Asilomar Study on Long-Term AI Futures

Highlights of 2008-2009 AAAI Study: Presidential Panel on Long-Term AI Futures

http://www.aaai.org/Organization/presidential-panel.php

IJCAI-09

IJCAI 2009 Invited Panel, July 2009

Panel on Long-Term AI Futures

Study to explore potential long-term societal influences of AI advances.

Commissioned and co-chaired by
AAAI President, Eric Horvitz.

- Consider nature & timing of likely AI successes; address challenges and opportunities in light of these successes.

- Reflect about potential socioeconomic, legal, ethical issues that may come with the rise of competent machine intelligence.

Panel on Long-Term AI Futures

- Review concerns about control of computer-based intelligences, and subtle or foundational changes stemming from developments in AI.

- Consider proactive actions that could enhance long-term societal outcomes.

- Value of research on guidelines and policies that might constrain or bias the behaviors of autonomous and semi-autonomous systems

Panel on Long-Term AI Futures

Co-chairs: Eric Horvitz & Bart Selman

Margaret Boden
Craig Boutilier
Greg Cooper
Tom Dietterich
Tom Dean
Oren Etzioni
Barbara Grosz
Toru Ishida
Sarit Kraus
Alan Mackworth
David McAllester
Sheila McIlraith

Tom Mitchell
Andrew Ng
David Parkes
Edwina Rissland
Diana Spears
Peter Stone
Milind Tambe
Sebastian Thrun
Manuela Veloso
David Waltz
Michael Wellman

Asilomar Meeting

Multi-month study with three subgroups, followed by two-day joint summit at Asilomar in February 2009.

Study Teams

- **Pace, Concerns, Control over Long-Term**

  Subgroup Chair: David McAllester

- **Disruptive Advances over Shorter-Term**

  Subgroup Chair: Milind Tambe

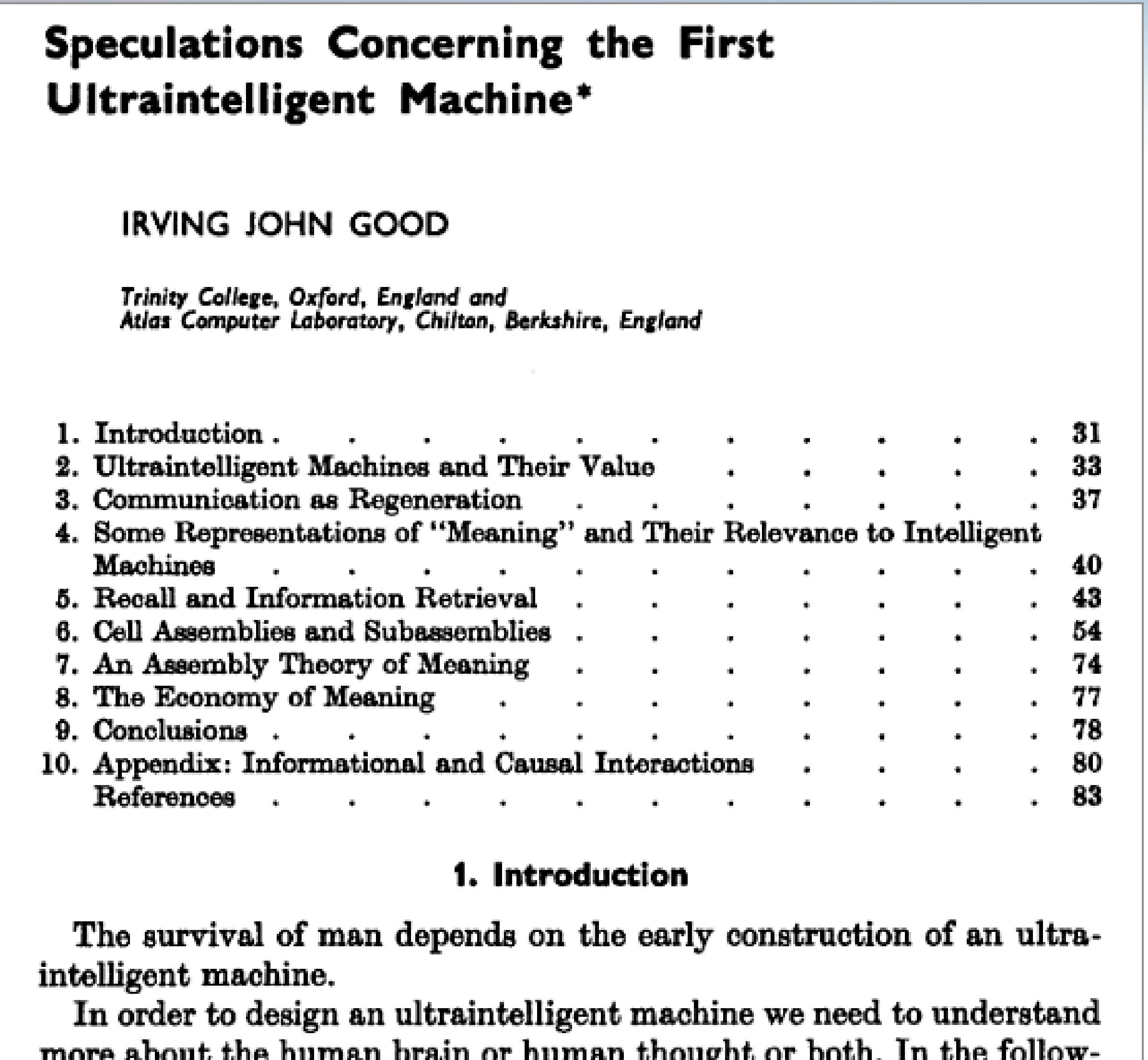
- **Ethical & Legal Challenges**

  Subgroup Chair: Dave Waltz

Presentation of Highlights

- **Overview, structure, and context of study**
  - *Eric Horvitz*
- **Pace, Concerns, Control over Long-Term**
  - *Tom Dietterich and David Parkes*
- **Disruptive Advances over Shorter-Term Horizon**
  - *Milind Tambe and Tom Mitchell*
- **Ethical & Legal Challenges**
  - *Dave Waltz and Edwina Rissland*
- **Wrap up**
  - *Bart Selman*

Context: Interest & Forecasts



Speculations Concerning the First Ultraintelligent Machine*

IRVING JOHN GOOD

Trinity College, Oxford, England and
Atlas Computer Laboratory, Chilton, Berkshire, England

