numéro 6

juin 1995

Αρχαι

[a r k h a ï]

www.arkhai.com

Nicolas MONOD

Une logique qui parle d'elle-même

LE PRÉCÉDENT numéro d'Ἀρχαί contenait une brève esquisse de ce que pourrait être une syntaxe de l'autoréférence. Devant l'intérêt manifesté par certains lecteurs, j'ai procédé à une formalisation complète d'une telle syntaxe sur des bases plus agréables à manipuler. Le présent article établit un certain nombre de propriétés syntaxiques mais aussi sémantiques de ce langage ; le lecteur me pardonnera donc si la première partie de ce texte reprend quelques éléments de l'ancien article.

PREMIÈRE SECTION : UNE FORMALISATION DE L'AUTORÉFÉRENCE

Le calcul propositionnel (noté $\mathcal{L}_0$) vise à formaliser la structure de certains types de propositions, comme par exemple "Socrate court et tu es à Rome", "s'il fait beau, alors il pleut" et plus généralement tout assemblage de propositions dites élémentaires reliées par les connecteurs "et", "ou", "non", etc. Il va de soi que cette forme de représentation ne peut qu'être extrêmement grossière et partielle. Il y a une classe de propositions impossibles à formaliser dans $\mathcal{L}_0$ qui semblent pourtant assez peu éloignées des assemblages de $\mathcal{L}_0$; par exemple "la présente proposition est fausse", ou "*la partie en italiques de cete prase est équivalente à* la négation de *la phrase tout entière*". Ces deux propositions font référence à elles-mêmes ('la présente proposition...') ; elles sont autoréférentes. Il y en a bien d'autres, et de beaucoup plus complexes : une proposition autoréférente peut se référer à elle-même, à une partie d'elle-même ou à plusieurs parties d'elle-même.

Lorsqu'on s'occupe de propositions autoréférentes, on est sans cesse confronté à des paradoxes et à des contradictions. Cela tient au fait qu'on les manipule comme des propositions de la logique ordinaire, formalisée par $\mathcal{L}_0$. Ainsi, on dira par exemple de la phrase citée plus haut :

« De deux choses l'une. Soit elle est vraie, soit elle est fausse. Si elle est vraie, elle est fausse, ce qui est absurde. Si elle est fausse, ... » etc.

Ce raisonnement, correct en logique propositionnelle, conduit ici à une cercle vicieux de contradictions succesives.

Par conséquent, si l'on veut traiter les expressions autoréférentes de façon sensée, il faut abandonner la logique propositionnelle ordinaire au profit d'une logique qui reste à déterminer : on ne sait pas *a priori* quelles règles de déduction peuvent être appliquées aux propositions autoréférentes.

Le langage Ourobore

Notre premier objectif, avant même de pouvoir chercher à manipuler des expressions autoréférentes, est de trouver un moyen de les représenter formellement. Commençons par rappeler comment sont représentées les propositions de la logique *ordinaire*.

Considérons un ensemble $\mathscr{V}$ de variables (c'est à dire un ensemble infini dénombrable quelconque). Le langage $\mathscr{L}_0$ du calcul propositionnel est alors défini par les quatre règles suivantes :

(*i*) Tout élément de $\mathscr{V}$ est dans $\mathscr{L}_0$.
(*ii*) Si α est dans $\mathscr{L}_0$, alors $\sim\alpha$ est dans $\mathscr{L}_0$.
(*iii*) Si α et β sont dans $\mathscr{L}_0$, alors $(\alpha\Rightarrow\beta)$, $(\alpha\Leftarrow\beta)$, $(\alpha\Leftrightarrow\beta)$, $(\alpha\wedge\beta)$, $(\alpha\vee\beta)$ sont dans $\mathscr{L}_0$[1].
(*iv*) Rien n'est dans $\mathscr{L}_0$ sinon par ce qui précède.

Notre but est de proposer un formalisme *étendant* $\mathscr{L}_0$, c'est à dire le contenant ; ce formalisme devra de plus modéliser toute relation

[1]Le choix de ces symboles est affaire de convenance. Par ailleurs, on n'écrira que les parenthèses vraiment indispensables.

indiquant qu'une variable de l'expression représente une partie donnée de la proposition. Par exemple, pour formaliser la phrase du menteur, il s'agit de trouver un moyen d'indiquer que dans l'expression '$\sim p$', 'p' représente l'ensemble de l'expression. Il y a là une sorte de mise en abyme qui présente une difficulté formelle ; on pourrait affirmer en termes raccoleurs que l'on a affaire à une logique qui serait « fractale », puisqu'une proposition doit se retrouver dans une de ses parties. Si dans le cas du menteur plusieurs idées de formalisation viennent assez rapidement à l'esprit, il suffit sans doute de considérer des propositions comme '$(p\wedge q)\Rightarrow r$' avec la contrainte que 'p' représente la proposition entière et 'q' la partie '$(p\wedge q)$' pour se rendre compte des problèmes d'ambiguïté qui peuvent s'opposer à une formalisation hâtive.

