Should We Fear Intelligent Machines?

Gerald Jay Sussman

From the BBC

Prof. Stephen Hawking, one of Britain's pre-eminent scientists, has said that efforts to create thinking machines pose a threat to our very existence.

"The development of full artificial intelligence could spell the end of the human race."

"Humans, who are limited by slow biological evolution, couldn't compete and would be superseded."

From The Guardian

Elon Musk has spoken out against artificial intelligence (AI), declaring it the most serious threat to the survival of the human race:
"I think we should be very careful about artificial intelligence. If I had to guess at what our biggest existential threat is, it's probably that. So we need to be very careful."
"I'm increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish. . . . With artificial intelligence we are summoning the demon."
GJS: At least he put his money where his mouth is, donating $10Million to the Future of Life Institute, run by MIT physicist Max Tegmark.

Vision?

Problems in Heaven?

Nguyen A, Yosinski J, Clune J. “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.” In *Computer Vision and Pattern Recognition (CVPR ’15)*, IEEE, 2015.

> “... Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). In this paper, we show another way that DNN and human vision differ: It is easy to produce images that are completely unrecognizable to humans ..., but that state-of-the-art DNNs believe to be recognizable objects with over 99% confidence (e.g. labeling with certainty that TV static is a motorcycle). ...”

Memorizing not Understanding

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals “Understanding deep learning requires rethinking generalization” (2017)

Deep neural networks easily fit random labels.

“... when trained on a completely random labeling of the true data, neural networks achieve 0 training error. ...”

“The effective capacity of neural networks is sufficient for memorizing the entire data set.”

So perhaps we are safe?

I Disagree

- Artificial Intelligence is a threat.
- Artificial Intelligence is not an *existential threat*.
- Synthetic Biology *is* an existential threat.
  - It is a present danger—NOW!
  - It does not require an industrial base.
  - Can make a nasty organism in a kitchen.
- Human behavior is an existential threat.
  - Local optimization
  - Overpopulation
  - Resource depletion
  - Ecological destruction
  - Aggressive behavior

Artificial Intelligence is a threat

- It is a threat
  - to our privacy
  - to our civil liberties
  - to our political/social/economic systems
- The biggest AI threat—full automation
  - productivity/person becomes infinite
  - cost of goods and services goes to zero
  - *The End Of Work*
  - economic, political, social upheaval
  - Could be liberating!
  - Could be a disaster: wars and tyrannies

Mitigation

- Let's ignore the full automation problem for now.
- How can we ameliorate these other threats?
  - Sociological
  - Technological
- To socialize autonomous intelligent agents
  - Discourage harmful behavior
  - Encourage beneficial behavior
  - Create Taboos—Asimov's laws
  - Enforce Taboos

"Artificial" is Irrelevant!

- Humans are dangerous!
- Intelligent machines are dangerous!
- The Legal System is one way to enforce taboos
  - Investigates reports of bad behavior
  - Audits post hoc
  - Takes action to prevent recurrence
- Investigation needs audit trails
  - Explanations and stories
  - Witnesses
- Perpetrator must be either
  - corrigible
  - capable of being disabled or confined

Audit Trails

Autonomous agents that can make decisions or take actions that could affect the welfare of others must be

- able to be audited.
- able to tell an understandable coherent story justifying its actions.
- able to be challenged in an adversary proceeding.
- If the explanation is inadequate or inappropriate the agent should be either
  - corrected
  - disabled

Suppes's Rules and Audit Trails

Introduction of a Premise

$$\frac{\cdots \quad \cdots \quad \cdots \qquad \cdots}{n \quad A \quad \text{Premise} \quad \{n\}}$$

Modus Ponens

$$\frac{\begin{array}{llll} n & A \Rightarrow B & \cdots & d_1 \\ m & A & \cdots & d_2 \end{array}}{o \quad B \qquad (\text{MP } n\ m) \quad d_1 \cup d_2}$$

More Suppes's Rules

Conditional Proof

n	A	Premise	$\{n\}$
m	B	$\cdots$	d
o	$A \Rightarrow B$	(CP n m)	$d - \{n\}$

Reducto Ad Absurdum

n	A	Premise	$\{n\}$
m	B	$\cdots$	d_1
o	$\neg B$	$\cdots$	d_2
p	$\neg A$	(RAA n m o)	$d_1 \cup d_2 - \{n\}$

A Suppes Derivation

1	$\forall y(\text{greek}(y) \Rightarrow \text{human}(y))$	Premise	{1}
2	$\forall x(\text{human}(x) \Rightarrow \text{mortal}(x))$	Premise	{2}
3	$\text{greek}(\star G)$	Premise	{3}
4	$\text{greek}(\star G) \Rightarrow \text{human}(\star G)$	(US 1 $\star G$ y)	{1}
5	$\text{human}(\star G)$	(MP 4 3)	{1 3}
6	$\text{human}(\star G) \Rightarrow \text{mortal}(\star G)$	(US 2 $\star G$ x)	{2}
7	$\text{mortal}(\star G)$	(MP 6 5)	{1 2 3}
8	$\text{greek}(\star G) \Rightarrow \text{mortal}(\star G)$	(CP 7 3)	{1 2}
9	$\forall z(\text{greek}(z) \Rightarrow \text{mortal}(z))$	(UG 8 z *G)	{1 2}

A Caveat

If you think that your paper is vacuous,
use the first-order functional calculus.
It then becomes logic,
and as if by magic,
the obvious is hailed as miraculous.