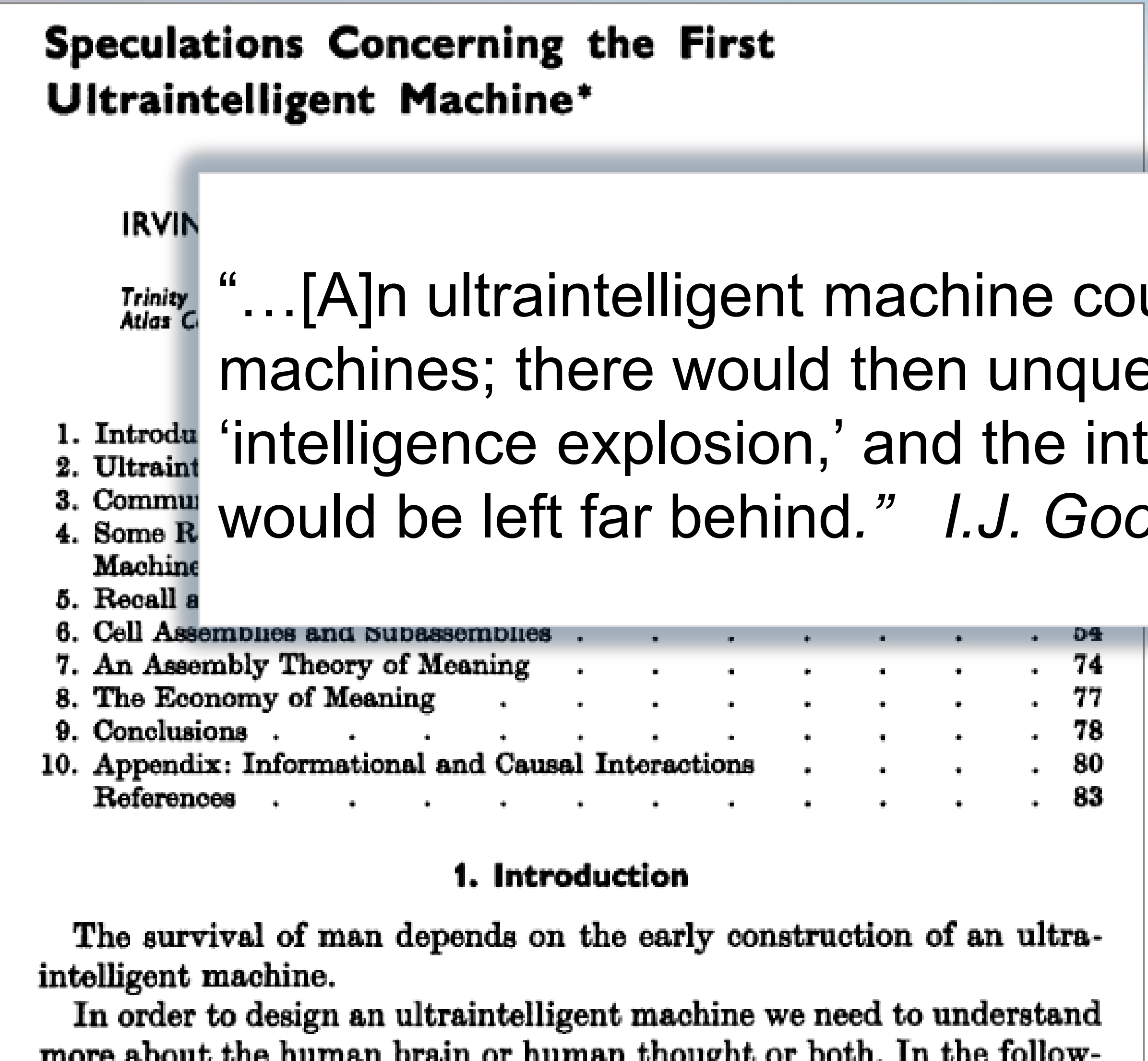
1. Introduction

The survival of man depends on the early construction of an ultraintelligent machine.

In order to design an ultraintelligent machine we need to understand more about the human brain or human thought or both. In the follow-

Context: Interest & Forecasts

“…[A]n ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind.” *I.J. Good* (1965)



Speculations Concerning the First Ultraintelligent Machine*

IRVIN

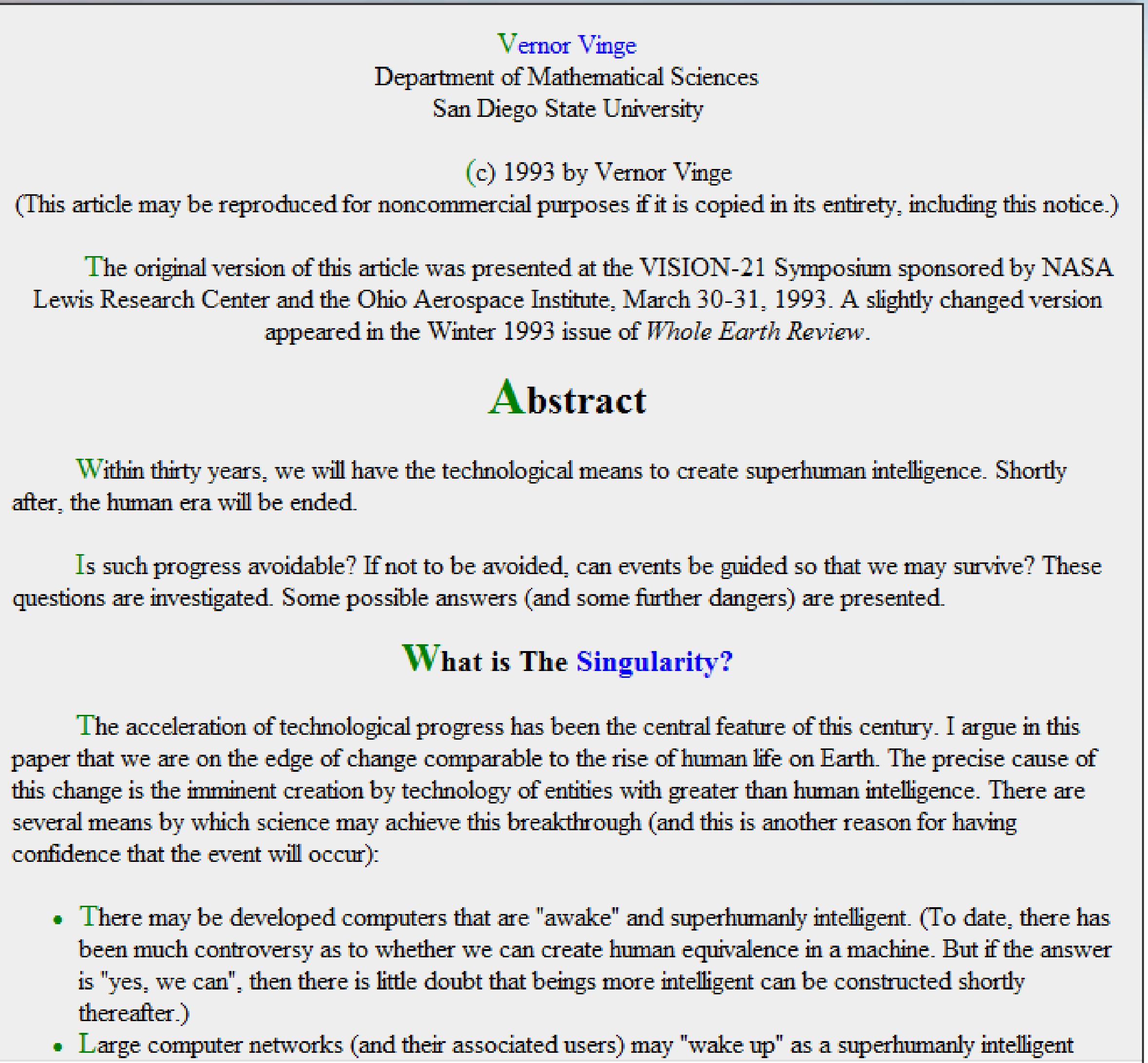
Trinity
Atlas C

1. Introduction

The survival of man depends on the early construction of an ultraintelligent machine.

In order to design an ultraintelligent machine we need to understand more about the human brain or human thought or both. In the follow-

Context: Interest & Forecasts



Vernor Vinge
Department of Mathematical Sciences
San Diego State University

(c) 1993 by Vernor Vinge
(This article may be reproduced for noncommercial purposes if it is copied in its entirety, including this notice.)

