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”

Gauthier 1961

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

Chisholm 1957, 1964

Toulmin 1958

Gettier, Goldman, Klein, Lehrer, Paxton, Swain . . .

2. In Artificial Intelligence:

Doyle 1979
McCarthy 1980
Reiter 1980
McDermott and Doyle 1980

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 1986” (Pollock, 2007)
4. Pollock’s work:

- How to reason defeasibly, *Artificial Intelligence*, 1992
- Justification and defeat, *Artificial Intelligence*, 1994
- Defeasible reasoning with variable degrees of justification, *Artificial Intelligence*, 2002
- A recursive semantics for defeasible reasoning, in *Argumentation in Artificial Intelligence*, Rahwan and Simari (eds) 2009
Basic concepts

1. There are ordinary statements:

   \[
   \begin{align*}
   &Penguin(Tweety) \\
   &Quaker(Nixon) \\
   &Republican(Nixon)
   \end{align*}
   \]

   along with strict rules of inference:

   \[
   \begin{align*}
   &A \land B \Rightarrow A \\
   &Triangular(x) \Rightarrow Trilateral(x) \\
   &Penguin(x) \Rightarrow Bird(x)
   \end{align*}
   \]

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

   \[
   \begin{align*}
   &Looks.red(x) \rightarrow Is.red(x) \\
   &Bird(x) \rightarrow Flws(x) \\
   &Penguin(x) \rightarrow \neg Fly(x) \\
   &Quaker(x) \rightarrow Pacifist(x) \\
   &Republican(x) \rightarrow \neg Pacifist(x)
   \end{align*}
   \]

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 default theory is a structure of the form

\[
\langle \mathcal{W}, \mathcal{D}, < \rangle
\]

The question is: What should we conclude from a given default theory?
3. Default theories can be represented through inference graphs

Example (Tweety Triangle):

\[
\begin{align*}
W &= \{P, P \Rightarrow B\} \\
D &= \{r_1, r_2\} \\
r_1 &= B \rightarrow F \\
r_2 &= P \rightarrow \neg F \\
r_1 &< r_2
\end{align*}
\]

\((P = \text{Penguin}, \ B = \text{Bird}, \ F = \text{Flies})\)
Another example (Nixon Diamond):

\[
\begin{align*}
\mathcal{W} & = \{Q, R\} \\
\mathcal{D} & = \{r_1, r_2\} \\
r_1 & = Q \rightarrow P \\
r_2 & = R \rightarrow \neg P \\
< & = \emptyset.
\end{align*}
\]

\((Q = \text{Quaker}, R = \text{Republican}, P = \text{Pacifist})\)
4. Two different 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 \( \infty \)

(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 \( P_{j_1} \ldots P_{j_n} \) by defeasible rule \( r \), in which case
   i. \( J_i \) is \( r \)
   ii. \( L_i \) is \( \{j_1 \ldots j_n\} \)
   iii. \( S_i \) is the weakest of \( \{S_{j_1} \ldots S_{j_n}\} \cup \{r\} \)
6. Examples:

From the Tweety Triangle

Argument $\alpha$

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

Argument $\beta$

1. $\langle P, W, \emptyset, \infty \rangle$
2. $\langle \neg F, r_2, \{1\}, r_2 \rangle$
From the Nixon Diamond

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

Argument $\beta$
1. $\langle R, W, \emptyset, \infty \rangle$
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 $\alpha$ is rebutted by an argument $\beta$ iff some line of $\beta$ rebuts some line of $\alpha$.

- 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 $\text{Out}(r)$, and
  - $\neg(S' < S)$

An argument $\alpha$ is undercut by an argument $\beta$ iff some line of $\beta$ undercut some line of $\alpha$. 
8. Example (Drug 1):

\[ W = \{LR, D1\} \]
\[ D = \{r_1, r_2\} \]
\[ r_1 = LR \rightarrow R \]
\[ r_2 = D1 \rightarrow Out(r_1) \]
\[ r_1 < r_2 \]

\((LR = \text{Looks.red}, R = \text{Is.red}, D1 = \text{Drug 1})\)

Here, \(\alpha\) is undercut by \(\beta\)

Argument \(\alpha\)

1. \(\langle LR, W, \emptyset, \infty \rangle\)
2. \(\langle R, r_1, \{1\}, r_1 \rangle\)

Argument \(\beta\)

1. \(\langle D1, W, \emptyset, \infty \rangle\)
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 (¬Q > ¬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 ≫ Q ....