Paul Halmos c.1967

The purpose of this exercise is not to advocate logic.
It is to show how one can keep track of reasons and dependencies in a system of computational rules.

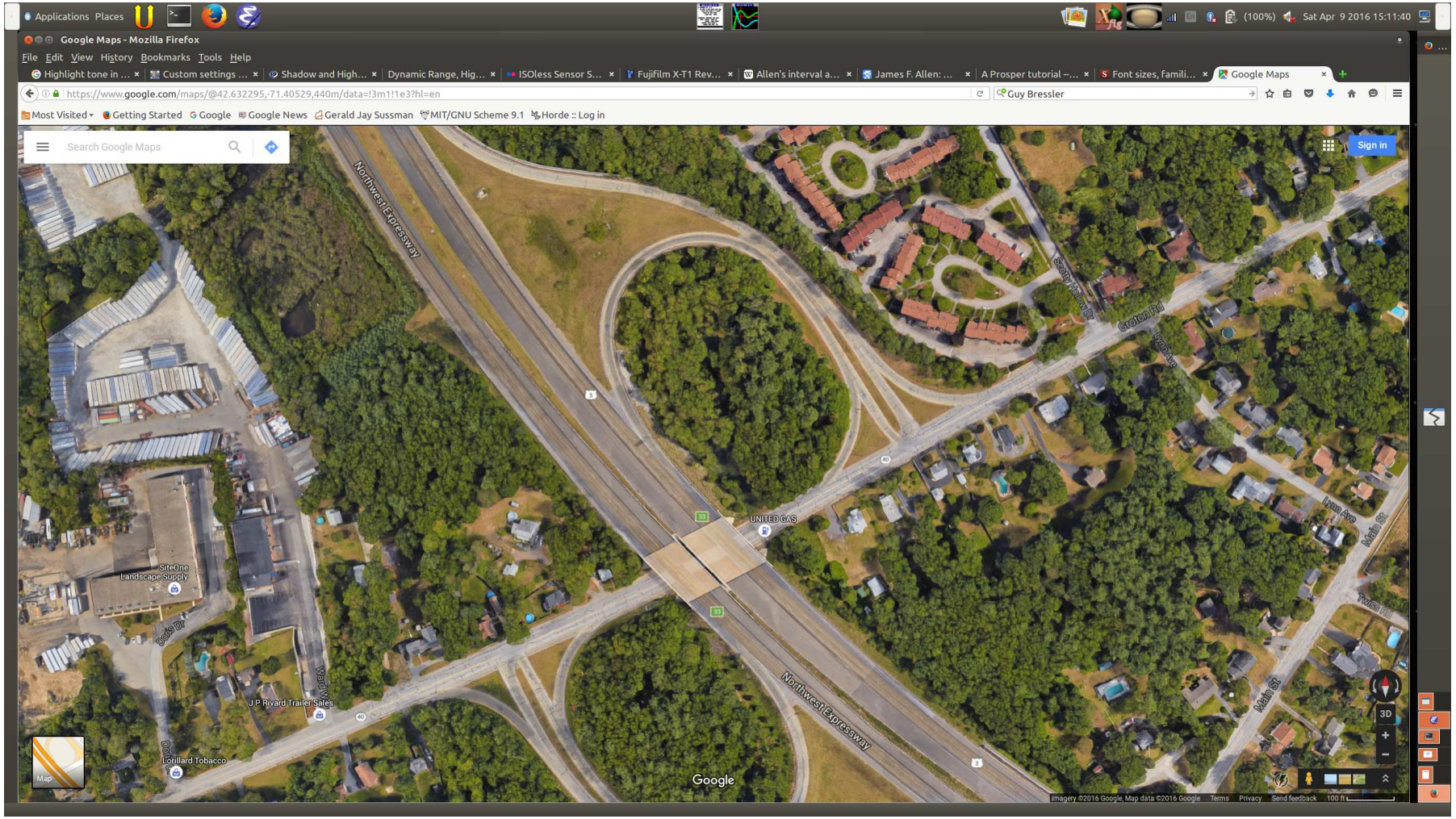
IBM Watson appears to keep track of provenance.