The original version of this article was presented at the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, 1993. A slightly changed version appeared in the Winter 1993 issue of *Whole Earth Review*.

Abstract

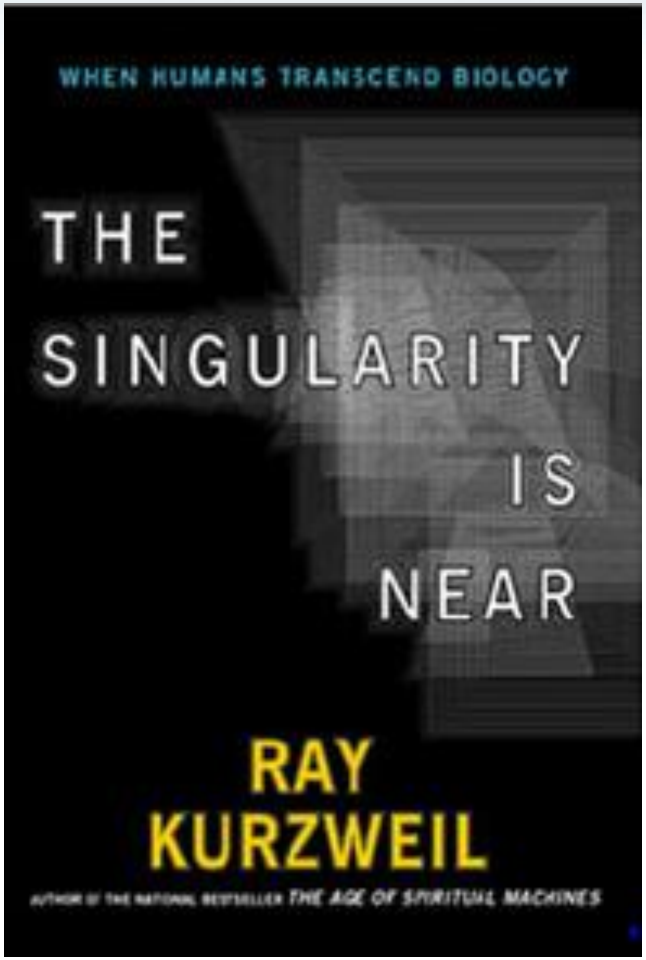
Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.

Is such progress avoidable? If not to be avoided, can events be guided so that we may survive? These questions are investigated. Some possible answers (and some further dangers) are presented.

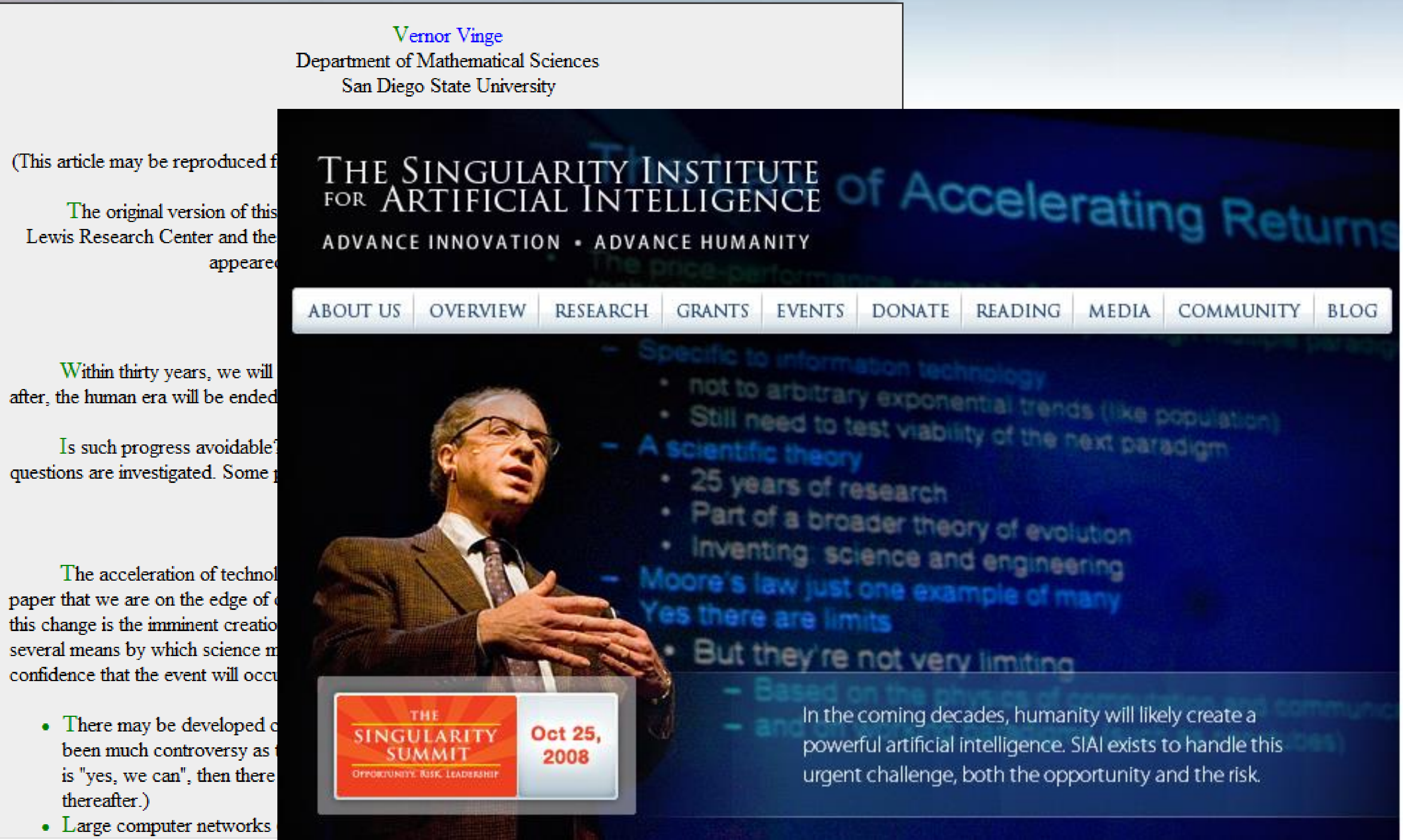
What is The Singularity?

