# Fallbacks and push-ons

Krister Segerberg

September 22, 2006

## 1 Background

Grove showed in effect that, in the theory of belief change initiated by Alchourrón, Gärdenfors and Makinson, belief states may be represented as sphere systems in the sense of David Lewis ([1], [2], [3]). A sphere system is essentially a linearly ordered family of subsets of a given space. In Lewis's own favourite interpretation of his sphere systems the elements corresponded to degrees of similarity with a given point ("possible world"), an interpretation that is not well suited to the theory of belief change.[1] Lindström and Rabinowicz proposed instead a more congenial interpretation under which the elements of a sphere system represent theories of varying strength: the strongest one is the current belief set of the agent, while the others are weaker theories on which the agent may fall back if he is challenged by new information that is inconsistent with what he currently believes ([4]). Those weaker theories were accordingly called fallbacks.

The term "fallback" is apt. A rational agent not in possession of indubitable knowledge must be prepared to give up some of his beliefs if new information makes his doxastic position untenable. He would then look for the most suitable fallback—a default theory, weaker than the theory representing his current beliefs, but as strong as possible under the circumstances.

But isn't it also possible, even for a rational agent, to favour certain views (conjectures, "hunches") without believing that they are true? If fallbacks model an agent's disposition to shed old beliefs, couldn't there be a dual concept that models his disposition to subscribe to new beliefs? Probably many students of Grove's modelling have, at some time or other, toyed with the idea of extending

[1] Which is not to rule out the use of sphere systems in an analysis of verisimilitude.

This paper was written while the author was a fellow-in-residence at N.I.A.S. (Netherlands Institute for Advanced Study)



Printed from: *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. www.fil.lu.se/hommageawlodek

the sphere systems by introducing "push-ons" ("inside" the belief set) to correspond to the fallbacks (which are "outside" the belief set).[2] It is certainly possible to develop an abstract modelling of that kind; here we offer one example.

## 2 Formalia

Let $(U, T)$ be a Stone space (that is, a compact, totally separated topological space).[3] We say that $(I, O)$ is an *onion pair* if

(i) $I$ and $O$ are nonempty families of closed subsets of $U$,

(ii) $O$ is linearly ordered by the subset relation and closed under arbitrary nonempty intersection,

(iii) there is a subset $B$ of $U$ such that $B \in I \cap O$ and $B = \bigcup I = \bigcap O$,

(iv) there is a subset $K$ of $U$ such that $K \in O$ and $K = \mathsf{C} \bigcup O$.

The sets $B$ and $K$ are called the *belief set* and the *commitment set*, respectively. The elements of $O$ other than $B$ are called *fallbacks* and those of $I$ other than $B$ *push-ons*. If $C$ is a maximal chain of elements of $I$ we call $\bigcap C$ a *possible conjecture*.[4] If $(I, O)$ is an onion pair we refer to $I$ as the *inner onion* and to $O$ as the *outer onion*.

A *frame* is a structure $(U, T, Q, R, E, J)$ such that $(U, T)$ is a Stone topology, $Q$ is a quantity of onion pairs, while $R = \{R^P : P$ is a clopen subset of $U\}$ and $E = \{E^P : P$ is a clopen subset of $U\}$ and $J = \{J^P : P$ is a clopen subset of $U\}$ are families of binary relations in $Q$ satisfying certain conditions:

> if $((I, O), (I', O')) \in R^P$ then
>
> either $O$ overlaps with $P$ and $\bigcap O' = P \cap Z$,
> where $Z$ is the smallest element in $\{X \in O : P \cap X \neq \varnothing\}$,
>
> or else $O$ does not overlap with $P$ and $O' = \{\varnothing, K\}$;

[2] Our perspective is semantic, not syntactic!

[3] For more in the way of explanation of and motivation for the technical terminology in this paper, see for example [5] or [6]. We use the term 'onion' in place of David Lewis's more dignified but unwieldy 'sphere system'. Our onions differ from his in some respects.