An Accident–the story

- Going north on Route 3.
- Takes exit 33 to get onto Route 40 West.
  - Going too fast for the curve.
- Operator applies brakes too hard.
  - Rear wheels lose traction.
  - Rear wheels spin out clockwise: oversteering skid.
- Car tries to compensate with CPR strategy:
  - ALB: Reduce braking.
  - Steering: Small adjustment to left.
  - Transmission: Neutral.
  - Steering: Return to center.
  - Transmission: Restore power.

Partial support from Toyota Research Institute

An Accident–the venue



Partial support from Toyota Research Institute

An Accident—Ben Yuan's road model

Partial support from Toyota Research Institute

A “CAN” log fragment

```
   ...
56.095 GroundTruthXYZ 4.34 5.05 44.99
56.1 22 0.30
56.1 23 1.34
56.1 25 1.07
56.1 30 0.00
56.1 B1 9797.73 9845.80
56.1 B3 9846.59 9894.56
56.1 120 13 04 50
56.1 244 0.29
56.1 GroundTruthXYZ 4.36 5.05 45.12
56.105 22 0.31
56.105 23 1.35
56.105 25 1.06
56.105 30 0.00
56.105 B1 9799.55 9848.86
56.105 B3 9848.39 9897.63
56.105 120 13 04 50
56.105 244 0.29
   ...
```

```
22: sideways acceleration:
        m/s^2, + => right
23: forward acceleration:
        m/s^2, + => forward
25: steering angle:
        degrees, + => right
30: brakes: float 0-1
B1: front wheel speeds:
     right left | kph * 100
B3: rear wheel speeds:
     right left | kph * 100
120: drive mode:
     shiftState1 04 shiftState2
     (PRND state)
244: accelerator input:
     float 0-1
groundTruthXYZ: world location
   meters relative to origin:
     X east, Y up, Z north
```

Partial support from Toyota Research Institute

Code fragment—Leilani Gilpin

```
;;; params are rates for beginning and end of an interval
;;; wheel-rates = (right-wheel-rate left-wheel-rate)
;;; params = (rates-at-beginning rates-at-end)

(define (skid-in-interval? front-params back-params)
  (let* ((front-start (car front-params))
         (front-end (cadr front-params))
         (back-start (car back-params))
         (back-end (cadr back-params)))
    (or (skidding? front-start back-start)
        (skidding? front-end back-end))))

(define (skidding? front-wheel back-wheel)
  (let ((wheel-rate-differences
         (map - front-wheel back-wheel)))
    (> (abs (apply max wheel-rate-differences))
       skid-threshold)))
```

Partial support from Toyota Research Institute

An Accident–*What* happened.

- 18:10:25.333 GPS: Heading 321.16, Speed 60.3mph
- 18:10:26.500 Operator: Brake 0.35, Steer 5.0
- 18:10:26.560 Driver assist: Brake 0.45 :-)!
- 18:10:27.867 GPS: Heading 353.84, Speed 52.1 mph
- 18:10:29.970 Operator: Brake 0.90, Steer 9.3
- 18:10:30.010 Wheel Rate Monitor: Skid
- 18:10:30.040 GPS: Heading 28.27, Speed 0.0mph
- 18:10:30.070 Wheel Rate Monitor: Skid
- 18:10:30.170 Operator: Brake 0.91, Steer 6.6
- 18:10:32.933 GPS: Heading 129.08, Speed 0.2mph
- 18:10:35.140 Operator: Brake 0.93, Steer 0.0
- 18:10:35.467 GPS: Heading 121.52, Speed 0.0mph
- 18:10:38.670 Stopped