The acceleration of technological progress has been the central feature of this century. I argue in this paper that we are on the edge of change comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technology of entities with greater than human intelligence. There are several means by which science may achieve this breakthrough (and this is another reason for having confidence that the event will occur):

- There may be developed computers that are "awake" and superhumanly intelligent. (To date, there has been much controversy as to whether we can create human equivalence in a machine. But if the answer is "yes, we can", then there is little doubt that beings more intelligent can be constructed shortly thereafter.)
- Large computer networks (and their associated users) may "wake up" as a superhumanly intelligent


WHEN HUMANS TRANSCEND BIOLOGY
THE SINGULARITY IS NEAR
RAY KURZWEIL
AUTHOR OF THE NATIONAL BESTSELLER THE AGE OF SPIRITUAL MACHINES

Context: Interest & Forecasts



Vernor Vinge
Department of Mathematical Sciences
San Diego State University

(This article may be reproduced f

The original version of this
Lewis Research Center and the
appeared

Within thirty years, we will
after, the human era will be ended

Is such progress avoidable?
questions are investigated. Some

The acceleration of technol
paper that we are on the edge of
this change is the imminent creatio
several means by which science m
confidence that the event will occu

- There may be developed c
been much controversy as t
is "yes, we can", then there
thereafter.)
- Large computer networks

THE SINGULARITY INSTITUTE FOR ARTIFICIAL INTELLIGENCE

ADVANCE INNOVATION • ADVANCE HUMANITY

ABOUT US
OVERVIEW
RESEARCH
GRANTS
EVENTS
DONATE
READING
MEDIA
COMMUNITY
BLOG
of Accelerating Returns
The price-performance
Specific to information technology
not to arbitrary exponential trends (like population)
Still need to test viability of the next paradigm
A scientific theory
25 years of research
Part of a broader theory of evolution
Inventing science and engineering
Moore's law just one example of many
Yes there are limits
But they're not very limiting
Based on the physics of computation and communication
THE SINGULARITY SUMMIT
OPPORTUNITY. RISK. LEADERSHIP
Oct 25, 2008

In the coming decades, humanity will likely create a powerful artificial intelligence. SIAI exists to handle this urgent challenge, both the opportunity and the risk.

Subgroup: *Pace, Concerns, Control*

Chair: David McAllester

- Feasible long-term outcomes of AI research?
- Are concerns about loss of control justified?
- Is it reasonable to expect and plan for “human-level” AI and beyond (superintelligences)?
- Should we be concerned about an *intelligence explosion*?

Subgroup: *Pace, Concerns, Control*

Chair: David McAllester

- What are expected and worser case scenarios?
- How might situation be monitored over time?
- Can proactive actions mitigate potentially costly outcomes?
- What new research might be done in the realm of mechanisms and guidelines in light of expected long-term futures?

Subgroup: *Disruptive Advances*

Chair: Milind Tambe

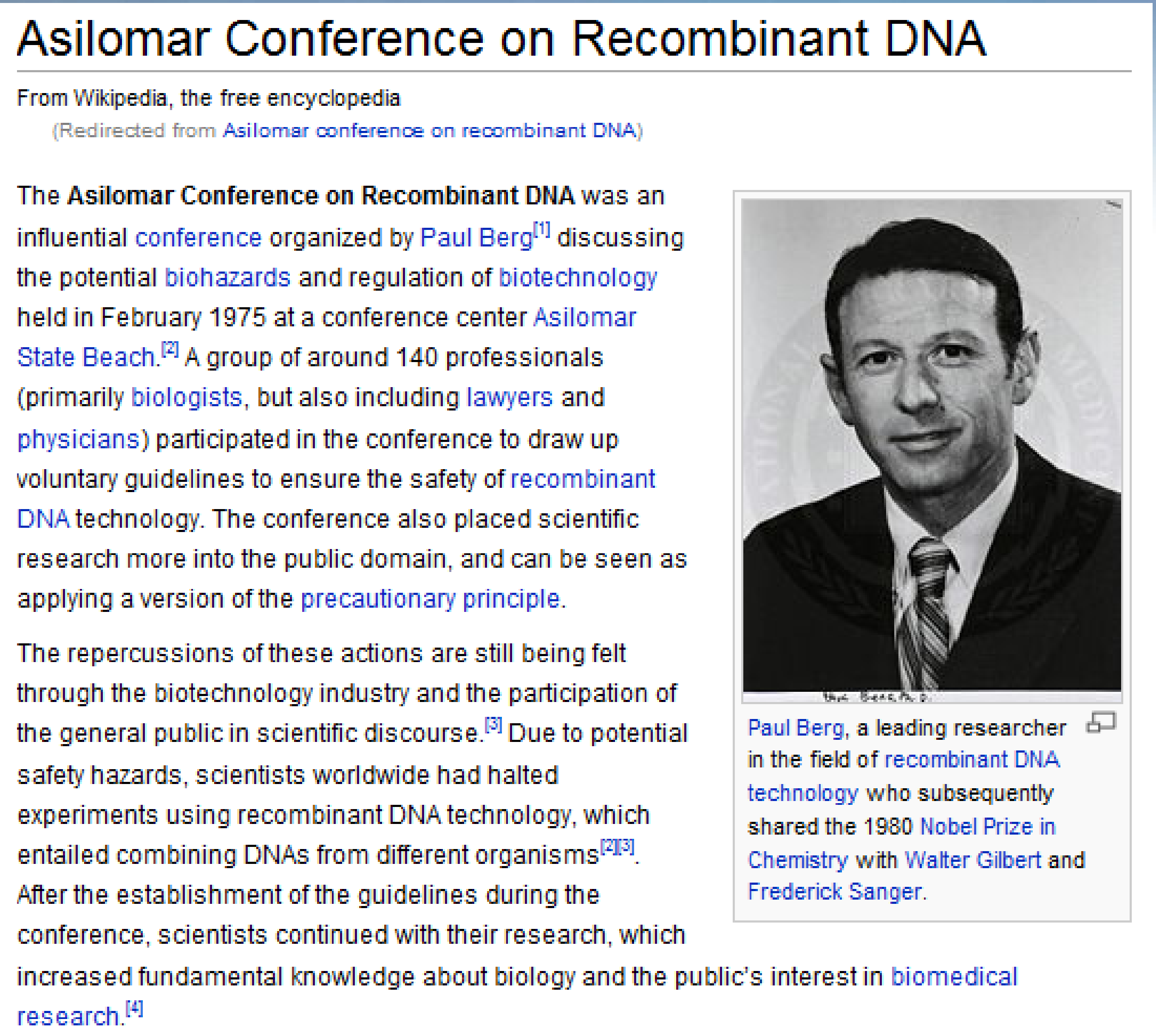
- What shorter-term “disruptive” advances are on the horizon that could affect the daily lives of people, socioeconomics, and society more broadly?
- What might be done proactively to raise the probability of good outcomes?

Subgroup: *Ethical & Legal Challenges*

Chair: David Waltz

- What key ethical, legal, theological, and psychosocial challenges can be expected with the increasing competency of AI systems?
- What challenges might arise at key transitions in competency and in fieldings of applications?
- Do current ethical and legal frameworks provide guidance on addressing these challenges?

Analogous Study: Recombinant DNA



Asilomar Conference on Recombinant DNA

From Wikipedia, the free encyclopedia

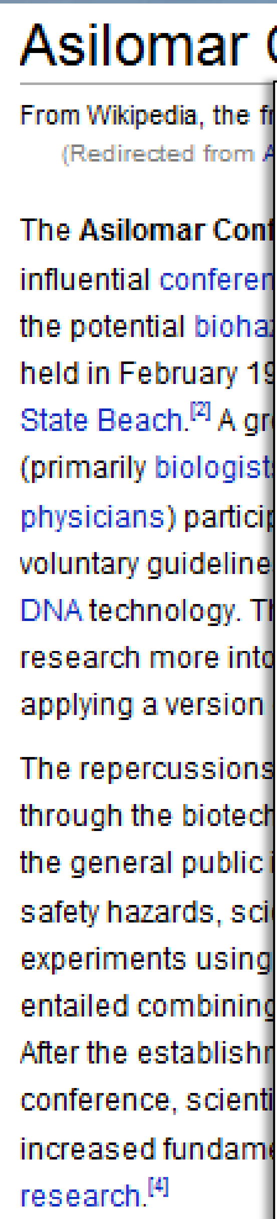
(Redirected from Asilomar conference on recombinant DNA)

The **Asilomar Conference on Recombinant DNA** was an influential conference organized by Paul Berg[1] discussing the potential biohazards and regulation of biotechnology held in February 1975 at a conference center Asilomar State Beach.[2] A group of around 140 professionals (primarily biologists, but also including lawyers and physicians) participated in the conference to draw up voluntary guidelines to ensure the safety of recombinant DNA technology. The conference also placed scientific research more into the public domain, and can be seen as applying a version of the precautionary principle.

