John Pollock's Work on Defeasible Reasoning

John Horty University of Maryland

A bit of history

1. In Philosophy:

Hart 1948

When the student has learnt that in English law there are positive conditions [for] a valid contract . . . his understanding is still incomplete . . .

He still has to learn what can *defeat* a claim that there is a valid contract . . .

This characteristic of legal concepts is one for which no word exists in ordinary English ... but the law has a word which with some hesitation I borrow and extend: this is the word "defeasible", used of a legal interest in property, which is subject to termination or "defeat"

Chisholm 1957, 1964

Toulman 1958

Gauthier 1961

Practical principles are defeasible. (I take the term from Professor H L A Hart.)

Firth, Goldman, Klein, Lehrer, Paxton, Pollock, Sosa, Swain ... in the 1960's and 1970's

No attempt at formalization ...

2. In Computer Science:

Doyle 1979 McCarthy 1980 Reiter 1980 McDermott and Doyle 1980

Reiter 1978 Minker 1983

Touretzky 1986 Horty, Thomason, Touretzky 1987 Loui 1987 Pollock 1987

Then, an explosion of research ...

3. A quote:

I believe that I developed the first formal semantics for defeasible reasoning in 1979, but I did not initially publish it because, being ignorant of AI, I did not think anyone would be interested. That semantics was finally published in 1987 (Pollock, 2007)

- 4. Pollock's work:
 - Defeasible reasoning, *Cognitive Science*, 1987
 - A theory of defeasible reasoning, *International Journal of Intelligent Sytems*, 1991
 - Self-defeating arguments, *Minds and Machines*, 1991
 - How to reason defeasibly, *Artificial Intelligence*, 1992
 - Justification and defeat, *Artificial Intelligence*, 1994
 - Cognitive Carpentry, MIT Press, 1995
 - Defeasible reasoning with variable degrees of justification, *Artificial Intelligence*, 2002
 - Defeasible reasoning, in *Reasoning: Studies of Human Inference and its Foundations*, Adler and Rips (eds), 2007
 - A recursive semantics for defeasible reasoning, in *Argumentation in Artificial Intelligence*, Rahwan and Simari (eds) 2009

Basic concepts

1. There are ordinary statements:

Penguin(Tweety) Quaker(Nixon) Republican(Nixon)

along with strict rules of inference:

 $A \land B \Rightarrow A$ $Triangular(x) \Rightarrow Trilateral(x)$ $Penguin(x) \Rightarrow Bird(x)$

But there are also defeasible (or default, or prima facie) rules:

 $Looks.red(x) \rightarrow Is.red(x)$ $Bird(x) \rightarrow Flys(x)$ $Penguin(x) \rightarrow \neg Fly(x)$ $Quaker(x) \rightarrow Pacifist(x)$ $Republican(x) \rightarrow \neg Pacifist(x)$

2. Where \mathcal{W} is a set of ordinary facts and rules, \mathcal{D} a set of defeasible rules, and < an ordering on the defeasible rules, a *prioritized default theory* is a structure of the form

$$\langle \mathcal{W}, \mathcal{D}, <
angle$$

The question is : What should we conclude from a given prioritized default theory?



3. Default theories can be represented through inference graphs

Example (Tweety Triangle):

$$\mathcal{W} = \{P, P \Rightarrow B\}$$
$$\mathcal{D} = \{r_1, r_2\}$$
$$r_1 = B \rightarrow F$$
$$r_2 = P \rightarrow \neg F$$
$$r_1 < r_2$$

(P = Penguin, B = Bird, F = Flies)



Another example (Nixon Diamond):

$$\mathcal{W} = \{Q, R\}$$
$$\mathcal{D} = \{r_1, r_2\}$$
$$r_1 = Q \to P$$
$$r_2 = R \to \neg P$$
$$< = \emptyset.$$

(Q = Quaker, R = Republican, P = Pacifist)

4. Two different graph structures:

Inference graphs

Defeat graphs

Defeat graphs depict defeat relations among arguments, so we begin with . . .

