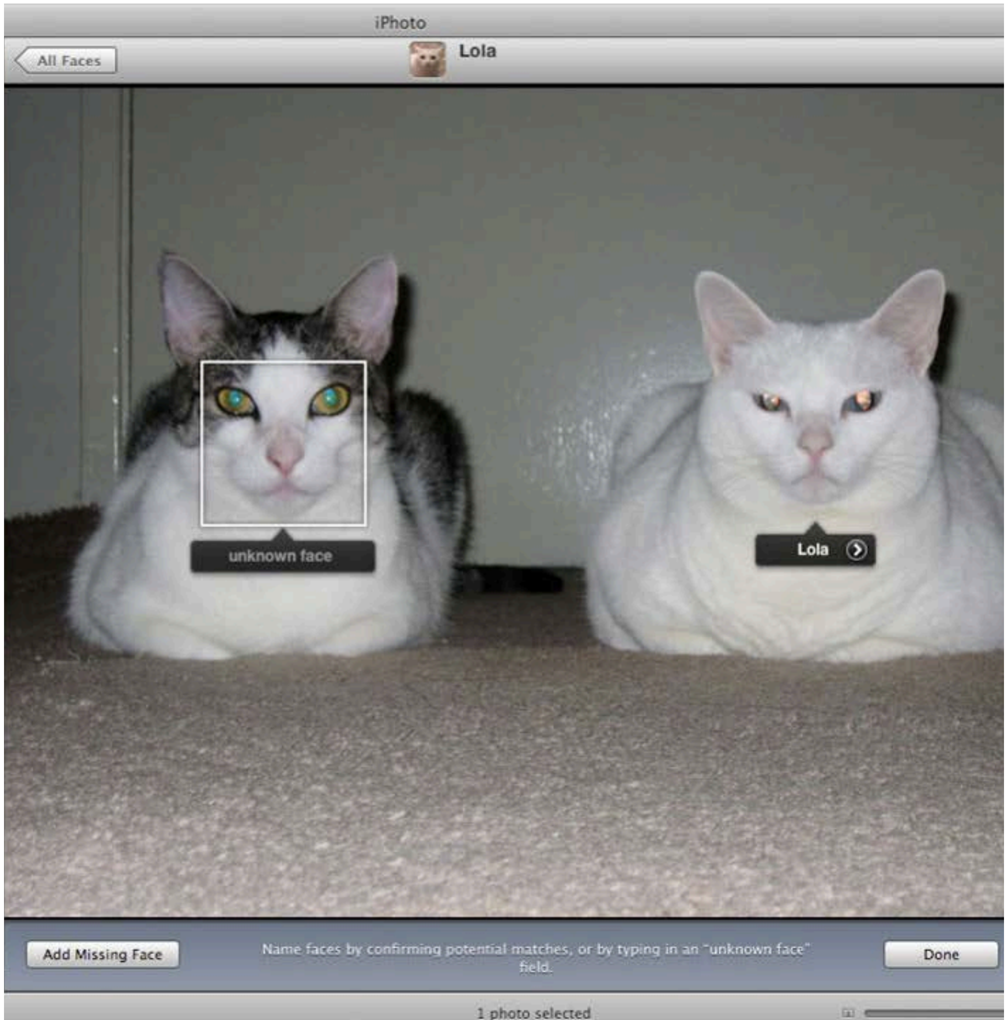
An example for a machine learning technique that can be used in creative ways is “face recognition”.



http://www.geeksugar.com/Use-Faces-iPhoto-26342264

- **Apple iPhoto** allows to automatically tag people by facial recognition

An example for a machine learning technique that can be used in creative ways is “face recognition”.



http://www.maclife.com/article/news/iphotos_faces_recognizes_cats

- **Apple iPhoto** allows to automatically tag people by facial recognition
- not restricted to humans

Finally, due to their **mobility**, agents have to tolerate faults and prevent attacks, even when there is no technically savvy operator present.

SAFETY

Faulty components can lead to system malfunctions of sensing and acting components, possibly causing harm.

