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 mak
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 neu physiological level

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

We aim at human level AI, and the key phenomeno the common sense informatic situation.

THE COMMON SENSE INFORMATIC SITUATIO

- Involves approximate entities.

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

- Reasoning must often be non-monotonic.

Common sense theories therefore contrast with for scientific theories and most present AI theories. Scie is embedded in common sense.

A LOGICAL ROAD TO HUMAN LEVEL AI

- Use *Drosophilas* that illustrate aspects of represen tion and reasoning problems.

- Concepts, context, circumscription, counterfactu 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 pas
that car, there would have been a head-on collision."

Such counterfactuals

- Are not usefully regarded as material conditionals v false antecedents. Believing the above as a tautol would not suggest driving more carefully.

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

- Can have non-counterfactuals as consequences.

- Permit learning from experiences you don't have would rather not have.

- Counterfactuals about specific circumstances ext *case based reasoning*.

- Counterfactuals hold within theories.

- In order to provide for counterfactuals, the theories m be partial.

- The car-passing theory does not say whether anot car will come over the hill.

- "If another car had come over the hill when you pass there would have been a head-on collision."

- (1) $Carcomes(Present) \succ Collision(Present)$.

- Why believe it or disbelieve it?

- Some computer systems could measure and comp but the unaided humans must estimate how close he to the top of the hill.

- Consequence of believing (1):
$(\forall s)(Similar(s, Present) \land Carcomes(s)$
$\rightarrow Occurs(Collision, s))$

IF ANOTHER CAR HAD COME OVER THE HILL-

A CARTESIAN COUNTERFACTUAL

$s = \sqrt{x^2 + y^2 + z^2}$ is the distance from a point P (x, y, z) to the origin.

Let $P0 = (1, 2, 1)$. be our current world. We ask wheth

$$y = 3 \succ s = \sqrt{19}.$$

Our cartesian structure implies that x and z hold t 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. Howe 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 counterfactu are true.

EXAMPLES

- If Caesar had been in charge in Korea he would h used nuclear weapons.

- * “If Caesar had been in charge in Korea he would h used catapults.” is not useful.

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

- If I had bought the stock promptly when the proc 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, $2^{2^5}+1$ were prime twic
  would be prime.

- If all algebraic integer domains had unique factor
  tion, Kronecker would have proved the Fermat c
  jecture.

- A mathematical counterfactual is true in a partial t
  ory, 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 fal

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

- The *theory of skiing lessons* says that skiers with m lessons bend their knees when they should.

POSSIBLE WORLDDS

- Metric structures are not often as useful as Cartes structures.

- The theory of counterfactuals needs to be based incomplete structures.

APPROXIMATE OBJECTS AND THEORIES

- Counterfactuals inhabit approximate theories.

- Counterfactuals can become cartesian in suitable proximate theories.

- Article in KR-2000, also www.formal.stanford.edu/jmc/approximate.html.

- The theory of the car passing incident does not t into account what might make a car come over the h

- The simple skiing theory doesn't take into account w might make the skier bend his knees. The theory ab skiing lessons does.

VERY APPROXIMATE THEORIES ARE WHAT PEOPLE USE

- The car-crash counterfactual is complicated by be situated in a partially observable actual situation doesn't take into account the actual speeds of c coming over the hill.

CONCLUSIONS

- Some counterfactuals are useful.

- Useful counterfactuals often have non-counterfact consequences.

- Cartesian counterfactuals are the easiest.

- Counterfactuals inhabit approximate theories.

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