5. Arguments:

An *argument* is a sequence of tuples $\langle P_i, J_i, L_i, S_i \rangle$ such that for each *i* either

- (a) P_i is an axiom or a member of \mathcal{W} , in which case
 - i. J_i is either axiom or \mathcal{W}
 - ii. L_i is \emptyset
 - iii. S_i is ∞
- (b) P_i follows from previous P_{j_1} and P_{j_2} by MP, in which case
 - i. J_i is MP
 - ii. L_i is $\{j_1, j_2\}$
 - iii. S_i is the weaker of $\{S_{j_1}, S_{j_2}\}$
- (c) P_i follows from previous $P_{j_1} \dots P_{j_n}$ by defeasible rule r, in which case
 - i. J_i is r
 - ii. L_i is $\{j_1 \dots j_n\}$
 - iii. S_i is the weakest of $\{S_{j_1}\dots S_{j_n}\}\cup \{r\}$



6. Examples:

From the Tweety Triangle

Argument α

- 1. $\langle P, \mathcal{W}, \emptyset, \infty \rangle$
- 2. $\langle P \Rightarrow B, \mathcal{W}, \emptyset, \infty \rangle$
- **3**. $\langle B, MP, \{1, 2\}, \infty \rangle$
- 4. $(F, r_1, \{3\}, r_1)$

Argument β

1.
$$\langle P, \mathcal{W}, \emptyset, \infty \rangle$$

2. $\langle \neg F, r_2, \{1\}, r_2 \rangle$



From the Nixon Diamond

Argument α

- 1. $\langle Q, \mathcal{W}, \emptyset, \infty \rangle$
- 2. $\langle P, r_1, \{1\}, r_1 \rangle$

Argument β

1. $\langle R, \mathcal{W}, \emptyset, \infty \rangle$ 2. $\langle \neg P, r_2, \{1\}, r_2 \rangle$

7. Defeat:

Two kinds: rebutting and undercutting

- An argument line $\langle P,J,L,S\rangle$ is rebutted by an argument line $\langle P',J',L',S'\rangle$ iff
 - J is a defeasible rule, and
 - P' is $\neg P$, and
 - $\neg (S' < S)$

An argument α is rebutted by an argument β iff some line of β rebuts some line of α .

- An argument line $\langle P,J,L,S\rangle$ is undercut by an argument line $\langle P',J',L',S'\rangle$ iff
 - J is a defeasible rule r, and
 - P' is Out(r), and
 - $\neg (S' < S)$

An argument α is undercut by an argument β iff some line of β undercut some line of α .



8. Example (Drug 1):

$$\mathcal{W} = \{LR, D1\}$$

$$\mathcal{D} = \{r_1, r_2\}$$

$$r_1 = LR \rightarrow R$$

$$r_2 = D1 \rightarrow Out(r_1)$$

$$r_1 < r_2$$

(LR = Looks.red, R = Is.red, D1 = Drug 1)

Here, α is undercut by β

Argument α

- 1. $\langle LR, \mathcal{W}, \emptyset, \infty \rangle$
- 2. $\langle R, r_1, \{1\}, r_1 \rangle$

Argument β

1. $\langle D1, \mathcal{W}, \emptyset, \infty \rangle$ 2. $\langle Out(r_1), r_2, \{1\}, r_2 \rangle$ 9. An aside on undercutting defeat

Pollock 1987:

R is an undercutting defeater for P as a prima facie reason for Q iff R is a reason for denying that P wouldn't be true unless Q were true

And then:

"P wouldn't be true unless Q were true" is clearly some kind of conditional

I used to maintain that [it] was analyzable as $(\neg Q > \neg P)$, where > is the so-called "simple subjunctive"

And then later, in 1991:

"P wouldn't be true unless Q were true" is some kind of conditional, and I will symbolize it as $P \gg Q \ldots$

And still later, in 1992:

Symbolizing "It is false that P wouldn't be true unless Q were true" as $P \bigotimes Q \ldots$

10. The aside continues

On my treatment, nothing is denied

Instead, with

$$r_1 = LR \rightarrow R$$

 $r_2 = D1 \rightarrow Out(r_1)$

the default r_1 says that LR as a reason for concluding R, then r_2 says that D1 is a reason for taking this first reason out of consideration

11. More aside: Bayesian analyses?

Undercutting vs. exclusion (Raz)

• Undercutting:

The object looks red

Drug 1 makes everything look red

• Exclusion: Colin's son, private school?