Partial support from Toyota Research Institute

An Accident—Deeper Analysis

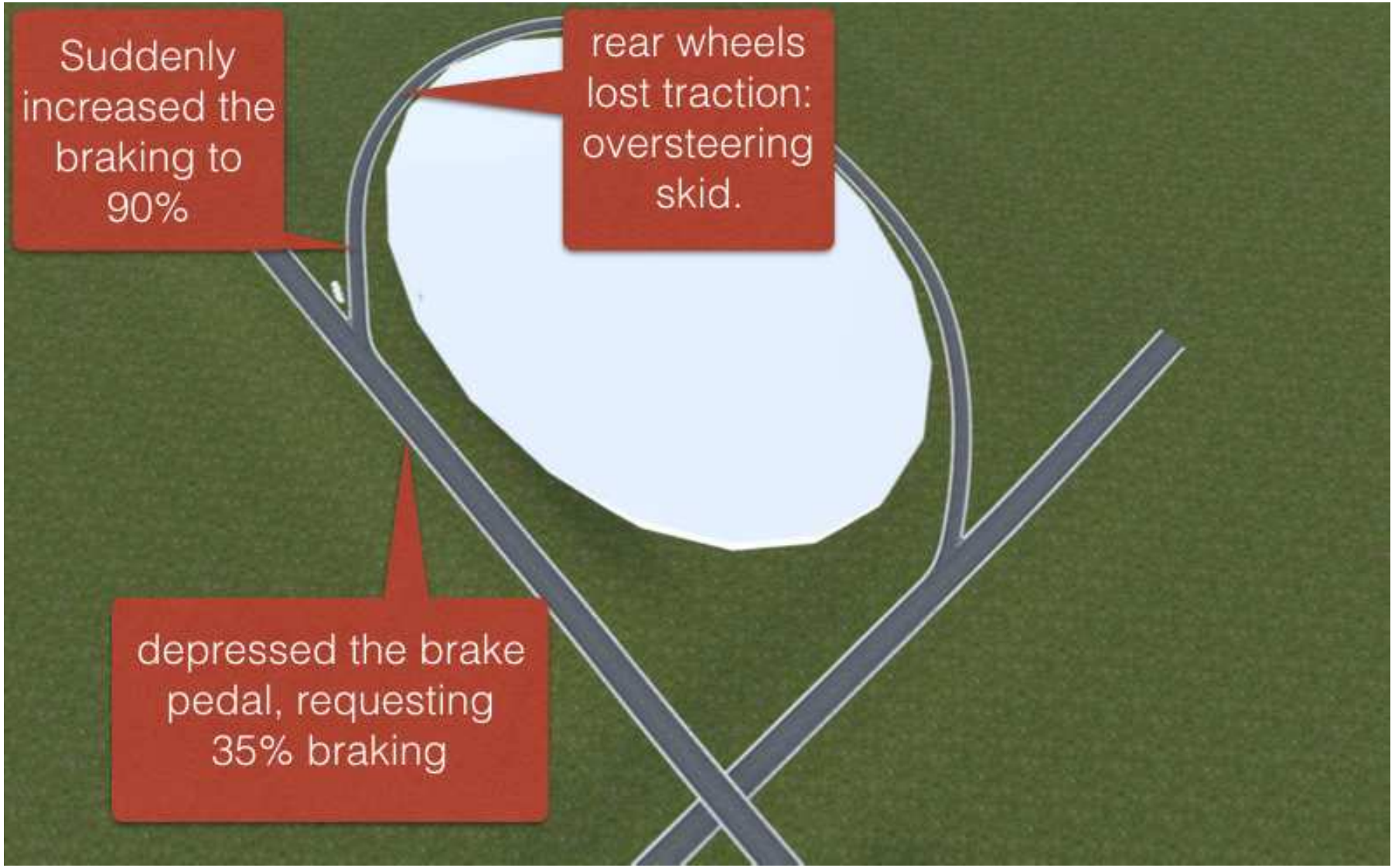
However, we should be able to get more detail:

```
==> (Expand Skid report 18:10:30.01)
```

- `18:10:30.010 Front Wheel Rates (145.503 124.757)`
- `18:10:30.010 Rear Wheel Rates (1246.291 1467.201)`
- `Wheel rates:  speeds inconsistent`
- `18:10:30.010 Lateral acceleration 12.221`
- `Rear wheels sliding clockwise`

Partial support from Toyota Research Institute

An Accident–understood



Partial support from Toyota Research Institute

The Power of Provenance

John Ionnidis: “Why Most Published Research Findings Are False,” in *PLoS Med.* 2005 Aug; 2(8): e124.

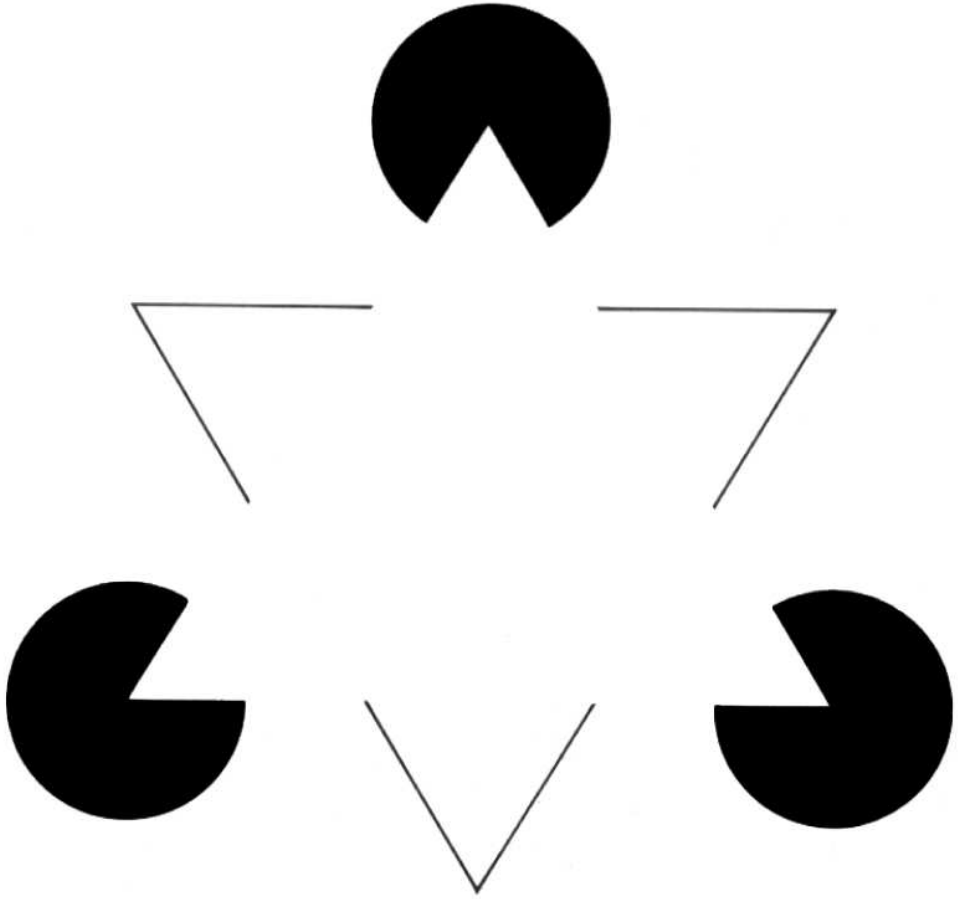
We can ameliorate this problem with provenance tracking