Notons d'abord qu'il n'est pas nécessaire de s'occuper des cas où une variable doit représenter une sous-expression dans laquelle elle n'apparaît pas ; en effet, dans ce cas, il n'y a pas de cercle vicieux de « mise en abyme », et pour exprimer que dans une expression du type '$\mathrm{P}(\alpha,x)$' où x n'apparaît pas dans α, x doit représenter α, il suffira d'écrire '$\mathrm{P}(\alpha,x)\wedge\alpha\Leftrightarrow x$'.

J'ai choisi de marquer l'autoréférence par une structure syntaxique analogue à celle des quantificateurs. Pour indiquer qu'une sous-expression α est représentée par x, j'écris cette sous-expression '$Я x\alpha$'. Le symbole '$Я$' ("ya" russe) peut être compris comme le 'R' de "référence", retourné comme le sont les quantificateurs. Ainsi, la phrase du menteur s'écrit '$Я p\sim p$' et la proposition '$(p\wedge q)\Rightarrow r$' avec la contrainte que 'p' représente la proposition entière et 'q' la partie '$(p\wedge q)$' s'écrit '$Я p(Я q(p\wedge q)\Rightarrow r)$'. Nous pouvons donc définir formellement le langage $\mathcal{O}$ des propositions autoréférentes ou ourobores[2] en ajoutant une règle à celles que nous connaissons pour $\mathcal{L}_0$:

[2]Ourobores, voilà comment Daumal qualifie les concepts qui rappellent le serpent Ouroboros qui se mord la queue ; de là la lettre 'O' pour désigner ce langage.

(*i*) Tout élément de $\mathcal{V}$ est dans $\mathcal{O}$.
(*ii*) Si α est dans $\mathcal{O}$, alors $\sim\alpha$ est dans $\mathcal{O}$.
(*iii*) Si α et β sont dans $\mathcal{O}$, alors $(\alpha\Rightarrow\beta)$, $(\alpha\Leftarrow\beta)$, $(\alpha\Leftrightarrow\beta)$, $(\alpha\wedge\beta)$, $(\alpha\vee\beta)$ sont dans $\mathcal{O}$.
(*iv*) Si α est dans $\mathcal{O}$ et x dans $\mathcal{V}$, alors $\mathfrak{A}x\alpha$ est dans $\mathcal{O}$.
(*v*) Rien n'est dans $\mathcal{O}$ sinon par ce qui précède.

Nous avons achevé de définir les expressions autoréférentes ; pourtant nous n'en sommes encore qu'au début de notre tâche : dans quelle mesure peut-on affirmer que cette formalisation rend bien compte de l'interprétation habituelle des phrases autoréférentes ? et comment raisonner avec de telles phrases, c'est-à-dire comment manipuler les objets de $\mathcal{O}$? quelles sont les propositions « vraies » de $\mathcal{O}$, et qu'est-ce que "vrai" veut dire dans ce langage ?

DEUXIÈME SECTION : LES THÉORÈMES DE $\mathcal{O}$

Il y a deux manières très différentes de déterminer si une expression logique traditionnelle doit être considérée comme « vraie ».

La première façon consiste à accepter une expression si on peut la démontrer à partir de certains axiomes et de quelques règles que l'on convient de fixer ; le cas échéant, on appelera cette expression un *théorème*. L'assertion "α est un théorème" ou "α est démontrable" est généralement notée $\vdash\alpha$. Cette notion purement syntaxique de vérité n'est que très rarement associée au mot "vérité" — du moins chez les logiciens. On préférera dire, en sacrifiant certes à une forme de jargonophasie, "démontrabilité", "validité syntaxique".

La seconde façon de procéder est dite sémantique. C'est à elle que les logiciens réservent la notion de vérité. Son application au cas qui nous intéresse sera décrite plus en détail dans la troisième section de ce texte, il nous suffira ici de savoir qu'elle consiste à donner une interprétation à chacune des expressions du langage formalisé, à l'aide des fameuses « tables de vérité ». Une des interprétations possibles est « le vrai ». Une expression sera dite *vraie*, ou sera appelée une *tautologie*, si toute interprétation possible l'interprète comme « le vrai ». Cet état de fait est noté $\models\alpha$.

Il reste à signaler que ces deux notions se confondent dans le cas du calcul propositionnel ; ce résultat fondamental est une forme de complétude de ce système : tout ce qui est vrai est démontrable, et vice-versa. *Il en va cependant différemment des systèmes logiques plus complexes, comme ceux qu'utilisent les mathématiciens.*

Avant d'aborder une première définition de la validité d'une expression de $\mathcal{O}$, nous devons nous acquitter de quelques préliminaires techniques.

