John Pollock’s Work on Defeasible Reasoning

John Horty

University of Maryland
(and currently, Stanford CASBS)
A bit of history

1. In Philosophy:

Hart 1948

“It is not usually possible to define a legal concept such as ... “contract” by specifying the necessary and sufficient conditions for its application ... such concepts can only be explained with the aid of a list of exceptions or negative examples ...”

Chisholm 1957, 1964

“These questions concern the defeasibility of moral requirements.”

(Loui, interview: Chisholm says that his colleague, John Ladd, “who was quite taken by Hart,” had used the term)

Toulmin 1958

Gauthier 1961

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

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


Rescher 1977
2. In Artificial Intelligence:

Doyle 1979
McCarthy 1980
Reiter 1980
McDermott and Doyle 1980

Touretzky 1986
Horty, Thomason, Touretzky 1987
Loui 1987
Pollock 1987
Nute 1988

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
Pollock’s theory: constant features

1. There are ordinary statements:

   \(\text{Penguin}(Tweety)\)
   \(\text{Quaker}(Nixon)\)
   \(\text{Republican}(Nixon)\)

   along with strict rules of inference:

   \(A \land B \Rightarrow A\)
   \(\text{Triangular}(x) \Rightarrow \text{Trilateral}(x)\)
   \(\text{Penguin}(x) \Rightarrow \text{Bird}(x)\)

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

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

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, we can refer to the structure \(\langle \mathcal{W}, \mathcal{D}, < \rangle\) as a default theory.

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

Example (Nixon Diamond):

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

\((Q = \text{Quaker}, \ R = \text{Republican}, \ P = \text{Pacifist})\)
Another example (Tweety Triangle):

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

\((P = \text{Penguin}, B = \text{Bird}, F = \text{Flies})\)
4. Two different structures:

   Inference graphs

   Defeat graphs

We've just seen some inference 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 Nixon Diamond

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

Argument $\beta$
1. $\langle R, \mathcal{W}, \emptyset, \infty \rangle$
2. $\langle \neg P, r_2, \{1\}, r_2 \rangle$
From the Tweety Triangle

Argument $\alpha$

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. $\langle F, r_1, \{3\}, r_1 \rangle$

Argument $\beta$

1. $\langle P, \mathcal{W}, \emptyset, \infty \rangle$
2. $\langle \neg F, r_2, \{3\}, r_2 \rangle$
7. Rebutting defeat:

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 $\alpha$.

8. Undercutting defeat:

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 $\alpha$ is undercut by an argument $\beta$ iff some line of $\beta$ undercut some line of $\alpha$. 
9. 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 = Looks.red, R = Is.red, D1 = 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. \( \langle Out(r_1), r_2, \{1\}, r_2 \rangle \)
10. Defeat:

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

11. Defeat graph: a graph with

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

$\alpha \sim \beta$ means: $\alpha$ defeats $\beta$
Admissibility semantics

1. Due to Phan Minh Dung, 1995

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

3. Defense:
   - $\Gamma$ defends $\alpha$ means: $\forall \beta (\beta \sim \alpha \Rightarrow \Gamma \sim \beta)$
   - $F(\Gamma) = \{ \alpha : \Gamma$ defends $\alpha \}$
     Note: $F$ is monotonic
     $\Gamma \subseteq \Gamma' \Rightarrow F(\Gamma) \subseteq F(\Gamma')$

4. Admissible sets:
   - $\Gamma$ admissible means: consistent and $F(\Gamma) \subseteq \Gamma$
     (ie, $\Gamma$ defends all of its own members)
5. 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, \( D_1 = \) Drug1, \( D_2 = \) Drug2 )
6. Various extension concepts:

An admissible set $\Gamma$ is:

- Preferred iff: $\Gamma$ is a maximal admissible set
- Complete iff: $F(\Gamma) = \Gamma$
- Grounded iff: the minimal complete set
- Stable iff: $\alpha \notin \Gamma \Rightarrow \Gamma \not\sim \alpha$