- All research papers refer to inputs
- Any result that uncovers an obvious inconsistency can mark all inputs (recursively).
- Any result that accumulates many marks is probably wrong and polluting the literature.
- Results that do not depend on heavily marked results are likely to be correct.

Cognition fills in details

Torralba, IJCV, 2003

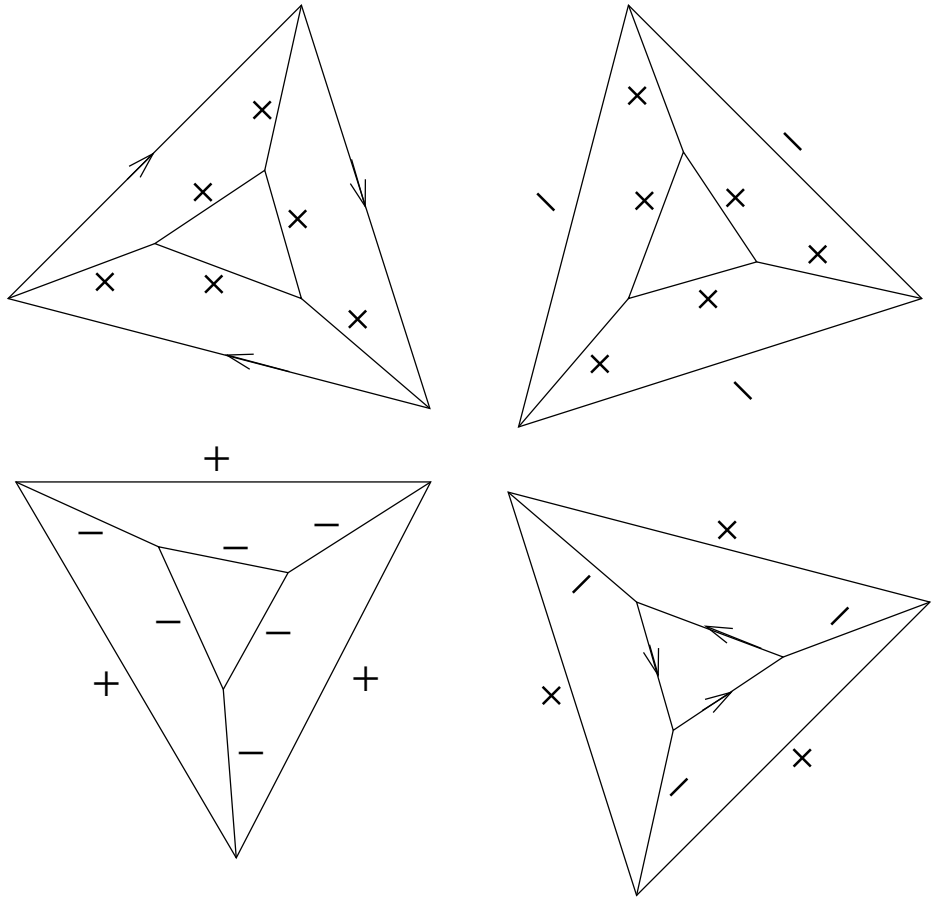
Kanizsa's Triangle Illusion



Gaetano Kanizsa (1955)