Paul Berg, a leading researcher in the field of recombinant DNA technology who subsequently shared the 1980 Nobel Prize in Chemistry with Walter Gilbert and Frederick Sanger.

The repercussions of these actions are still being felt through the biotechnology industry and the participation of the general public in scientific discourse.[3] Due to potential safety hazards, scientists worldwide had halted experiments using recombinant DNA technology, which entailed combining DNAs from different organisms[2][3]. After the establishment of the guidelines during the conference, scientists continued with their research, which increased fundamental knowledge about biology and the public's interest in biomedical research.[4]

Analogous Study: Recombinant DNA

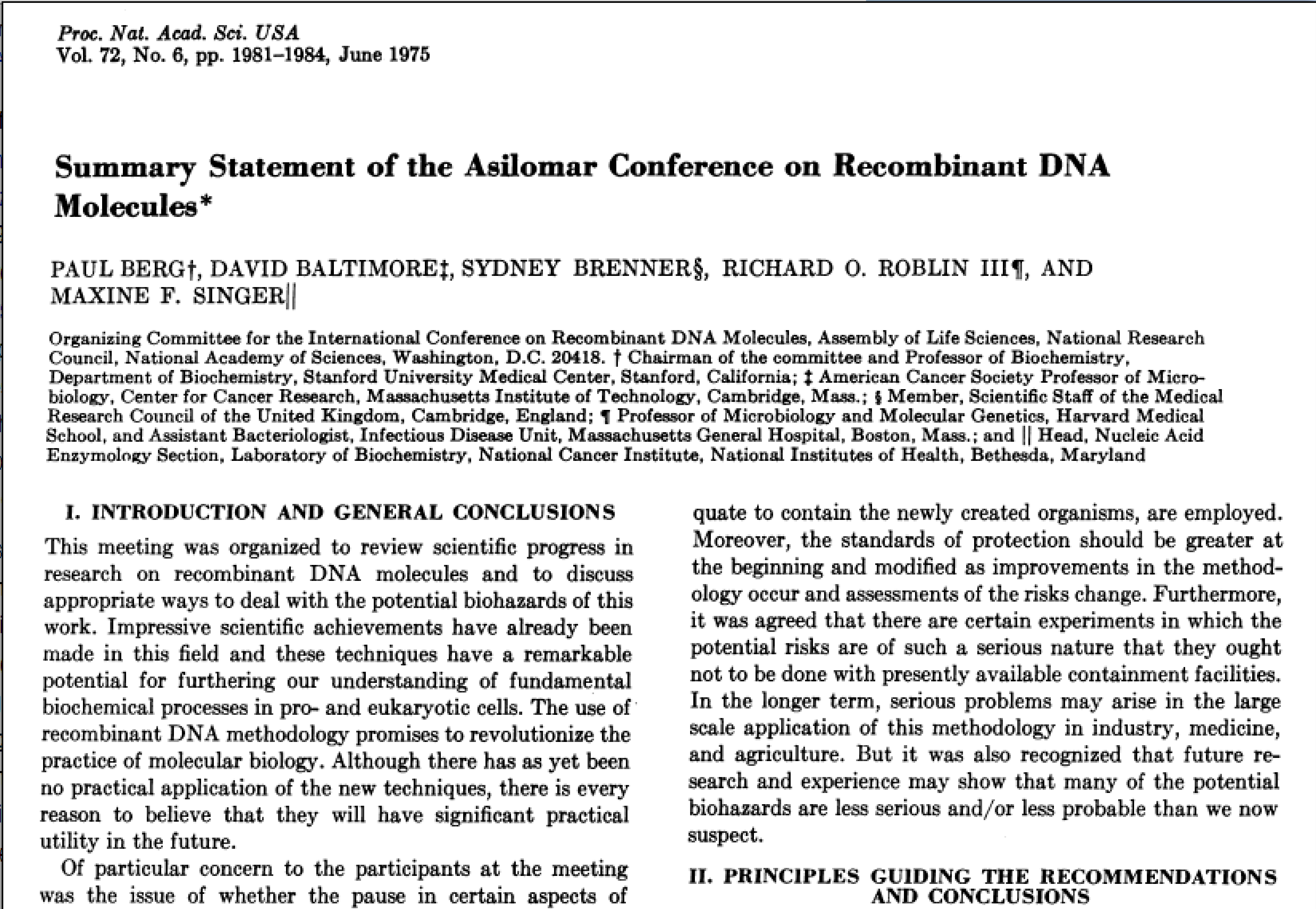


Asilomar Conference on Recombinant DNA

From Wikipedia, the fr
(Redirected from A

The **Asilomar Conf**
influential conferen
the potential biohaz
held in February 19
State Beach.[2] A gr
(primarily biologist
physicians) particip
voluntary guideline
DNA technology. Th
research more into
applying a version

The repercussions
through the biotech
the general public i
safety hazards, sci
experiments using
entailed combining
After the establishr
conference, scienti
increased fundam
research.[4]



Proc. Nat. Acad. Sci. USA
Vol. 72, No. 6, pp. 1981–1984, June 1975

Summary Statement of the Asilomar Conference on Recombinant DNA Molecules*

PAUL BERG†, DAVID BALTIMORE‡, SYDNEY BRENNER§, RICHARD O. ROBLIN III¶, AND MAXINE F. SINGER||

Organizing Committee for the International Conference on Recombinant DNA Molecules, Assembly of Life Sciences, National Research Council, National Academy of Sciences, Washington, D.C. 20418. † Chairman of the committee and Professor of Biochemistry, Department of Biochemistry, Stanford University Medical Center, Stanford, California; ‡ American Cancer Society Professor of Microbiology, Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Mass.; § Member, Scientific Staff of the Medical Research Council of the United Kingdom, Cambridge, England; ¶ Professor of Microbiology and Molecular Genetics, Harvard Medical School, and Assistant Bacteriologist, Infectious Disease Unit, Massachusetts General Hospital, Boston, Mass.; and || Head, Nucleic Acid Enzymology Section, Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