Note 1: Only the grounded extension is unique

Note 2: If everything is finite, let

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

Then grounded extension can be calculated as

$$\Gamma = \bigcup_n \Gamma_n$$
7. Status assignment: assigns to zero or more arguments either \textit{in} or \textit{out}, subject to conditions:

- An argument is \textit{in} iff all defeaters are \textit{out},
- An argument is \textit{out} iff some defeater is \textit{in}

8. Correlation with extension concepts:

- Complete: any status assignment
- Grounded: maximal undecided
- Preferred: maximal \textit{in} (maximal \textit{out})
- Stable: every argument \textit{in} or \textit{out}
9. Exercise: calculate all complete extensions for each of these graphs. What complete extensions are grounded, stable, preferred?
Pollock’s semantics

1. The 1987 theory, preliminary version:

   • 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 is justified (or: ultimately undefeated) iff there is some \( m \) such that, for every \( n \geq m \), the argument is in at level \( n \)

2. A slight reformulation:

Define

\[
G(\Gamma) = \{ \alpha : \neg (\Gamma \leadsto \alpha) \}
\]

Where \( \Delta \) is entire set of arguments, let

\[
\Delta_0 = \Delta \\
\Delta_i+1 = G(\Delta_i)
\]

Then define

\( \alpha \) justified iff \( \exists m \forall n \geq m (\alpha \in \Delta_n) \)
3. Example (Drugs 1 and 2, again):
4. Fact (Dung, 1995):

An argument is justified iff it belongs to the grounded extension

Recall:

The grounded extension is the minimal fixed point of $F$, where

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

$$= \{\alpha : \forall \beta (\beta \leadsto \alpha \Rightarrow \Gamma \leadsto \beta)\}$$

Compare

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

$$= \{\alpha : \forall \beta (\beta \in \Gamma \Rightarrow \neg(\beta \leadsto \alpha))\}$$

and note that

1. $F(\Gamma) = G(G(\Gamma))$
2. $G(\Delta) = F(\emptyset)$,

where $\Delta$ is the entire set of arguments
5. Example (zombie arguments):

\[ \mathcal{W} = \{Q, R\} \]
\[ \mathcal{D} = \{r_1, r_2, r_3, r_4\} \]
\[ r_1 = Q \rightarrow P \]
\[ r_2 = R \rightarrow \neg P \]
\[ r_3 = P \rightarrow S \]
\[ r_4 = \top \rightarrow \neg S \]
\[ < = \emptyset. \]
6. 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. \( \langle \neg F, \text{logic}, \{2, 4\}, S \rangle \)

Argument self-defeating iff one step defeats another. Pollock’s initial solution: simply rule out all self-defeating arguments. This is 1987.
7. Now, what’s wrong with 1987?

Example (Unreliable John):

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

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

Now have:

1. \( \langle JA(U(j)), fact, \emptyset, S \rangle \)
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
8. The 1994/1995 theory:

A mapping $\sigma$ of nodes from an inference graph to $in$ and $out$ is a partial status assignment iff

- $\sigma$ assigns $in$ to each initial node;
- $\sigma$ assigns $in$ to a non-initial node iff it assigns $in$ to all immediate ancestors and $out$ to all defeaters;
- $\sigma$ assigns $out$ to a non-initial node iff it assigns $out$ to an immediate ancestor or $in$ to a defeater.

A status assignment is a maximal partial status assignment

A node is: justified iff $in$ in all status assignments; overruled iff $out$ or undecided in all status assignments; defensible iff $in$ in some but not all status assignments
9. Pollock writes:

“In earlier publications, I proposed that defeat could be analyzed as defeat among arguments, rather than inference nodes ... I now feel that obscured the proper treatment of self-defeat. 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” (Pollock 1994, p 393)

10. But there is a way (Jakobovits, 1999):

Pollock’s 1994/1995 analysis for inference graphs is equivalent to Dung’s preferred semantics for defeat graphs
11. Now what motivates the further change?
We approach in stages …