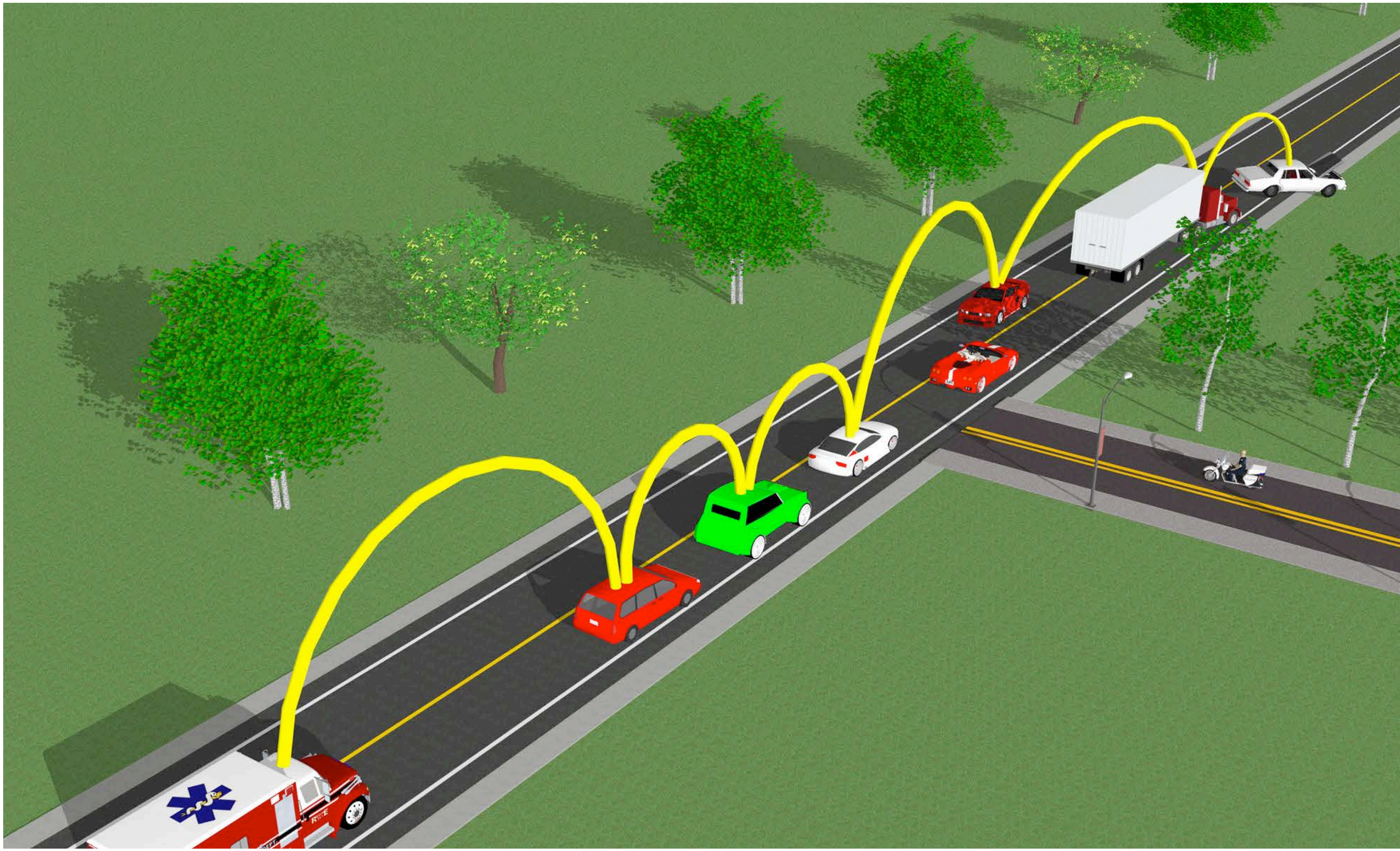
- **redundancy**
  use multiple components to achieve a task
- **diversity**
  use different components to achieve a task

SECURITY

Agent design has to take into account the possibility of malicious participants (third parties as well as the owner!)