The school provides good education Meet fancy friends The school is expensive Promise: only consider son's interests ...



12. Defeat (end of aside):

An argument α is defeated by an argument β iff α is rebutted or undercut by β

- 13. Defeat graph: a graph with
 - Nodes $\alpha, \beta, \gamma, \ldots$ representing arguments
 - Edges \rightsquigarrow representing defeat relations

 $\alpha \rightsquigarrow \beta$ means: α defeats β

Pollock's 1987 theory

1. Preliminary version: given a defeat graph, define

Status at a level:

- All arguments are *in* at level 0
- An argument is *in* at level n + 1 iff it is not defeated by any argument *in* at level n

Justification:

• An argument α is *justified* iff

 $\exists m \forall n \geq m(\alpha \text{ is } in \text{ at level } n)$



2. Example (Drugs 1 and 2):

${\mathcal W}$	=	${D1, D2, LR}$
${\cal D}$	=	$\{r_1, r_2, r_3\}$
r_1	=	$LR \rightarrow R$
r_2	=	$D1 \rightarrow Out(r_1)$
r_3	=	$D2 \rightarrow Out(r_2)$
$r_1 < r_2 < r_3$		

(LR = Looks.red, R = Is.red, D1 = Drug1, D2 = Drug2)



3. Another example (Nixon again)



4. Yet another example (zombie arguments):

$$\mathcal{W} = \{Q, R\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4\}$$

$$r_1 = Q \rightarrow P$$

$$r_2 = R \rightarrow \neg P$$

$$r_3 = P \rightarrow S$$

$$r_4 = \top \rightarrow \neg S$$

$$< = \emptyset.$$



5. Refining the preliminary version: consider

$$\mathcal{W} = \{Q, R, B\}$$
$$\mathcal{D} = \{r_1, r_2\}$$
$$r_1 = Q \to P$$
$$r_2 = R \to \neg P$$
$$r_3 = B \to F$$
$$< = \emptyset.$$

Now have:

1.
$$\langle Q, W, \emptyset, S \rangle$$

2. $\langle P, r_1, \{1\}, S \rangle$
3. $\langle R, W, \emptyset, S \rangle$
4. $\langle \neg P, r_2, \{3\}, S \rangle$
5. $\langle \neg F, logic, \{2, 4\}, S \rangle$

Problem: self-defeating arguments

Pollock's initial solution: simply remove them all!

Abstract argumentation

- 1. Due to Phan Minh Dung, 1995
- 2. Some simple definitions:
 - We've seen $\alpha \rightsquigarrow \beta$
 - $\Gamma \rightsquigarrow \beta$ means: $\exists \alpha (\alpha \in \Gamma \land \alpha \rightsquigarrow \beta)$
 - $\Gamma \rightsquigarrow \Delta$ means: $\exists \alpha \exists \beta (\alpha \in \Gamma \land \beta \in \Delta \land \alpha \rightsquigarrow \beta)$
 - Γ is consistent means: $\neg(\Gamma \rightsquigarrow \Gamma)$
- 3. Defense:
 - Γ defends α means: $\forall \beta (\beta \rightsquigarrow \alpha \Rightarrow \Gamma \rightsquigarrow \beta)$
 - $F(\Gamma) = \{ \alpha : \Gamma \text{ defends } \alpha \}$
 - $\Gamma \subseteq \Delta \Rightarrow F(\Gamma) \subseteq F(\Delta)$
- 4. Admissible sets:
 - Γ admissible iff: Γ consistent and $\Gamma \subseteq F(\Gamma)$
 - Γ complete iff: Γ consistent and $F(\Gamma) = \Gamma$

complete preferred T stable

5. Various solution concepts:

An admissible set Γ is:

- Grounded iff: Γ is a minimal complete set
- Preferred iff: Γ is a maximal complete set
- Stable iff: $\alpha \not\in \Gamma \Rightarrow \Gamma \rightsquigarrow \alpha$

Note 1: Stable may not exist, others do

Note 2: Only the grounded solution is unique

Note 3: If everything is finite, let

$$\begin{array}{rcl} \Gamma_0 & = & \emptyset \\ \Gamma_{i+1} & = & F(\Gamma_i) \end{array}$$

Then grounded solution is

$$\overline{\Gamma} = \bigcup_n \Gamma_n$$



6. Exercise: calculate all complete solutions for each of these graphs. What complete solutions are grounded, stable, preferred?

$$\left(\begin{array}{c} x \\ y \\ y \end{array}\right) = \left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} x \\ y \\ y \end{array}\right) = \left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{$$

7. Alternative approach #1: labelingsA complete labeling is a total function

 $L: Arguments \rightarrow \{in, out, u\}$

subject to conditions

•
$$L(\alpha) = out$$
 iff $\exists \beta (\beta \rightsquigarrow \alpha \land L(\beta) = in)$

- $L(\alpha) = in \text{ iff } \forall \beta(\beta \rightsquigarrow \alpha \Rightarrow L(\beta) = out)$
- 8. A complete labeling is:
 - Grounded iff: maximal undecided
 - Preferred iff: maximal in (maximal out)
 - Stable iff: no undecided
- 9. Fact: If a complete labeling L is X, then the set {α : L(α) = in}
 is X, for X = Grounded, Preferred, Stable

10. Alternative approach #2: dispute games

Example: Grounded dispute game

- Proponent states an argument
- Proponent and opponent then take turns:
 - Each new argument must defeat previous
 - Proponent cannot repeat argument
 - Proponent arguments must strictly defeat
- Proponent wins if opponent cannot move

Fact: An argument is in grounded solution iff proponent always wins grounded dispute game starting with that argument



11. Example: Extensive form of grounded dispute game



12. Alternative solution concepts

Pollock in abstract argumentation

- 1. The 1987 theory, again:
 - All arguments are *in* at level 0
 - An argument is *in* at level n + 1 iff it is not defeated by any argument *in* at level n
 - An argument α is *justified* iff

 $\exists m \forall n \geq m(\alpha \text{ is } in \text{ at level } n)$

2. A slight reformulation:

Define

$$G(\Gamma) = \{ \alpha : \neg(\Gamma \rightsquigarrow \alpha) \}$$

Where Δ is entire set of arguments, let

$$\Delta_0 = \Delta$$
$$\Delta_{i+1} = G(\Delta_i)$$

Then define

 α justified iff $\exists m \forall n \geq m (\alpha \in \Delta_n)$

3. Fact:

An argument is justified iff it belongs to the grounded solution

Recall:

$$F(\Gamma) = \{\alpha : \Gamma \text{ defends } \alpha\}$$

and then

$$\begin{aligned}
 \Gamma_0 &= \emptyset \\
 \Gamma_{i+1} &= F(\Gamma_i)
 \end{aligned}$$

and

 $\overline{\Gamma} = \bigcup_i \Gamma_i$

is the grounded solution

So what fact says is that

$$lpha\in\overline{\mathsf{\Gamma}}$$
 iff $\exists morall n\geq m(lpha\in \Delta_n)$

Verification depends on

(1) $G(\Delta) = F(\emptyset), \quad \Delta \text{ all arguments}$

(2)
$$F(\Gamma) = G(G(\Gamma))$$

$$(3) \quad \Gamma = F(\Gamma) \quad \Rightarrow \quad \Gamma \subseteq G(\Gamma)$$

Now, why move to 1994/95?