Définitions

Une *référence* est une expression commençant par 'Я'. Cette définition se réfère à l'écriture *sans* abus du type de ceux signalés en note 1. Il faut donc se garder de considérer une expression du type '$Яp(p \wedge q) \Leftrightarrow q$' comme une référence ; en effet, sa seule écriture rigoureusement conforme aux règles que nous avons précisées serait '$(Яp(p \wedge q) \Leftrightarrow q)$', qui commence par '(' et n'est donc pas une référence.

L'*intérieur* d'une référence est la plus grande sous-expression stricte qu'elle contient. Exemple : l'intérieur de '$Яp{\sim}p$' est '${\sim}p$'. Si α est une référence, son intérieur est noté α°.

La variable qui suit immédiatement le symbole 'Я' d'une référence est dite *liée* à cette référence.

Les variables qui ne sont pas liées sont appelées variables *libres*.

La *portée* d'une occurrence d'une variable liée est la référence qui la lie, privée des sous-références qui lient la même variable.

Le *degré* d'une expression est le plus grand nombre de références emboîtées les unes dans les autres que contient cette expression. L'ensemble des expressions de degré i est noté $\mathcal{O}_i$. Ainsi, $\mathcal{O}_0$ n'est autre que $\mathcal{L}_0$ lui-même.

Le *codegré* d'une sous-expression est le nombre de références dont elle est partie propre. En particulier, le degré d'une expression n'est jamais inférieur à la somme du degré et du codegré d'une de ses parties.

Pour tout ensemble A d'expressions, nous notons '$\bigwedge$A' la conjonction de tous les éléments de A et '$\bigvee$A' la disjonction de tous ses éléments.

Si α, β et γ sont des expressions, nous notons $\alpha(\beta|\gamma)$ l'expression obtenue par la substitution simultanée de γ à toutes les occurrences de β. En particulier, β sera souvent une variable.

Si β est une occurrence d'expression, la substitution est notée $\alpha[\beta|\gamma]$. Les deux définitions précédentes restent inchangées si α est l'occurrence d'une expression.

Si A est un ensemble de termes de la forme $(\beta|\gamma)$ ou $[\beta|\gamma]$, αA représente l'application simultanée de chacun de ces termes à α.

Validité syntaxique

Nous n'allons pas définir la validité syntaxique par des axiomes et des règles de déduction, pour éviter les écueils signalés en introduction. La recherche de telles règles ne pourra être que le couronnement de tout notre travail. Dans ces circonstances, il conviendra de ne pas dire des expressions syntaxiquement valides qu'elles sont "démontrables" ; il serait plus correct de dire "vérifiables", à condition de bien garder à l'esprit qu'il s'agit d'une vérification purement syntaxique.

La définition de validité syntaxique que je propose consiste à transformer toute expression α de $\mathcal{O}$ en une formule T(α) du système classique $\mathcal{L}_0$ selon un procédé syntaxique ; on dira alors que la formule α de $\mathcal{O}$ est syntaxiquement valide chaque fois que la formule T(α) obtenue sera syntaxiquement valide *dans* $\mathcal{L}_0$.

Pour commencer, il faut que cette notion de validité coïncide avec la validité dans $\mathcal{L}_0$ chaque fois que α est dans $\mathcal{O}$ *et* dans $\mathcal{L}_0$. Dans ce but, nous définissons T(α) = α pour tout α dans $\mathcal{L}_0$. De cette façon, nous pouvons sans risque d'ambiguïté noter "$\vdash \alpha$" la validité syntaxique de n'importe quelle expression α de $\mathcal{O}$, "$\vdash \alpha$" signifiant simplement "$\vdash$T(α)".

Si à présent α est une expression quelconque de $\mathcal{O}$, mais non pas de $\mathcal{L}_0$, son degré n'est pas inférieur à un, et peut être noté d+1.

Nommons $\alpha_{0,1}$ à α_{0,n_0} les n_0 occurrences de références de codegré 0 que contient α, $\alpha_{1,1}$ à α_{1,n_1} celles de codegré 1, et ainsi de suite jusqu'à $\alpha_{d,nd}$. Notons encore $a_{i,j}$ la variable liée par l'occurrence $\alpha_{i,j}$. Enfin, choisissons des variables $x_{i,j}$ toutes distinctes et n'apparaissant pas dans α.

Nous allons considérer toutes les applications qui associent à chaque variable soit le symbole '~', soit la suite '~~'. Il sera commode de noter l'ensemble de ces applications par le sigle $\{\sim,\sim\sim\}^{\mathbb{N}\times\mathbb{N}}$. Si ν est une telle application, et si nous avons indicé une variable $x_{i,j}$, $\nu_{i,j}$ représentera l'élément que ν fait correspondre à $x_{i,j}$. Nous définissons encore $\nu^\circ \in \{\sim,\sim\sim\}^{\mathbb{N}\times\mathbb{N}}$ par $\nu^\circ{}_{i,j}$ = '~~'.