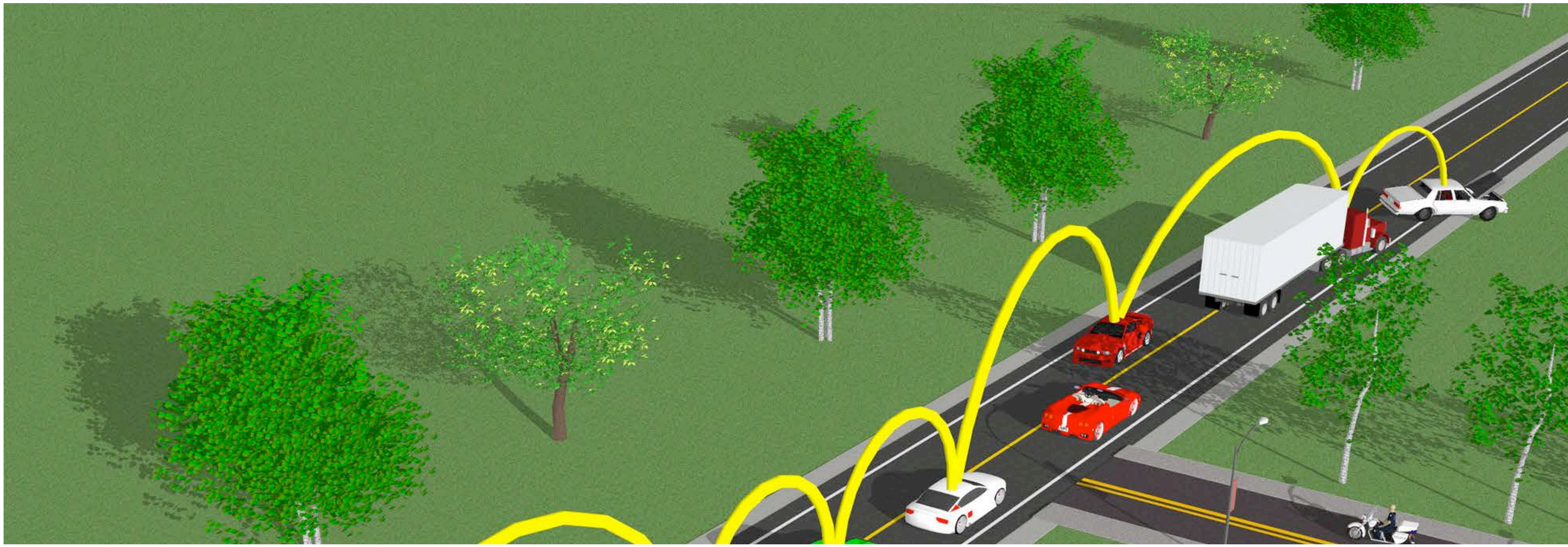
- **integrity and authenticity** to prevent forging of messages
- **confidentiality** to prevent information disclosure
- **veracity** to validate plausibility of sensor readings

Cars are going to communicate with each other over WiFi networks
Perception of agents is not limited to direct line of sight any more.



http://www.dlr.de/Portaldata/1/Resources/verkehr/car2car1.jpg

Cars are going to communicate with each other over WiFi networks
Perception of agents is not limited to direct line of sight any more.



Risks of malicious interference

- **injection** of faked warnings
- **suppression** of warnings
- remote control and shutdown

Liability issue: **plausible deniability**

Protective measures:

- integrity
- authenticity
- veracity

for detection of faked messages

http://www.dlr.de/Portaldata/1/Resources/verkehr/car2car1.jpg

Open Questions

How can engineers ensure controllability?

Conventional software quality control techniques are not sufficient. Safeguards and restrictions are an important feature of intelligent agents.

CONVENTIONAL SOFTWARE

Software engineering process ensures high quality:

- defining requirements
- specification
- software design
- coding
- test

Supervision by operators and users.