And still later, in 1992:

Symbolizing “It is false that P wouldn’t be true unless Q were true” as P ⊕ Q ....
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
  
  He’ll meet fancy friends
  
  The school is expensive
  
  Promise: only consider son’s interests …
12. Defeat (end of aside):

An argument $\alpha$ is defeated by an argument $\beta$ iff $\alpha$ is rebutted or undercut by $\beta$

13. Defeat graph: a graph with

- Nodes $\alpha, \beta, \gamma, \ldots$ representing arguments
- Edges $\sim$ representing defeat relations

$\alpha \sim \beta$ means: $\alpha$ defeats $\beta$
Pollock’s 1987 theory

1. Preliminary version: given a defeat graph, define

Status at a level:

- All arguments are \textit{in} at level 0

- An argument is \textit{in} at level \( n + 1 \) iff it is not defeated by any argument \textit{in} at level \( n \)

Justification:

- An argument \( \alpha \) is \textit{justified} iff

\[ \exists m \forall n \geq m (\alpha \text{ is } \textit{in} \text{ at level } n) \]
2. Example (Drugs 1 and 2):

\[ W = \{ D_1, D_2, LR \} \]
\[ D = \{ r_1, r_2, r_3 \} \]
\[ r_1 = LR \rightarrow R \]
\[ r_2 = D_1 \rightarrow Out(r_1) \]
\[ r_3 = D_2 \rightarrow Out(r_2) \]
\[ r_1 < r_2 < r_3 \]

(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

\[ W = \{Q, R, B\} \]
\[ D = \{r_1, r_2\} \]
\[ r_1 = Q \rightarrow P \]
\[ r_2 = R \rightarrow \neg P \]
\[ r_3 = B \rightarrow F \]
\[ < = \emptyset. \]

Now have:

1. \( \langle Q, \text{fact}, \emptyset, S \rangle \)
2. \( \langle P, r_1, \{1\}, S \rangle \)
3. \( \langle R, \text{fact}, \emptyset, S \rangle \)
4. \( \langle \neg P, r_2, \{3\}, S \rangle \)
5. \( \langle \neg F, \text{logic}, \{2, 4\}, S \rangle \)

Problem: self-defeating arguments
Pollock’s initial solution: simply remove them all!
Admissibility semantics

1. Due to Phan Minh Dung, 1995

2. Some simple definitions:
   - We’ve seen $\alpha \leadsto \beta$
   - $\Gamma \leadsto \beta$ means: $\exists \alpha (\alpha \in \Gamma \land \alpha \leadsto \beta)$
   - $\Gamma \leadsto \Delta$ means: $\exists \alpha \exists \beta (\alpha \in \Gamma \land \beta \in \Delta \land \alpha \leadsto \beta)$
   - $\Gamma$ is consistent means: $\neg (\Gamma \leadsto \Gamma)$

3. Defense:
   - $\Gamma$ defends $\alpha$ means: $\forall \beta (\beta \leadsto \alpha \Rightarrow \Gamma \leadsto \beta)$
   - $F(\Gamma) = \{ \alpha : \Gamma$ defends $\alpha \}$
   - $\Gamma \subseteq \Delta \Rightarrow F(\Gamma) \subseteq F(\Delta)$

4. Admissible sets:
   - $\Gamma$ admissible iff: $\Gamma$ consistent and $\Gamma \subseteq F(\Gamma)$
   - $\Gamma$ complete iff: $\Gamma$ consistent and $F(\Gamma) = \Gamma$
5. Various extension concepts:

An admissible set $\Gamma$ is:

- Grounded iff: $\Gamma$ is a minimal complete set
- Preferred iff: $\Gamma$ is a maximal complete set
- Stable iff: $\alpha \notin \Gamma \Rightarrow \Gamma \sim \alpha$