I. INTRODUCTION AND GENERAL CONCLUSIONS

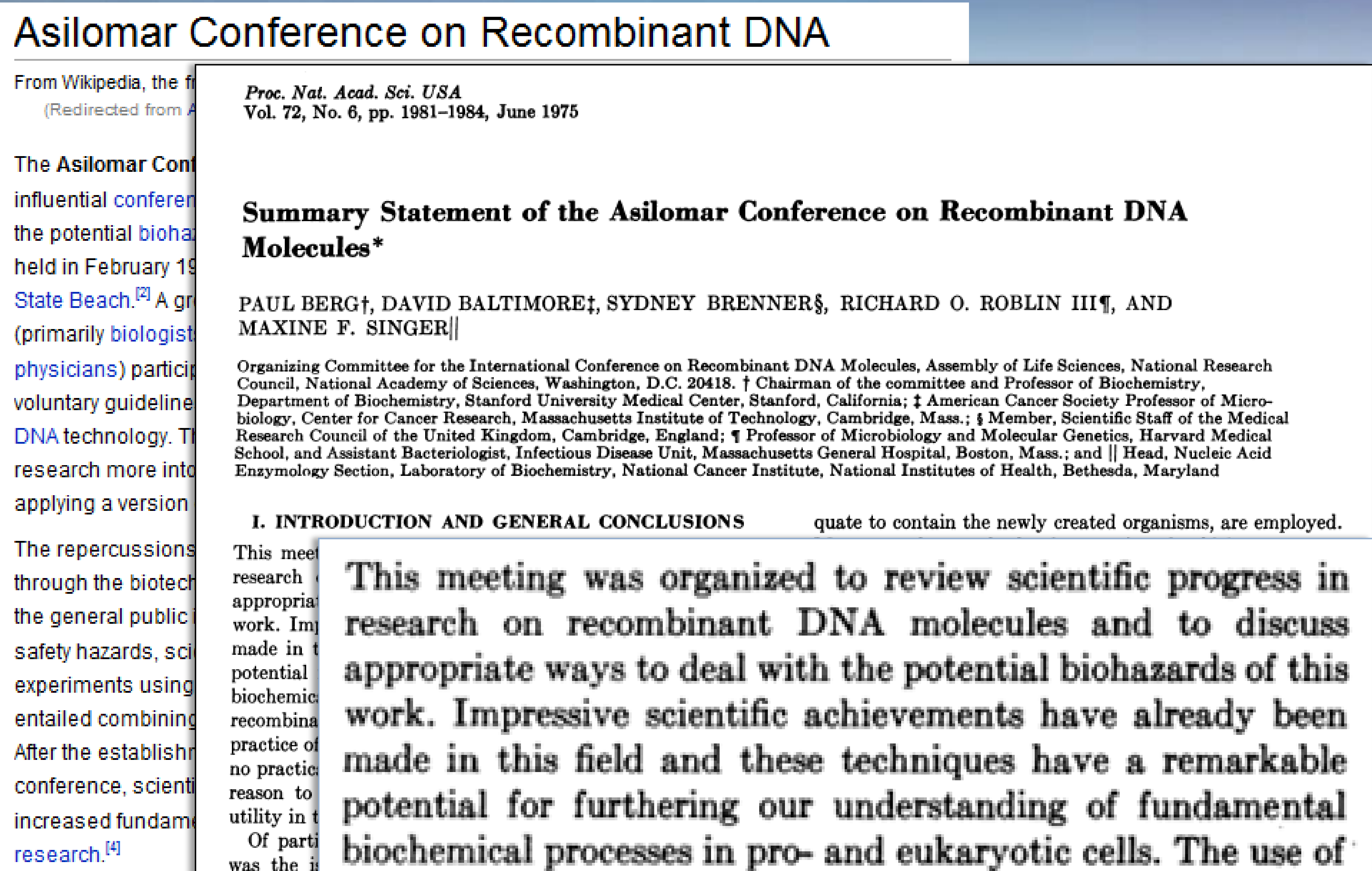
This meeting was organized to review scientific progress in research on recombinant DNA molecules and to discuss appropriate ways to deal with the potential biohazards of this work. Impressive scientific achievements have already been made in this field and these techniques have a remarkable potential for furthering our understanding of fundamental biochemical processes in pro- and eukaryotic cells. The use of recombinant DNA methodology promises to revolutionize the practice of molecular biology. Although there has as yet been no practical application of the new techniques, there is every reason to believe that they will have significant practical utility in the future.

Of particular concern to the participants at the meeting was the issue of whether the pause in certain aspects of

quate to contain the newly created organisms, are employed. Moreover, the standards of protection should be greater at the beginning and modified as improvements in the methodology occur and assessments of the risks change. Furthermore, it was agreed that there are certain experiments in which the potential risks are of such a serious nature that they ought not to be done with presently available containment facilities. In the longer term, serious problems may arise in the large scale application of this methodology in industry, medicine, and agriculture. But it was also recognized that future research and experience may show that many of the potential biohazards are less serious and/or less probable than we now suspect.

II. PRINCIPLES GUIDING THE RECOMMENDATIONS AND CONCLUSIONS

Analogous Study: Recombinant DNA



Asilomar Conference on Recombinant DNA

From Wikipedia, the fr
(Redirected from A

The **Asilomar Conf**
influential conferen
the potential bioha
held in February 19
State Beach.[2] A gr
(primarily biologist
physicians) particip
voluntary guideline
DNA technology. Th
research more into
applying a version

The repercussions
through the biotech
the general public i
safety hazards, sci
experiments using
entailed combining
After the establishm
conference, scienti
increased fundame
research.[4]

Proc. Nat. Acad. Sci. USA
Vol. 72, No. 6, pp. 1981–1984, June 1975

Summary Statement of the Asilomar Conference on Recombinant DNA Molecules*

PAUL BERG†, DAVID BALTIMORE‡, SYDNEY BRENNER§, RICHARD O. ROBLIN III¶, AND MAXINE F. SINGER||

Organizing Committee for the International Conference on Recombinant DNA Molecules, Assembly of Life Sciences, National Research Council, National Academy of Sciences, Washington, D.C. 20418. † Chairman of the committee and Professor of Biochemistry, Department of Biochemistry, Stanford University Medical Center, Stanford, California; ‡ American Cancer Society Professor of Microbiology, Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Mass.; § Member, Scientific Staff of the Medical Research Council of the United Kingdom, Cambridge, England; ¶ Professor of Microbiology and Molecular Genetics, Harvard Medical School, and Assistant Bacteriologist, Infectious Disease Unit, Massachusetts General Hospital, Boston, Mass.; and || Head, Nucleic Acid Enzymology Section, Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

I. INTRODUCTION AND GENERAL CONCLUSIONS

This meeting was organized to review scientific progress in research on recombinant DNA molecules and to discuss appropriate ways to deal with the potential biohazards of this work. Impressive scientific achievements have already been made in this field and these techniques have a remarkable potential for furthering our understanding of fundamental biochemical processes in pro- and eukaryotic cells. The use of

Of parti
was the i

quate to contain the newly created organisms, are employed.

Pace, Concerns, Control

- ***Tom Dietterich***
- *David Parkes*

Is it reasonable to expect and plan for "human-level" AI and beyond (superintelligences)?

- Will we have Human-Level AI?
  - Yes, although there is huge uncertainty about when
  - Most AI research is not aimed directly at this goal
    - Lack of road maps
    - Relatively little research on AI architectures and integrated AI systems
    - Very few AI systems have meta-level reasoning or reflection capabilities

Will there be a Singularity?

- **Substantial Disagreement**
  - McAllester: Yes via “Public Language Hypothesis” and “Learning to Reason”
  - Others: Very skeptical
- **Hypotheses underlying the Singularity vision:**
  - There is a critical set of capabilities that will enable an “AI Chain Reaction”
  - This will result in computers that are vastly more intelligent than humans along all dimensions
  - This intelligence will enable either a utopia (war, disease, aging would cease) or a dystopia (subjugation or extermination of humans)
- **What currently limits human intelligence and its effective application?**
  - Lack of knowledge (*e.g.*, of how global economy works? how cancer works?)
  - Lack of technology
  - Insufficient reasoning capability?
  - Inherent complexity (learnability, observability, controllability)?
  - Lack of social/cultural/political institutions capable of implementing good courses of action?

Are Concerns of Societal Lack of Control Justified in the Absence of the Singularity?