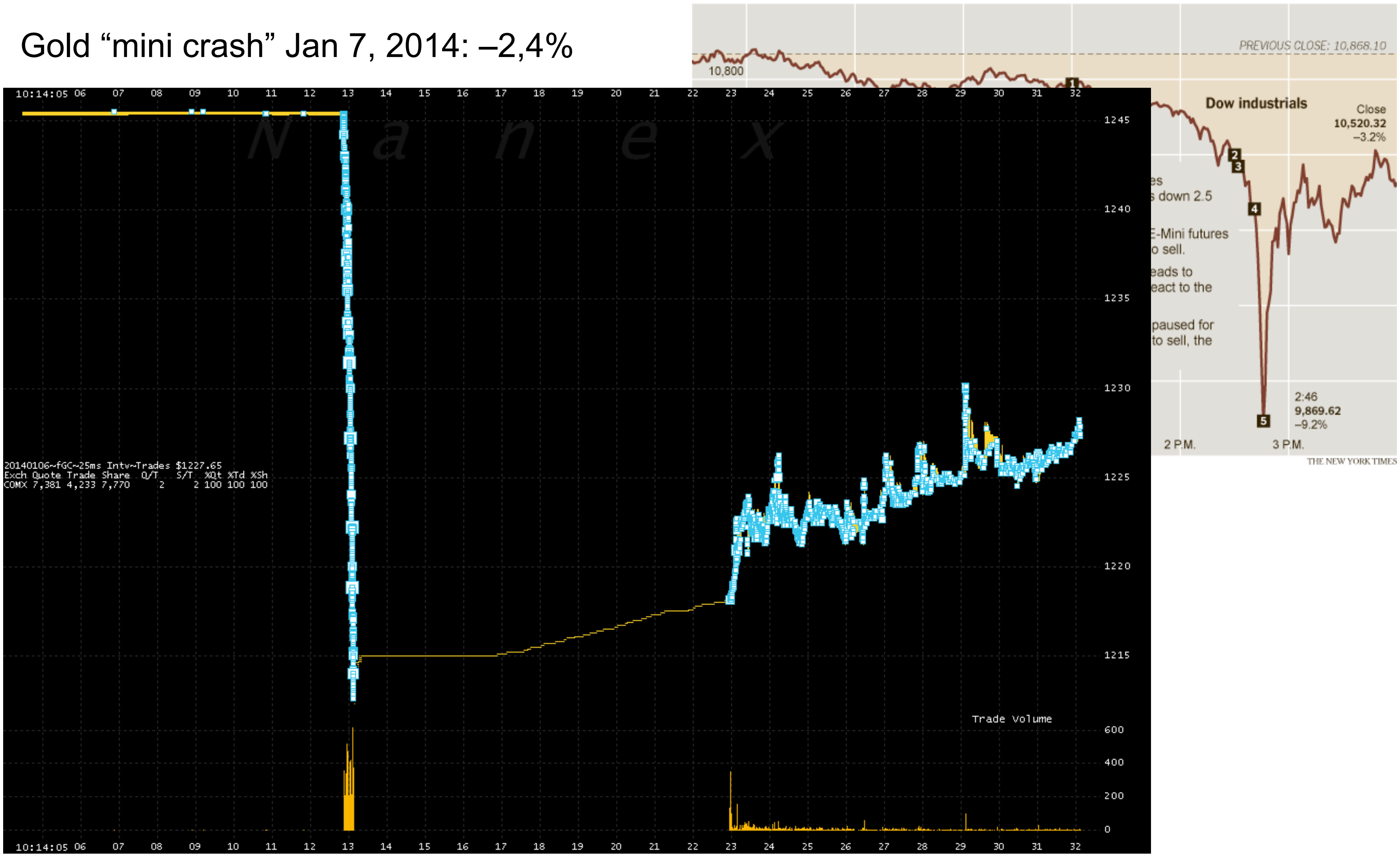
AGENTS

Monitoring component that aggregates multiple readings of different sensors and evaluates the actions of the agent (research field: moral/ethical computing).

Definite whitelist of explicitly allowed behavior, everything else initiates a safe shutdown of the system.