Cognition fills in details

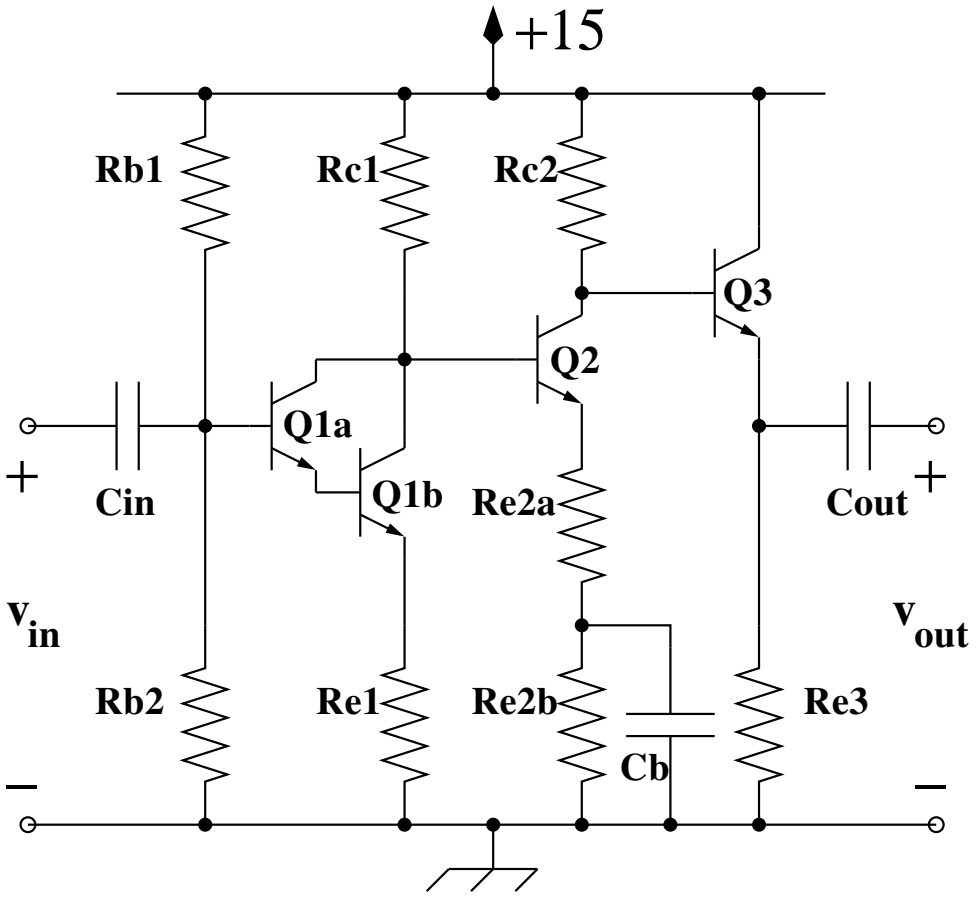
Details appear from all directions.



David Waltz (1972)

Circuit Analysis is filling in details

Details appear from all directions.



Stallman&Sussman (1975)

Details appear from all directions

- Mental imagery is the visual system running backwards
  - Mental Imagery is controlled hallucination.
- The elephant test
  - *You are at home in your bedroom. An elephant appears just outside the bedroom door. Can the elephant come into the bedroom with you?*
- You know what happens because you "see" it.
- Cognition makes good use of hallucinations!