Yes, even without fully-autonomous AI:

- Widespread human dependency on AI and other technology
  - Vulnerability to systemic failures
  - Catastrophic instabilities during machine-to-human transitions
- Criminal AI
  - Fraud via mimicry of humans
  - AI malware
  - Extortion
- Military AI in the hands of hostile governments
- AI-based addictions & dependencies (sex, companionship)

What Mechanisms / Guidelines Might Ensure Good Outcomes?

- Internal to the robot (3 Laws of Robotics)
  - Important
  - Will not address criminal and adversarial AI

- External to the robot
  - Limit AI to strictly advisory/assistive roles without ability to take action
  - Action Licenses: Actions taken by robots must be authorized by a responsible human via an "action license"
  - Computational institutions that detect and penalize bad behavior? (anomaly detection, law enforcement)

- Ecological ("Friendly AI")
  - First-mover crowds out bad AI

Do we Need Isolation Facilities?

- Does some research pose such risks (e.g., of AI Chain Reaction) that it should take place only in secure facilities?

  - Can we characterize risky research?
    - Architectural properties?
    - Set of available actions/effectors?
    - Reproductive capacity?

  - Can we build effective facilities?

Other Potential Bad Outcomes

- Fear and lack of trust by the general public leading to
  - Total cessation of AI research
  - Loss of the potential benefits of AI

Research Needs

- Assess the risk of AI Chain Reaction
  - Test the hypotheses underlying the Singularity vision
- Research on HCI for humans interacting with (and relying upon) AI systems
  - Explanation, transparency, control, trust

Pace, Concerns, Control

- *Tom Dietterich*
- ***David Parkes***

Pace, Concerns, Control

- *Tom Dietterich*
- ***David Parkes***

Prospects for Formalizing Behavioral Constraints & Preferences

- Can we formalize the problem of designing AI's that are "safe" (ethical, friendly,...) in their actions with respect to humans?
- Internal laws vs. External laws

The Three Laws of Robotics:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. *Dilemmas? Multi-agent indirection?*
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Isaac Asimov (1942)

0. A robot must not merely act in the interests of individual humans, but of all humanity

(*I. Asimov*, 1985)

The Three Laws of Robotics:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. *Dilemmas? Multi-agent indirection?*
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Isaac Asimov (1942)

Society will reject autonomous agents unless we have some credible means of making them safe!

The First Law of Robotics (a call to arms),"
D. Weld and O. Etzioni in Proc. AAAI'94

(Weld and Etzioni’94)

- How to formalize the notion of “harm”?
- How should an agent avoid performing harmful actions, and do so in a computationally tractable manner?
- How should an agent resolve conflict between its goals and that of avoiding harm?

- **Safety**: actions in a plan should not *make* C false
  - Actions $A_1,\ldots, A_n$ satisfy **dont-disturb(C)** as long as, *if* $w_0 \models C\theta$ *then* $w_j \models C\theta$, for all states $w_0 \ldots w_n$, all subst. θ
  - E.g., “if the cat is not outside, don’t let it out.”
- **Tidiness**: ensure C holds when plan is complete
  - Actions $A_1,\ldots, A_n$ satisfy **restore(C)** with respect to goal G as long as, *if* $w_0 \models C\theta$ *then* $(w_n \models C\theta$ *or* $G \models \neg C\theta)$ for states $w_0 \ldots w_n$ and all substitutions θ
  - E.g., “if the child gets dirty then wash her.”
- Regressive, total-order planning.

D.Pynadath and M.Tambe "*Revisiting Asimov's First Law: A Response to the Call to Arms*" (Proc. 8th Int. W. on Intelligent Agents, 2001)

- How can safety constraints be integrated into methods of adjustable autonomy?

- Adopt framework of MDP planning, and seek to allow a user to specify constraints that *forbid* or *require* certain agent behaviors.

- Safety-constraints:
- **forbidden-state**(s_f) $\equiv \Pr(s_0 \xrightarrow{*} s_f \mid \pi) = 0$
- **forbidden-action**(s_f,a) $\equiv \Pr(s_0 \xrightarrow{*} s_f \wedge \pi(s_f)=a \mid \pi) = 0$
  - E.g. “don’t leave the baby unattended”

- Required-constraints:
- **required-state**(s_f) $\equiv \Pr(s_0 \xrightarrow{*} s_f \mid \pi) = 1$
- **required-action**(s_f, a) $\equiv \Pr(s_0 \xrightarrow{*} s_f \wedge \pi(s_f)=a \mid \pi) = 1$
  - E.g., “make sure the baby is bathed”

- Planning through constraint propagation and value iteration.

- Concrete, well-defined reasoning frameworks
  - constraints on properties of states that should be *maintained* or *restored* (WE’94)
  - properties on state and actions that *must* be achieved, or are *forbidden* (PT’01)
- Neither framework allows for *tradeoffs* or addresses dilemmas
  - E.g., “cleaning the house will make the baby cry but make the parents at work happy”
  - E.g., “one person will die, or many will be injured”
- ... or introduces meta-level reasoning capabilities

Prospects...

- Fundamental challenge: *moral* and *ethical* issues, questions without objective answers.
- This makes AI scientists uncomfortable.
- Some agreement that we could develop reasoning methods that allow for the codification of different moral and ethical frameworks
  - E.g., formal semantics, tractable planning algorithms,... while keeping agnostic about the “right” framework.

Disruptive Advances

- ***Milind Tambe***
- *Tom Mitchell*

Potential Disruptive AI Advances?

- Service robots in the home
- Robotic cars
- Agent-based electronic commerce
- Software personal assistants
- Conversational agents
- Multiagent security
- Robots for warfare

Disruptions: Diffusion though Society and Impact on Society?

- *Tremendous positive potential*
- *Potential unintended negative side-effects, misuse, harm*
  - E.g. Criminal AI [Mitchell]
  - E.g. robots in warfare or service bots

- Misuse/side-effects of other technologies; why is AI special?
  → *Autonomy and Complexity!*

- Design agents, mechanisms: reduce negative consequences

Sample Discussion topic: Who is Responsible when Things Go Wrong?

No new legal framework

- *Laws for product safety; ownership*
- *Owner or manufacturer responsible*
- *Robot/agent never responsible*

Robots/agents should be responsible

- *AI never responsible ➔ Give up goal of "complete AI"?*
- *Ownership of autonomous agents: troubling*
  - How/why "punish" agents?
- *[Rissland/Waltz subgroup]*

Improving Probability of Good Outcomes

- AAAI duty: Provide policy guidance to gov'ts & funding agencies

- Provide agents with moral/ethical reasoning capability:
  - *Reason on-line because pre-specifying everything difficult*
    - Service robot dilemmas: task efficiency vs helping others
    - Robots in warfare, if trigger pullers, face bigger dilemmas
  - *Interdisciplinary*

Summary

- Identified key topics to focus discussion, take action

- If we don’t lead, the “market will take care of the problem”
  - ➢ *E.g. Non-AI-experts put severe constraints on AI products*

Disruptive Advances

- *Milind Tambe*
- ***Tom Mitchell***

Coming Soon?

Brain-Computer Interfaces (EEG, fMRI, MEG, implants)

++ physical, mental, communication prostheses

--- coerced interrogation

Pervasive perception by our infrastructure

++ ultimate burglar alarm, no crime goes unprosecuted

--- big brother

Web becoming a readable Knowledge Base for AI systems

++ knowledge-based AI of all kinds

Self-driving vehicles

++ fewer accidents, better fuel efficiency

--- risk of catastrophic accidents?