[4] Those who are reluctant to use the term 'conjecture' for a theory may wish to refer to $\bigcap C$ as a *conjectured theory*.

if $((I,O),(I',O')) \in E^P$ then
$I' = \{P \cap X : X \in I\} \cup \{\mathsf{C}\bigcup_{X\in I}(P \cap X)\}$ and $O \cup \{\mathsf{C}\bigcup_{X\in I}(P \cup X)\} \subseteq O'$;

if $((I,O),(I',O')) \in J^P$ then $I' = \{X \in I : X \subseteq B'\}$,
where $B' = \mathsf{C}\bigcup\{X \in I : X \subseteq P\}$.

The set $U$ is called the *universe* of the frame. The elements of $R$ are referred to a *revision* relations, those of $E$ as *expansion* relations and those of $J$ as *jumps* (or *jumping-to-conclusion* relations).

We assume a language for classical propositional logic with some additional modal (unary propositional) operators: $\mathbf{B}, \mathbf{K}, \mathbf{C}, \mathbf{D}$ (doxastic) and $[*\ldots], [+\ldots]$, $[\oplus\ldots]$ (dynamic). Two restrictions apply: (i) a doxastic operator operates only on pure Boolean formulæ, and (ii) in a dynamic operator the three dots must be replaced by a pure Boolean formula. (A pure Boolean formula is a formula all of whose operators are truth-functional.) We write $\mathbf{b}, \mathbf{k}, \mathbf{c}, \mathbf{d}, \langle *\ldots\rangle, \langle +\ldots\rangle, \langle \oplus\ldots\rangle$ for the duals of $\mathbf{B}, \mathbf{K}, \mathbf{C}, \mathbf{D}, [*\ldots], [+\ldots], [\oplus\ldots]$.

A *valuation* (in $(U,T)$) is a function from the set of propositional letters to the set of clopen subsets of $U$. Given a frame and a valuation, let us write $[\![\phi]\!]$ for the truth-set of a pure Boolean formula, that is, the subset of $U$ at which the formula in question is true under the valuation. (We omit details since only classical notions are involved so far.) By a *reference point* we mean a triple $(I,O,u)$ where $(I,O)$ is an onion pair and $u$ is an element of the universe of the frame. We define the notion of truth of a formula in a frame under a valuation at a reference point as follows (denoted by the symbol $\vDash$): for any onion pair $(I,O)$, point $u$ in $U$, pure Boolean formula $\phi$ and formula $\theta$,

$(I,O,u) \vDash \phi$ iff $u \in [\![\phi]\!]$, if $\phi$ is pure Boolean,

[obvious conditions for the truth-functional operators]

$(I,O,u) \vDash \mathbf{B}\phi$ iff $\bigcap O \subseteq [\![\phi]\!]$,

$(I,O,u) \vDash \mathbf{K}\phi$ iff $\bigcup O \subseteq [\![\phi]\!]$,

$(I,O,u) \vDash \mathbf{C}\phi$ iff, for all maximal chains $C$ in $I$, $\bigcap C \subseteq [\![\phi]\!]$,

$(I,O,u) \vDash \mathbf{D}\phi$ iff, for some maximal chain $C$ in $I$, $\bigcap C \subseteq [\![\phi]\!]$,

$(I,O,u) \vDash [+\phi]\theta$ iff it is the case that,
for all onion pairs $(I',O')$ such that $((I,O),(I',O')) \in E^{[\![\phi]\!]}$,
$(I',O',u) \vDash \theta$,

$(I, O, u) \vDash [\oplus\phi]\theta$ iff it is the case that,
for all onion pairs $(I', O')$ such that $((I, O), (I', O')) \in J^{[\![\phi]\!]}$,
$(I', O', u) \vDash \theta$,

$(I, O, u) \vDash [*\phi]\theta$ iff it is the case that,
for all onion pairs $(I', O')$ such that $((I, O), (I', O')) \in R^{[\![\phi]\!]}$,
$(I', O', u) \vDash \theta$.

A formula is *valid* if it is true at all reference points in all frames under all valuations.

The set of valid formulæ forms a modal logic in which each modal operator is normal.

# 3 Interpretation

Of our nonclassical operators, four have standard interpretations: for $\mathbf{B}\phi$ and $\mathbf{K}\phi$, read, respectively,

"the agent believes that $\phi$" or (equally accurately)
"it is one of the agent's revisable beliefs that $\phi$",

"the agent knows that $\phi$" or (more accurately)
"it is one of the agent's nonrevisable beliefs that $\phi$'.