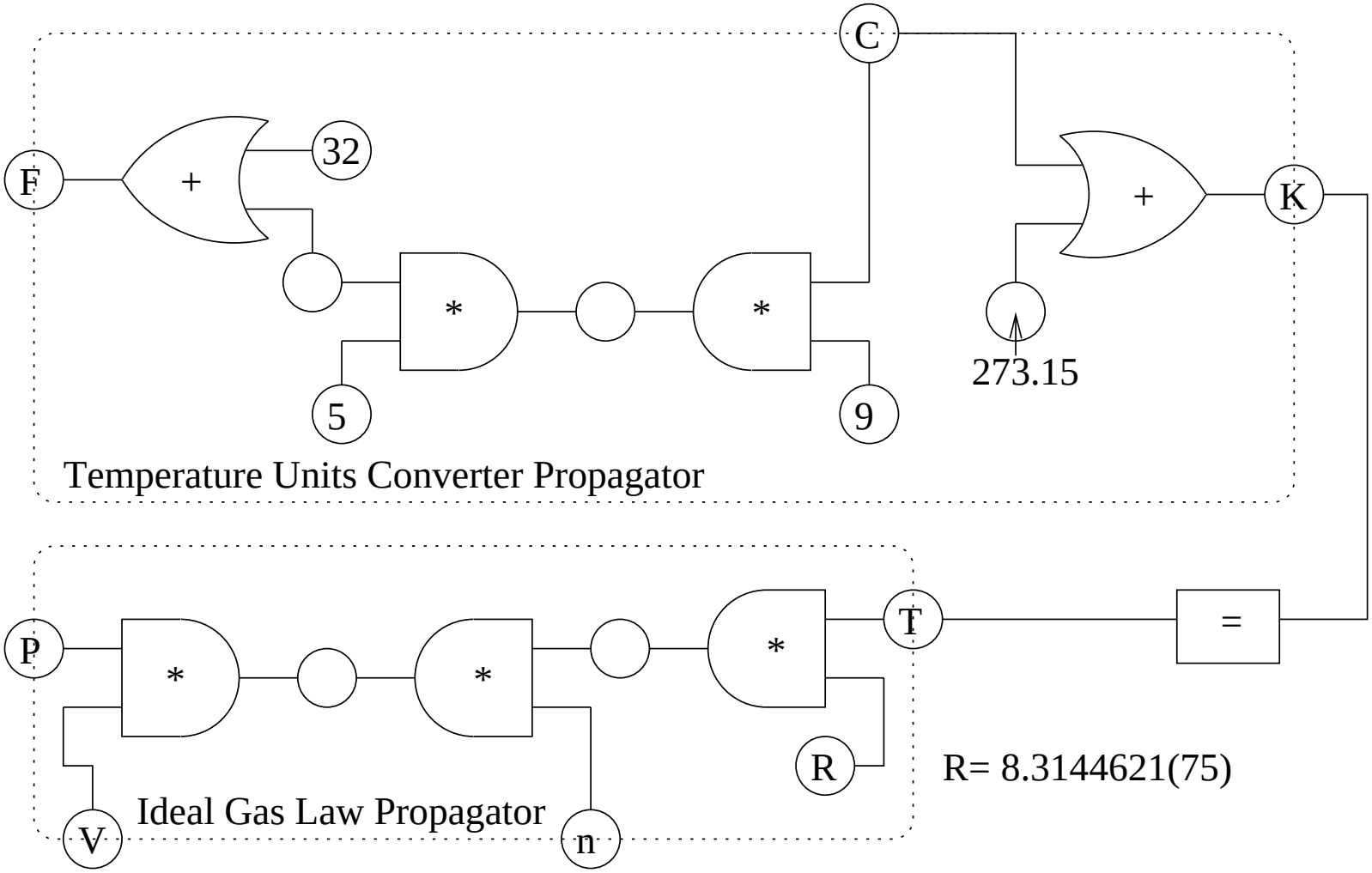
The Propagation Mechanism

a model for distributed, concurrent computation

- computational elements are “autonomous machines”
- interconnected by shared “cells”
- each propagator continuously examines its neighbor cells adding information to some, based on deductions it can make from the information in others

- scalable from multicore and multiprocessor through cluster and grid computing, to distributed computing over a network
- applicable to all levels, from hardware architecture to enterprise application software.

A Simple Example



Radul&Sussman (2009, 2010)

- a cell does not contain a value:
  a cell contains information about a value
- a cell generically merges new information with existing information to produce the most informative description
- the information in a cell is monotonically increasing
- many partial-information structures are available
  - numerical intervals merge by intersection
  - patterns merge by unification
  - algebraic expressions merge by equation-solving
  - ? probability distributions ?

A Constraint Network for Finance

```
(make-financial-entity 'Alyssa)
(make-financial-entity 'Ben)

;;; Ben and Alyssa are married
(make-financial-entity 'Ben-Alyssa)
(combine-financial-entities 'Ben-Alyssa 'Ben 'Alyssa)

;;; Ben and Alyssa file income tax jointly
(tell! (gross-income 'Ben-Alyssa) 427000 'IRS)

;;; Ben works at Gaggle as a software engineer.
(breakdown (gross-income 'Ben) 'Gaggle-salary 'investments)

;;; He gets paid a lot to make good apps.
(tell! (thing-of '(Gaggle-salary gross-income Ben))
       200000 'Gaggle)
```

Tracking Provenance

```
;;; Alyssa works as a PhD biochemist in big pharma.
(breakdown (gross-income 'Alyssa) 'GeneScam-salary 'investments)

;;; Biochemists are paid poorly.
(tell! (thing-of '(GeneScam-salary gross-income Alyssa))
       70000
       'GeneScam)

(tell! (thing-of '(investments gross-income Alyssa))
       (make-interval 30000 40000)
       'Alyssa)

(inquire (thing-of '(investments gross-income Ben)))
;Value: #(supported #[interval 117000 127000]
                    (gaggle genescam alyssa irs))
```

Tracking Provenance

```
;;; Ben is a tightwad
(tell! (thing-of '(expenses Ben))
       (make-interval 10000 20000)
       'Ben)

(inquire (thing-of '(net-income Ben)))
;Value: #(supported #[interval 297000 317000]
                    (ben genescam alyssa irs))

;;; But Alyssa is not cheap.  She likes luxury.
(tell! (thing-of '(expenses Alyssa))
       (make-interval 200000 215000)
       'Alyssa)

(inquire (thing-of '(net-income Alyssa)))
;Value: #(supported #[interval -115000 -90000]
                    (alyssa genescam))
```

Tracking Provenance

```
;;; But they are doing OK anyway!
(inquire (thing-of '(net-income Ben-Alyssa)))
;Value: #(supported #[interval 192000 217000]
                    (ben alyssa irs))
```

Notice that this conclusion does not depend on the details, such as information from Gaggle or GeneScam!