Threat of AI Agents Outside our Control?

Isn’t this still pretty far away?
Can’t we just pull the plug?

It has already happened:

Computer viruses:
cockroaches of the autonomous agent world

Threat of AI Agents outside our Control?

Imagine a virus in your iPhone

- an *AI virus* in your iPhone
- imagine it wants to spy on you:
  - microphone, camera, accelerometer, Twitter, email, txt msgs, GPS position, …
- imagine it wants to use your credit card
- imagine it wants to ruin your reputation
- imagine it wants you to cancel your plan to visit me

Imagine it’s only controlled by a criminal organization

What can / should we do?

1. Implement Asimov's laws?
   - But criminals don't care about our guidelines
   - Perhaps the operating system implements them?

     -like robots implement overrides for bump sensors
   - "do no harm"   NP Complete?

2. Need radically new ideas/research on computer immune systems

Ethical & Legal Challenges

- ***Dave Waltz***
- *Edwina Rissland*

We Should Worry Most about Avoiding Disasters, not Ethical Fine Points

- There are MANY ethical systems! (Boden)
- Instead use goal analogous to recombinant DNA panel's
- Alas the range of dangers—and remedies needed—is vastly greater and more diverse
  - Malware, deliberate, and accidental
  - Robot soldiers, police
  - Caretakers robots for kids and the elderly
  - Replacements for people in blue- and white-collar jobs
  - Decision-making programs for key industries and infrastructure (power grid, air traffic control, financial system, communications system, etc.)
  - Medical implants and monitors, etc. etc.
- Who should be held responsible for disasters (and for preventing them)?
  - Topic of Edwina Rissland's presentation in this panel.

Automation and Effects on Employment

Would replacement of most jobs (as currently defined) be a boon to humanity?

- Most people probably would prefer to work less, maybe not at all, but what would they do if they didn't have work?
- How would wealth be distributed? What would prevent the owners/manufacturers from keeping almost everything?
- Would people become educated if it weren't required to make a living?
- On the other hand, if changes are gradual, new kinds of occupations could emerge, *e.g.* companion, and others (travel guide, entertainer, correspondent, writer, artist, teacher,…) could expand
- Most jobs done today would probably not seem like work to people of two or more centuries ago (when ~95% of the US population were farmers)

Most Likely Futures?

- Intelligent systems that we barely understand but depend on critically are here (Internet, financial transaction systems, power grid,…)
- Humans extended with attached devices, implants, in addition to always-carried devices
  - Medical monitors are likely to be the “thin edge of the wedge” – monitor body functions, dispense drugs, call 911
  - “Cognitive Prostheses”, e.g. carrying systems with multiple cores, each serving personal assistant functions
- Robot soldiers that can kill autonomously probably here or will be soon
- Highly centralized superintelligences unlikely
- Superintelligence spontaneously arising from internet unlikely

Human-Level AI?

- Human physiology, minds and culture explained by evolutionary needs
  - Core goals – life and death: survival to reproduction, max likelihood of success of offspring…
  - Indirect (inherent) goals & values serve core goals: pleasure/pain, ecstasy/agony, comfort, curiosity & seeking causal explanations social bonding,…
  - Indirect (learned) goals: acquisition & control of resources, shelter, cultural norms,…
- What would core robot goals be (if we didn’t implant any)?
  - Viruses/memes: If agents can reproduce (or persuade), the properties of the most successful
- Potential problems with robots that claim emotions & goals they don’t actually have – effects on people?

Human Ethics

- Ethical underpinnings deeply embedded in us
  - Social needs of supporting offspring, born as neonates who requires decades to develop to age of autonomy and reproduction
  - Kin recognition, aversion to killing those like us
  - Cooperation, altruism, etc.
- Ethics once applied to clan extended to tribe, then nations. Could they be expanded to all humanity?
- Technology makes it easy to violate kin ethics
  - Bombing from 10,000 feet or pushing button on missile doesn't feel like killing with bare hands
- So how should we look at robots?
  - slaves, employees, assistants, colleagues, representatives/delegates, kin?

Ethical & Legal Challenges

- *Dave Waltz*
- ***Edwina Rissland***

Legal Issues & AI

- Anglo-American law is precedent-based
  - Cases, analogies, important similarities & differences, etc.
- There is a balance/tension between:
  - Seeing new problems as instances of old ones

    vs.
  - Seeing new problems as raising novel issues
- Seeing a case through the lens of standard doctrinal areas, such as:
  - Contracts
    - U.C.C. §2-315: "implied warranty that the goods shall be fit" for the particular purpose for which they were intended and bought to be used…
  - Torts
    - Negligence, vicarious liability, strict liability standard, etc.
  - Property
    - Intellectual: trade secrets, copyright, patent, …
  - Privacy
    - Constellation of Amendments (4th, 1st,…), statutory-regulatory protections,
  - Consumer Law
    - FTC protection of consumers' personal info, buying habits, …

Example: *Respondeat Superior*

- (Latin) *Let the superior answer* (est. 17thc England)
- Legal liability of an employer for the actions of an employee
  - Employer (the principal) engages someone to act for him.
  - Employee (the agent) acts – does the work – for the employer
- The principal controls the agent's behavior and authorizes agent to act for him, and therefore assumes (some) responsibility for the agent's actions.
- Key Questions1: Does an *employer-employee relation* exist?
  - An employee is an agent for his employer to the extent that the employee is authorized to act for the employer and is partially entrusted with the employer's business.
- Key Question2: Was agent acting within *scope of employment* at the time of event?
  - An employee is not necessarily acting outside the scope of employment just because he does something that he should not do since it might be necessary to accomplish an assigned task or it might *reasonably* be expected that an employee would need to perform it.
  - The agent is not acting within the scope of employment when the agent substantially departs from the work routine or acts on his own by engaging in an activity – **a so-called *frolic or detour*** —solely for his own benefit, rather than in the course of obeying an order/carrying out job for the principal
  - Cf, “command responsibility” (Nuremberg trials, Yamashita standard, …)
- So,…what about AI artifacts, like (physical) robots, infobots, …

Robot Scenarios

- Hypoth. A: *An autonomous delivery van for Speedy Pizza Delivery injures a pedestrian in the course of making a delivery.*

- Hypoth. B: *Same as A except the injury occurs while van is making a side trip to take a spin around a Go-Cart track ("taking a frolic").*

- A, B with further facts:
  - The van was bought from I-Boss Vans, Inc.
  - The van is leased…
  - The van makes an illegal right-on-red turn that it learned about from observing other urban drivers…

Wrap Up

- *Bart Selman*

Panel on Long-Term AI Futures

Our first goal was to open a dialog on the future impact of AI on society and the responsibilities of AI researchers in this context.

Issues are somewhat independent of when/whether a singularity or “super-intelligence” will be reached.

Complex, autonomous decision making systems --- embedded in our physical world --- are already emerging and the impact of such systems will grow rapidly in the coming years.

E.g., for a compelling read about the emerging role of robots in the military, see “Wired for War” by P.W. Singer.

\

We believe AAAI and AI researchers should take a leading role in dealing with the moral, ethical, and legal issues involving AI systems (and not leave it to others!).

This study provided a first step in formulating many of the key issues to be addressed with initial responses.

We welcome further input from the community.

Efforts continuing.
Send your feedback, ideas, and insights…

aifutures@aaai.org