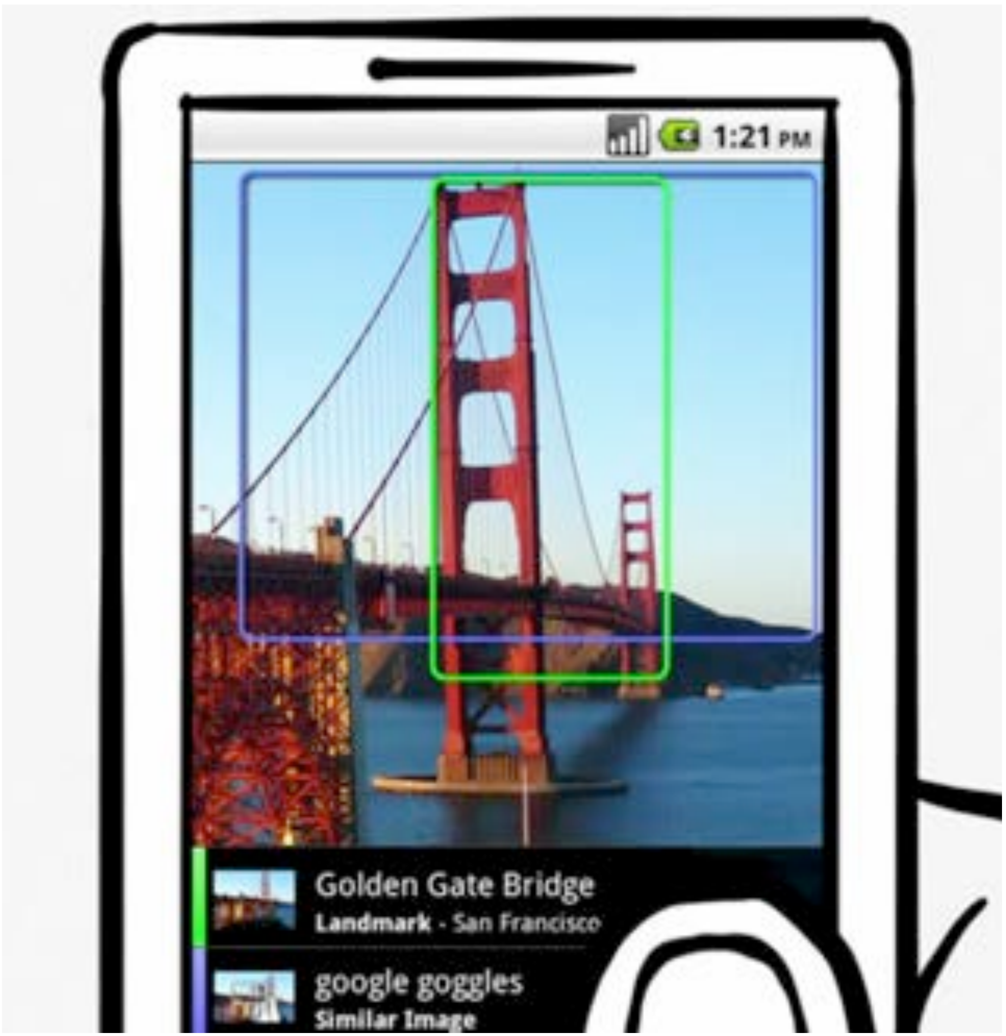
Stock markets have introduced automated circuit breakers to suspend trading when volatility rises sharply.

Gold “mini crash” Jan 7, 2014: –2,4%



http://www.ritholtz.com/blog/wp-content/uploads/2010/10/flash-crash-dow-popup.png and http://www.zerohedge.com/news/2014-01-06/gold-flash-crashes-halts-trading

Google has started to integrate safeguards into their applications.



http://www.hongkiat.com/blog/google-labs-experiments/

Google Goggles app can search for objects by taking a picture of them

- intentionally prevents submission of faces

Google Suggest uses a blacklist of words, which will not be suggested.

- however, blacklist approach needs human supervision

Topics addressed in this talk and conclusions

What are intelligent agents?
autonomous, mobile, adaptive sensing/acting, indeterministic

How are they different from conventional software?
behavior not solely based on initial program, but evolving

What techniques are used?
supervised learning: decision trees, art. neural networks, …

Why is it difficult to foresee their actions?
engineer is not a programmer any more, but a creator of an "software organism"; control is lost once agent is deployed

How can engineers ensure controllability?
"keep the human in the loop", monitoring and whitelists, safety and security measures

Dominik Herrmann (herrmann@informatik.uni-hamburg.de)
Slides: http://dhgo.to/agents

A. Matthias: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6: 175–183, 2004.