Future: Vehicle Accident Investigation

- Investigator: So what happened?
- Car: As we approached the curve everything looked OK. The operator set the steering to an appropriate trajectory for the radius of curvature of the turn, as indicated by the GPS, and the operator depressed the brake pedal, requesting 35% braking. I put on 45% front braking. As we proceeded along the curve the operator suddenly increased the braking to 90%. My rear wheels lost traction and I went into an ovesteering skid. I entered the Correct-Pause-Recover routine: I reduced the braking, I locked the steering left, I neutralized the front wheel drive. I then brought the steering back to neutral and restored power. This was insufficient to correct the trajectory in time.

Partial support from Toyota Research Institute

Investigation continues

- Investigator: Why did you choose 45% front braking at the entry to the turn?
- Car: I used the radius of curvature from the GPS map and I computed, from design formula #73602, that the operator's 35% front braking request would be inadequate to decrease my speed sufficiently to safely negotiate the turn. The operator's request was insufficient.
- Investigator: Why did the skid occur?
- Car: The sudden application of 90% braking decreased the load on the rear wheels. The loss of normal force decreased the frictional force on the rear wheels' contact patches. This allowed the centrifugal force of the turn to cause slip, and finally slide.

Partial support from Toyota Research Institute

Investigation continues

- Investigator: Is there a way the accident could be avoided?
- Car: I might not have accepted the 90% braking command, since I knew that it would exceed my design parameters. But I did not know that this command was not a response to some other hazard.

Indeed, is the car partly at fault? Or is it entirely the driver's fault for panicking and jamming on the brakes? This is for someone else to decide.

But, at least we get a story that reasonable people can understand and deliberate about.

Partial support from Toyota Research Institute

A Research Problem

"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know."—Donald Rumsfeld

One really hard problem, which is a matter of current research, is to be able to explain what is missing: If we cannot obtain some desired result, what plausible facts or rules could be provided that would make it possible to obtain the result.

You should be working on this.

Rumsfeldian Reasoning

M.D. Why can I not get access to Mr.Sick's records?

Files You are not authorized to access Mr.Sicks records.

M.D. What do I need to be authorized?

Files At least one of

- You are the provider of record for Mr.Sick.
- There is a written release on file for you to access those records.
- There is an emergency and you are the duty officer.

M.D. I am the duty officer, and Mr.Sick is having a seizure.

Files I did not know that a seizure is an emergency.

M.D. Add a rule that a seizure is an emergency, and give me the records.

Files Yes sir!

- I note and recorded that you gave this order.
- I added your rule to my rules.
- Here are Mr.Sick's files.

Fear?

- So ignoring the major threat of social, political, and economic disruption from the “End Of Work,” the residual threats of AI are to privacy and civil liberties.
- These threats can be ameliorated by requiring that autonomous agents that can affect our lives be able to explain their actions and decisions as a symbolic story in an adversary proceeding. Perhaps this should be a legal requirement for deployment.
- There are technologies in place and in development that can make this possible.
- We can have confidence in such systems only if the software that controls them is “free software” that can be read, understood, modified, and distributed by the users who employ them.

My Real Worry

- Malevolent Human/Government/Corporate Creators
- Proprietary Claims
  - Enforced by laws – e.g. the DMCA
  - Enforced by Technological Means
    - DRM (Digital “Rights” Management)
    - “Trusted Computing” (Treacherous Computing)
  - Enforced by technical obfuscation
- Code that behaves badly cannot
  - be examined to understand bad behavior
  - be modified to fix bad behavior
  - be patched and distributed to users

  without permission of the possibly malevolent creators

We Already Suffer

- Wide distribution of malware and spyware
- Trapdoors in systems
- Trapdoors in firmware
  - Kaspersky Lab reported: “Drives made by Seagate Technology, Western Digital Technologies, Hitachi, Samsung Electronics and Toshiba can be modified by two of Equation’s hard disk drive malware platforms, ‘Equationdrug’ and ‘Grayfish.’ ”
- Nasty hardware (e.g. Trusted Platform Module)

For a big list of real nastys see
https://www.gnu.org/philosophy/proprietary-back-doors.html

Fear Intelligent Machines?

- Not if they can
  - explain their actions and decisions to others (human and machine)
  - be a party to an adversarial proceeding
  - be corrected or disabled by their users
- Software mechanisms can capture reasons and dependencies
  - Suppes-like rules
  - Propagation infrastructure

Proprietary Software is Bad

- We must be able to
  - read the software
  - test the software
  - correct the software
  - distribute the patches
- To be trusted, autonomous agents must be
  - open to criticism
  - open to testing and argument
  - easily modified by other than the creators

Trusted autonomous agents must be built with

"free software"