1. Example (Unreliable John):

$$\mathcal{W} = \{JA(U(j))\} \\ \mathcal{D} = \{r_1, r_2\} \\ r_1 = JA(U(j)) \rightarrow U(j) \\ r_2 = U(j) \rightarrow Out(r_1) \\ < = \emptyset.$$

(JA(X) = John asserts X, U(y) = y is unreliable,j = John)

Now have:

1.
$$\langle JA(U(j)), W, \emptyset, S \rangle$$

2. $\langle U(j), r_1, \{1\}, S \rangle$
3. $\langle Out(r_1), r_2, \{2\}, S \rangle$

So this is a kind of self-defeat we can't rule out

2. Pollock writes:

In earlier publications, I proposed that defeat could be analyzed as defeat among arguments, rather than inference nodes I see no way to recast the present analysis in terms of a defeat relation between arguments, as opposed to nodes, which are argument steps rather than complete arguments. (1994, p 393)

3. But not so

Fact: The 1994/1995 theory is Dung's preferred solution



- 4. Multiple preferred solutions
- 5. Argument classification:

"Credulous"

 An argument is *defensible* iff it belongs to some preferred solution – ie, iff it gets *in* in some preferred labeling

"Skeptical"

• An argument is *justified* iff it belongs to every preferred solution – ie, iff it gets *in* in every preferred labeling

What motivates further change, to 2009?



1. We approach in stages ...

Stage A (John asserts P):

$$\mathcal{W} = \{JA(U(j)), JA(P)\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4\}$$

$$r_1 = JA(U(j)) \rightarrow U(j)$$

$$r_2 = U(j) \rightarrow Out(r_1)$$

$$r_3 = JA(P) \rightarrow P$$

$$r_4 = U(j) \rightarrow Out(r_3)$$

$$< = \emptyset.$$

(JA(X) = John asserts X, U(y) = y is unreliable,j = John)

2. Stage B (Adding Susan):

$$\mathcal{W} = \{JA(U(s)), SA(U(j)), JA(P)\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4, r_5, r_6\}$$

$$r_1 = JA(U(j)) \rightarrow U(s)$$

$$r_2 = U(j) \rightarrow Out(r_1)$$

$$r_3 = JA(P) \rightarrow P$$

$$r_4 = U(j) \rightarrow Out(r_3)$$

$$r_5 = SA(U(j)) \rightarrow U(j)$$

$$r_6 = U(s) \rightarrow Out(r_5)$$

$$< = \emptyset.$$

(JA(X) = John asserts X, SA(X) = Susan asserts X, U(y) = y is unreliable, j = John, s = Susan)

A quote:

We get the right answer, but we get it in a different way than before. This difference has always bothered me (Pollock, 2002)



3. Stage A1 (John and Donald):

$$\mathcal{W} = \{JA(U(j)), JA(P)\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4, r_7\}$$

$$r_1 = JA(U(j)) \rightarrow U(j)$$

$$r_2 = U(j) \rightarrow Out(r_1)$$

$$r_3 = JA(P) \rightarrow P$$

$$r_4 = U(j) \rightarrow Out(r_3)$$

$$r_7 = DA(\neg P) \rightarrow \neg P$$

$$< = \emptyset.$$

(JA(X) = John asserts X, DA(X) = Donald asserts X, U(y) = y is unreliable <math>j = John)

$$U(s) \qquad U(j) \qquad P \not = 7 \qquad P = 7 \qquad$$

4. Stage B1 (John, Susan, and Donald):

$$\mathcal{W} = \{JA(U(s)), SA(U(j)), JA(P)\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4, r_5, r_6, r_7\}$$

$$r_1 = JA(U(j)) \rightarrow U(s)$$

$$r_2 = U(j) \rightarrow Out(r_1)$$

$$r_3 = JA(P) \rightarrow P$$

$$r_4 = U(j) \rightarrow Out(r_3)$$

$$r_5 = SA(U(j)) \rightarrow U(j)$$

$$r_6 = U(s) \rightarrow Out(r_5)$$

$$r_7 = DA(\neg P) \rightarrow \neg P$$

$$< = \emptyset.$$

(JA(X) = John asserts X, SA(X) = Susan asserts X, DA(X) = Donald asserts X, U(y) = y is unreliable j = John, s = Susan)