And for $[*\phi]\theta$ and $[+\phi]\theta$ read, respectively,

"after the agent has revised his beliefs by $\phi$, it is the case that $\theta$',

"after the agent has expanded his beliefs by $\phi$, it is the case that $\theta$".

In the same vein we may suggest, for $\mathbf{C}\phi$ and $\mathbf{D}\phi$, respectively,

"according to the agent's conjectures it is the case that $\phi$",

"according to one of the agent's conjectures it is the case that $\phi$".

Finally, for $[\oplus\phi]\theta$ read

"after the agent has jumped to the conclusion that $\phi$, it is the case that $\theta$".

# 4 Comments

The presentation of our modelling leaves many questions unanswered, in particular the most important one: where do the inner onions come from? If $(I, O)$ and $(I', O')$ are related by some jump $J^P$, then our theory has nothing to say about $I'$ if $P$ happens to be incompatible with the agent's current beliefs (that is, if $P \cap \bigcap O = \varnothing$); even if $P$ is compatible, our theory says little. It is small comfort that the original AGM theory suffered from a similar limitation by failing to describe $O'$ in the case just envisaged (beyond the general requirement listed above which gives the new belief set but not the complete outer onion

In order to account for the inner onions one must widen the perspective and bring in new parameters or concepts. One alternative might be to turn to Reiter's default logic, which endows the agent with a repertoire of defeasible inference rules and which is known to be congenial to DDL ([8], [9]). A more ambitious alternative would be to try to conceive of a mechanism —a research programme? a line of investigation?— that would generate "plausibilities" (the inner onions) in some systematic manner. A beginning would be to represent such mechanisms as functions $\mathcal{R}$ from closed sets $X$ (possible belief sets) to families of clopen subsets of $X$ (inner onions). The difficulty would the be to find fruitful conditions to characterize $\mathcal{R}$. Perhaps the ideas in [7] might be useful here.[5]

# References


[1] Carlos Alchourrón, Peter Gärdenfors & David Makinson. "On the logic of theory change." *The journal of symbolic logic*, vol. 50 (1985), pp. 510-530.

[2] Adam Grove. "Two modellings for theory change." *Journal of philosophical logic*, vol. 17 (1988), pp. 157-170.

[3] David Lewis. *Counterfactuals.* Oxford: Blackwell, 1973.

[4] Sten Lindström & Wlodek Rabinowicz. "Epistemic entrenchment with incomparabilities and relational belief revision." In *The logic of theory change*, edited by André Fuhrmann & Michael Morreau, pp. 93-126. Lecture Notes in Artificial Intelligence, no. 465. Berlin: Springer-Verlag, 1990.


[5] This note was written while the author was a fellow-in-residence at N.I.A.S. (The Netherlands Institute for Advanced Study).

[5] Sten Lindström & Krister Segerberg. “Modal logic and philosophy.” To appear in *Handbook of modal logic*.

[6] Hannes Leitgeb & Krister Segerberg. “Doxastic dynamic logic: why, whether, how”. To appear in *Knowledge, rationality and action*.

[7] Erik J. Olsson & David Westlund. “On the role of research agenda in epistemic change.” In *Modality matters*, edited by Henrik Lagerlund, Sten Lindström & Rysiek Sliwinski, pp. 323-337. Uppsala philosophical studies, vol. 53 (2006).

[8] Ray Reiter. “A logic of default reasoning.” *Artificial intelligence*, vol. 13 (1980), pp. 81-132.

[9] Krister Segerberg. “Delfault logic as dynamic doxastic logic.” *Erkenntnis*, vol. 50 (1999), pp. 333-352.

***Envoi***

For a short period in the early 1990s Uppsala was one of the best places in the world for the study of the logic of belief change. When I arrived there in 1991, Wlodek Rabinowicz was the Head of Department (an office that did not visibly burden him), Sten Lindström and Sven Ove Hansson were already there, John Cantwell and Tor Sandqvist were soon to join as graduate students. Wlodek’s never ending enthusiasm (along with Sten’s never ending objections) created a unique and wonderful environment.

The issues dealt with in this paper could very well have been discussed during those exciting years, even though I don’t remember that they were. If they had been, this paper would already have been written.