USES OF COUNTERFACTUALS

John McCarthy
Computer Science Department
Stanford University
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/
September 25, 2002

Tom Costello, now at IBM Almaden Research Center,
co-author of the article on which this talk is based.

A slogan for AI:
Whatever a person can do, he should be able to make
computer do for him.

Almost all of my papers are on the above web page.
APPROACHES TO ARTIFICIAL INTELLIGENCE

**biological**—Humans are intelligent; imitate humans observe and imitate at either the psychological or neuro-physiological level

**engineering**—The world presents problems to intelligence. Study information and action available in the world.
1. Write programs using non-logical representations.
2. Represent facts about the world in logic and decide what to do by logical inference.

We aim at human level AI, and the key phenomenon is the common sense informatic situation.
THE COMMON SENSE INFORMATIC SITUATION

- Involves approximate entities.

- There is no limitation on what information may be relevant. Theories must be elaboration tolerant.

- Reasoning must often be non-monotonic.

Common sense theories therefore contrast with formal scientific theories and most present AI theories. Science is embedded in common sense.
A LOGICAL ROAD TO HUMAN LEVEL AI

• Use *Drosophilas* that illustrate aspects of representation and reasoning problems.

• Concepts, context, circumscription, counterfactuals, consciousness, creativity, approximation

• narrative, projection, planning

• mental situation calculus

• domain dependent control of reasoning
USEFUL COUNTERFACTUALS

“If another car had come over the hill when you passed that car, there would have been a head-on collision.”

Such counterfactuals

• Are not usefully regarded as material conditionals with false antecedents. Believing the above as a tautology would not suggest driving more carefully.

• Can often be inferred from non-counterfactuals—within a common sense theory.

• Can have non-counterfactuals as consequences.
• Permit learning from experiences you don’t have and would rather not have.

• Counterfactuals about specific circumstances extend case based reasoning.

• Counterfactuals hold within theories.

• In order to provide for counterfactuals, the theories must be partial.

• The car-passing theory does not say whether another car will come over the hill.
• “If another car had come over the hill when you passed, there would have been a head-on collision.”

• (1) $\text{Carcomes}(\text{Present}) \supset \text{Collision}(\text{Present})$.

• Why believe it or disbelieve it?

• Some computer systems could measure and compute, but the unaided humans must estimate how close he was to the top of the hill.

• Consequence of believing (1):

$(\forall s)(\text{Similar}(s, \text{Present}) \land \text{Carcomes}(s) \rightarrow \text{Occurs}(\text{Collision}, s))$
IF ANOTHER CAR HAD COME OVER THE HILL—
\[
s = \sqrt{x^2 + y^2 + z^2}
\]
is the distance from a point \(P(x, y, z)\) to the origin.

Let \(P_0 = (1, 2, 1)\) be our current world. We ask whether
\[
y = 3 \succ s = \sqrt{19}.
\]
Our cartesian structure implies that \(x\) and \(z\) hold their particular values 1, 1. Therefore we would have
\[
s = \sqrt{1 + 9 + 1} = \sqrt{11} \neq \sqrt{19}.
\]
and (1) is therefore an untrue counterfactual. However, the counterfactual \(y = 3 \succ s = \sqrt{11}\) is true.
A change of theory, i.e. of co-ordinate systems, e.g. $x' = x + 0.1y$, $y' = y$, $z' = z$, changes which counterfactuals are true.
If Caesar had been in charge in Korea he would have used nuclear weapons.

* "If Caesar had been in charge in Korea he would have used catapults." is not useful.

If Pickett’s charge at Gettysburg had succeeded, the Confederacy would exist today.

If I had bought the stock promptly when the product was announced I’d have made more money.

If wishes were horses beggars would ride.
MATHEMATICAL COUNTERFACTUALS

There are useful mathematical counterfactuals.

• If, as Fermat conjectured, $2^{25} + 1$ were prime twice, it would be prime.

• If all algebraic integer domains had unique factorization, Kronecker would have proved the Fermat conjecture.

• A mathematical counterfactual is true in a partial theory, maybe proof-theoretically partial.
SKIING

• The stick figure theory of skiing.

• If he had bent his knees he wouldn’t have fallen.

• No. If he had put his weight on his downhill ski wouldn’t have fallen.

• If he had taken two more lessons he wouldn’t have fallen.

• The *stick figure theory of skiing* is shared by the instructors arguing about why the skier fell. It infers that the student will fall if he doesn’t bend his knees or shift his weight properly but not why he does or doesn’t.

• The *theory of skiing lessons* says that skiers with more lessons bend their knees when they should.
POSSIBLE WORLDDS

• Metric structures are not often as useful as Cartesian structures.

• The theory of counterfactuals needs to be based in incomplete structures.
Counterfactuals inhabit approximate theories.

Counterfactuals can become cartesian in suitable approximate theories.


The theory of the car passing incident does not take into account what might make a car come over the hill.

The simple skiing theory doesn’t take into account what might make the skier bend his knees. The theory about skiing lessons does.
• The car-crash counterfactual is complicated by being situated in a partially observable actual situation. It doesn’t take into account the actual speeds of objects coming over the hill.
CONCLUSIONS

• Some counterfactuals are useful.

• Useful counterfactuals often have non-counterfactual consequences.

• Cartesian counterfactuals are the easiest.

• Counterfactuals inhabit approximate theories.

• This lecture advertises the article by Tom Costello and John McCarthy in *Electronic Transactions in Artificial Intelligence*. See http://www.ida.liu.se/ext/epa/ej/etai/1999/A/index.html. The article is also http://www.formal.stanford.edu/jmc/counterfactuals.html.