Stage A (John asserts $P$):

$$\mathcal{W} = \{JA(U(j)), JA(P)\}$$
$$\mathcal{D} = \{r_1, r_2, r_3, r_4\}$$
$$r_1 = JA(U(j)) \rightarrow U(j)$$
$$r_2 = U(j) \rightarrow Out(r_1)$$
$$r_3 = JA(P) \rightarrow P$$
$$r_4 = U(j) \rightarrow Out(r_3)$$
$$\langle = \emptyset.$$

($JA(X) =$ John asserts $X$, $U(y) =$ $y$ is unreliable, $j =$ John)
Stage B (Adding Susan):

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

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

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)”
Stage A1 (John and Donald):

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

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

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

(JA(X) = John asserts X, SA(X) = Susan asserts X, DA(X) = Donald asserts X, U(y) = y is unreliable \(j = \text{John}, s = \text{Susan}\))
12. 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)
13. 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

Odd 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: he misses some important generalizations, and ducks some difficult issues

(a) Pollock uses weakest-link ordering to calculate strengths, but in many normative domains the last-link ordering is more natural. Consider:

\[
\begin{align*}
\mathcal{W} &= \emptyset \\
\mathcal{D} &= \{r_1, r_2, r_3\} \\
Captain (r_1) &= T \rightarrow H \\
Major (r_2) &= T \rightarrow \neg W \\
Colonel (r_3) &= H \rightarrow W \\
r_1 < r_2 < r_3
\end{align*}
\]

\((H = \text{Heater on, } W = \text{Window open})\)

Does this work in some epistemic domains as well?
(b) More natural in normative domains to think of strengths as not totally ordered, which leads to many problems
Consider:

\[ W = \emptyset \]
\[ \mathcal{D} = \{ r_1, r_2, r_3, r_4 \} \]
\[ Captain (r_1) = T \rightarrow A \]
\[ Major (r_2) = T \rightarrow \neg A \]
\[ Priest (r_3) = T \rightarrow A \]
\[ Bishop (r_4) = T \rightarrow \neg A \]
\[ r_1 < r_2 \]
\[ r_3 < r_4 \]

Are there epistemic problems like this?
2. How do we reason about priorities?

\[ W = \{Q, R\} \]
\[ D = \{r_1, r_2, r_3, r_4\} \]
\[ r_1 = Q \rightarrow P \]
\[ r_2 = R \rightarrow \neg P \]
\[ r_3 = \top \rightarrow r_2 < r_1 \]
\[ r_4 = \top \rightarrow r_1 < r_2 \]
\[ < = \emptyset \text{ (initially)} \]
3. Is the reason relation necessarily “formal,” or can it be “material”?

A quote:

... prima facie reasons are supposed to be logical relationships between concepts. It is a necessary feature of the concept red that something’s looking red to me gives me a prima facie reason for thinking it is red.

By contrast, Reiters’s defaults often represent contingent generalizations. If we know that most birds can fly, then the inference from being a bird to flying may be adopted as a default. [In my theory] the latter inference is instead handled in terms of the following reason schema:

“Most A’s are B’s and this is an A” is a prima facie reason for “This is a B”
4. For Pollock: arguments all “present” at once, and then evaluated together

I prefer: interleaving argument construction and evaluation

\[ W = \{ \neg (A \land B) \} \]
\[ D = \{ r_1, r_2, r_3 \} \]
\[ Priest (r_1) = \top \rightarrow A \]
\[ Bishop (r_2) = \top \rightarrow B \]
\[ Cardinal (r_3) = A \rightarrow \neg B \]
\[ r_1 < r_2 < r_3 \]
5. Issues surrounding 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?
6. Undercutting by weaker defaults

Take \( \langle W, D \rangle \) with

\[
\begin{align*}
W &= \{A, B\} \\
D &= \{r_1, r_2\} \\
r_1 &= A \rightarrow \text{Out}(r_2) \\
r_2 &= B \rightarrow P \\
r_1 &< r_2
\end{align*}
\]

Can \( r_2 \) be undercut by \( r_1 \)?

Not for Pollock, because …

It seems apparent that any adequate account of justification must have the consequence that if a belief is unjustified relative to a particular degree of justification, then it is unjustified relative to any higher degree of justification. (Pollock 1995, p 104)

But I disagree
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. Many open issues in this area, and a lot of interesting work to do