Note 1: Stable may not exist, others do

Note 2: Only the grounded extension is unique

Note 3: If everything is finite, let

\[
\begin{align*}
\Gamma_0 &= \emptyset \\
\Gamma_{i+1} &= F(\Gamma_i)
\end{align*}
\]

Then grounded extension is

\[
\bar{\Gamma} = \bigcup_n \Gamma_n
\]
6. A complete labeling is a total function

\[ L : \text{Arguments} \rightarrow \{ \text{in}, \text{out}, \text{u} \} \]

subject to conditions

- \( L(\alpha) = \text{out} \) iff \( \exists \beta (\beta \sim \alpha \land L(\beta) = \text{in}) \)
- \( L(\alpha) = \text{in} \) iff \( \forall \beta (\beta \sim \alpha \Rightarrow L(\beta) = \text{out}) \)

7. A labeling is:

- Grounded iff: maximal undecided
- Preferred iff: maximal in (maximal out)
- Stable iff: no undecided

8. Fact: If a labeling \( L \) is \( X \), then the set

\[ \{ \alpha : L(\alpha) = \text{in} \} \]

is \( X \), for \( X = \text{Grounded, Preferred, Stable} \)
9. Exercise: calculate all complete extensions for each of these graphs. What complete extensions are grounded, stable, preferred?
Pollock in admissibility semantics

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 $\alpha$ is *justified* iff

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

2. A slight reformulation:

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

   Where $\Delta$ is entire set of arguments, let

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

   Then define

   $$\alpha \text{ justified iff } \exists m \forall n \geq m (\alpha \in \Delta_n)$$
3. Fact:

An argument is justified iff it belongs to the grounded extension

Recall:

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

and then

\[ \Gamma_0 = \emptyset \]
\[ \Gamma_{i+1} = F(\Gamma_i) \]

and

\[ \Gamma = \bigcup_i \Gamma_i \]

is the grounded extension

So what fact says is that

\[ \alpha \in \Gamma \text{ iff } \exists m \forall n \geq m (\alpha \in \Delta_n) \]

Verification depends on

\[ (1) \quad F(\Gamma) = G(G(\Gamma)) \]
\[ (2) \quad G(\Delta) = F(\emptyset), \]

where \( \Delta \) is the entire set of arguments
1. Example (Unreliable John):

\[
\begin{align*}
\mathcal{W} & = \{JA(U(j))\} \\
\mathcal{D} & = \{r_1, r_2\} \\
r_1 & = JA(U(j)) \rightarrow U(j) \\
r_2 & = U(j) \rightarrow Out(r_1) \\
< & = \emptyset.
\end{align*}
\]

\((JA(X) = \text{John asserts } X, \ U(y) = y \text{ is unreliable, } j = \text{John})\)

Now have:

1. \(\langle JA(U(j)), \text{fact}, \emptyset, S \rangle\)
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:

The 1994/1995 theory is Dung's preferred semantics
4. Multiple preferred extensions

5. Argument classification:

“Credulous”

• Each preferred extension constitutes an acceptable set of arguments

“Skeptical”

• An argument is justified iff it belongs to each preferred extension
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):

\[ \mathcal{W} = \{ \text{JA}(U(s)), \text{SA}(U(j)), \text{JA}(P) \} \]
\[ \mathcal{D} = \{ r_1, r_2, r_3, r_4, r_5, r_6 \} \]
\[ r_1 = \text{JA}(U(j)) \rightarrow U(s) \]
\[ r_2 = U(j) \rightarrow \text{Out}(r_1) \]
\[ r_3 = \text{JA}(P) \rightarrow P \]
\[ r_4 = U(j) \rightarrow \text{Out}(r_3) \]
\[ r_5 = \text{SA}(U(j)) \rightarrow U(j) \]
\[ r_6 = U(s) \rightarrow \text{Out}(r_5) \]
\[ < = \emptyset. \]

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

A quote:

We get the right answer, but we get it in a different way than we do for even-length defeat cycles ... This difference has always bothered me (Pollock, 2002)
3. Stage A1 (John and Donald):

\[ \mathcal{W} = \{JA(U(j)), JA(P)\} \]
\[ D = \{r_1, r_2, r_3, r_4, r_7\} \]
\[ r_1 = JA(U(j)) \rightarrow U(j) \]
\[ r_2 = U(j) \rightarrow Out(r_1) \]
\[ r_3 = JA(P) \rightarrow P \]
\[ r_4 = U(j) \rightarrow Out(r_3) \]
\[ r_7 = DA(\neg P) \rightarrow \neg P \]
\[ < = \emptyset. \]

\((JA(X) = \text{John asserts } X, DA(X) = \text{Donald asserts } X, U(y) = y \text{ is unreliable } j = \text{John})\)
4. Stage B1 (John, Susan, and Donald):

\[ \begin{align*}
\mathcal{W} & = \{ JA(U(s)), SA(U(j)), JA(P) \} \\
\mathcal{D} & = \{ r_1, r_2, r_3, r_4, r_5, r_6, r_7 \} \\
r_1 & = JA(U(j)) \rightarrow U(s) \\
r_2 & = U(j) \rightarrow Out(r_1) \\
r_3 & = JA(P) \rightarrow P \\
r_4 & = U(j) \rightarrow Out(r_3) \\
r_5 & = SA(U(j)) \rightarrow U(j) \\
r_6 & = U(s) \rightarrow Out(r_5) \\
r_7 & = DA(\neg P) \rightarrow \neg P \\
< & = \emptyset.
\end{align*} \]

\((JA(X) = \text{John asserts } X, SA(X) = \text{Susan asserts } X, DA(X) = \text{Donald asserts } X, U(y) = y \text{ is unreliable } j = \text{John, } s = \text{Susan})\)
5. So, the 2009 semantics is motivated by conflicting behaviors of odd and even defeat loops

Another quote:

“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. I feel that this is a misplaced concern. Odd and even cycles are just different: odd cycles are paradoxes, even cycles are dilemmas.

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. Although defeasible reasoning spans normative and non-normative domains (ethics, law, epistemology), Pollock concentrates only on non-normative reasoning.

As a result:

- Total order on default rules
- “Weakest-link” ordering on arguments

2. Generalize to:

- Strict partial ordering on default rules
- What is the appropriate argument ordering??
3. The problem:

- Lift partial order from defaults to sets of defaults
- Lift partial order from defaults to sequences of defaults

in a plausible way ...
4. Reasoning about priorities.

Problem: if priorities are decided as arguments are evaluated, how can the defeat graph be constructed prior to argument evaluation?

An idea: relativize defeat relation $\sim$ to an argument set $\Gamma$, so

$$\alpha \sim_\Gamma \beta$$
5. For Pollock: arguments all “present” at once, and then evaluated together

I prefer: interleaving argument construction and evaluation
6. A general issue: reason accrual

Consider: I am invited to a wedding and ...  

The wedding falls at an inconvenient time  
Aunt Olive will be there  
Aunt Petunia will be there  
Aunt Olive and Aunt Petunia will be there  

Is the fourth reason really independent of the second and third?
7. Undercutting by weaker defaults

Take $\langle \mathcal{W}, \mathcal{D} \rangle$ with

$\mathcal{W} = \{A, B\}$
$\mathcal{D} = \{r_1, r_2\}$
$r_1 = A \rightarrow \text{Out}(r_2)$
$r_2 = B \rightarrow P$
$r_1 < r_2$

Can $r_2$ be undercut by $r_1$?

Pollock: No

Me: hmmm ...
Conclusions

1. It helps to look at Pollock’s work on defeasibility through the lens of admissibility semantics

2. The 1987 theory is sensible

3. The 1994/1995 theory is also sensible

4. I’m not sure about the 2009 theory

5. 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 doesn’t *know* them; she only justifiably believes them.

Aurelien counters:
you may not have heard Clara correctly.

Are you sure she didn’t say *juggles beehives*?
Jane says Clara wouldn’t 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.