5. So, the 2009 theory is motivated by conflicting behaviors of odd and even defeat loops:

This, I take it, is a problem. Although it might not be clear which inference-graph is producing the right answer, the right answer ought to be the same in both inference graphs. Thus the semantics is getting one of them wrong (Pollock 2009)

6. Misplaced concern. Odd and even cycles are just different: odd cycles are paradoxes, even cycles are "pathological" but not paradoxical

This holds in many areas . . .

Odd cycle:

John: What I am now saying is false

Even cycle:

John: What Susan is now saying is false Susan: What John is now saying is false

Odd cycle:

John: What Susan is now saying is false Susan: What Jason is now saying is false Jason: What John is now saying is false

Even cycle:

John: What Susan is now saying is false Susan: What Jason is now saying is false Jason: What Sara is now saying is false Sara: What John is now saying is false

Etc



Some open issues

1. Problem #1: Defeasible reasoning spans different domains (epistemology, ethics, law), but Pollock concentrates only on epistemology

As a result:

- Total order on default rules
- "Weakest link" ordering on arguments
- 2. Weakest link trouble:

$$W = \emptyset$$

$$D = \{r_1, r_2, r_3\}$$

$$Captain (r_1) = \top \to H$$

$$Major (r_2) = \top \to \neg W$$

$$Colonel (r_3) = H \to W$$

$$r_1 < r_2 < r_3$$

(H = Heater on, W = Window open)



3. Weakest link trouble (epistemic)?

$$\mathcal{W} = \{V\}$$

$$\mathcal{D} = \{r_1, r_2, r_3\}$$

$$r_1 = V \rightarrow E$$

$$r_2 = E \rightarrow \neg H$$

$$r_3 = E \rightarrow H$$

$$r_1 < r_2 < r_3$$

(E = Visual, E = Egg, H = Healthy)



4. Problem #2: If priorities are decided as arguments are evaluated, how can the defeat graph be constructed prior to argument evaluation?

$$\mathcal{W} = \{Q, R, SA(r_2 < r_1), JA(r_1 < r_2)\}$$

$$\mathcal{D} = \{r_1, r_2, r_3, r_4\}$$

$$r_1 = Q \to P$$

$$r_2 = R \to \neg P$$

$$r_3 = SA(r_2 < r_1) \to r_2 < r_1$$

$$r_4 = JA(r_1 < r_2) \to r_1 < r_2$$

$$< = \emptyset \text{ (initially)}$$



4. An idea: relativize defeat relation \rightsquigarrow to an argument set $\Gamma,$ so

 $\alpha \leadsto_{\mathsf{F}} \beta$

Conclusions

- 1. It helps to look at Pollock's work on defeasibility through the lens of abstract argumentation
- 2. The 1987 theory is sensible
- 3. The 1994/1995 theory is also sensible
- 4. I'm not sure about the 2009 theory
- 5. The Pollock/Dung approach may have advantages over Reiter's original default logic (without priorities)
- But the advantages of prioritized default logic have not yet been incorporated into the Pollock/Dung approach
- 7. Many open issues in this area, and a lot of interesting work to do

Further conclusions

Homework Exercise – by Rachel Briggs

Jane says that everybody knows Richard is a liar.

Richard categorically denies the rumors. Clara, on the other hand, confirms the rumors,

but points out that she doesnt *know* them she only justifiably believes them.

Aurelien counters: you may not have heard Clara correctly.

Are you sure she didnt say *juggles beehives*? Jane says Clara wouldnt know a beehive from her own behind.

No, Aurelien is sure that Clara juggles dangerous objects of some sort

although it may have been torches. True, her mother told her never to play with fire

but her father said that if she was going to juggle it might as well be torches, why not burn the house down,

and Richard asked her to juggle the other day and Clara is the obliging type.

Bob says you can't trust Bob but everybody knows Bob is unreliable.