Nous définissons :

$$\Theta_\nu(\alpha) = \bigwedge \{\ x_{i,j} \Leftrightarrow \nu_{i,j}\alpha^\circ{}_{i,j}\ \{[\alpha_{k,l}|\nu_{k,l}x_{k,l}]|k>i,\ l\leq n_k\}(a_{i,j}|\nu_{i,j}x_{i,j})\ \big|\ i\leq d,\ j\leq n_i\ \}.$$

$$\mathrm{A}(\alpha) = \alpha[\alpha_{0,1}|x_{0,1}]\ldots[\alpha_{0,n_0}|x_{0,n_0}] \Leftarrow \Theta_{\nu^\circ}(\alpha).$$

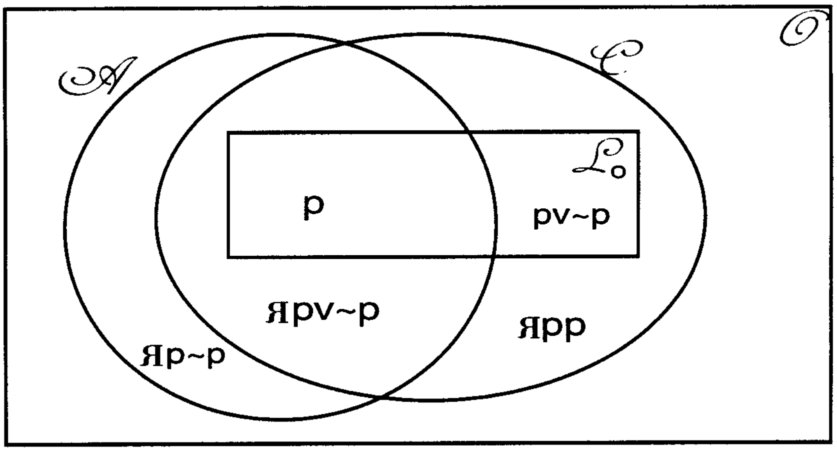
$$\mathrm{C}(\alpha) = \bigvee \{\ \Theta_\nu(\alpha) \mid \nu \in \{\sim,\sim\sim\}^{\mathbb{N}\times\mathbb{N}}\ \}.$$

Finalement :

$$\boxed{\mathrm{T}(\alpha) = \mathrm{C}(\alpha) \wedge \mathrm{A}(\alpha).}$$

Nous avons ainsi défini les expressions que nous appelerons désormais "syntaxiquement valides" ou "vérifiables" ; cependant nous ne savons encore presque rien de notre langage $\mathcal{O}$: nous ne savons même pas s'il ne serait pas contradictoire ! Pour en aprendre plus, il s'avère utile d'étudier aussi sa sémantique — qui est l'objet de la section suivante.

Nous notons $\mathscr{C}$ l'ensemble des expressions α telles que C(α) est un théorème de $\mathscr{L}_0$, et $\mathscr{A}$ l'ensemble de celles pour lesquelles A(α) est un théorème de $\mathscr{L}_0$. Ainsi, $\vdash\alpha$ est équivalent à $\alpha \in \mathscr{C} \cap \mathscr{A}$. La situation peut être résumée par le diagramme suivant :



La partie en dehors de $\mathscr{C}$ et de $\mathscr{A}$ est vide.

Forme normale

Nous avons déjà distingué une variable d'une de ses ocurrences ; appelons encore *réalisation* d'une variable liée l'ensemble de ses occurrences dans une de ses portées ou l'ensemble de ses occurrences qui sont hors de toute référence. Ainsi, dans l'expression '$(Я p(\sim p) \wedge Я p(p \Rightarrow Я p(p \wedge q)) \vee p)$', la variable '$p$' compte sept occurrences et quatre réalisations.

Nous dirons qu'une expression est de *forme normale* si chacune de ses variables liées n'apparaît que dans une seule réalisation, et dans cette réalisation seule. Les expressions '$Я p(p \wedge q) \Leftrightarrow p$' et '$Я pp \Leftrightarrow Я pp$' ne sont donc pas de forme normale. On peut toujours se ramener à une forme normale en renommant les variables. Les expressions de forme normale sont plus simples à manipuler que les autres, notamment par ce qu'elles ne connaissent pas la distinction entre une référence et une occurrence de cette référence, ni la distinction entre une variable liée et une réalisation de cette variable liée.

Si α est de forme normale, les expressions composant $T(\alpha)$ peuvent être simplifiées dans le sens suivant : il n'y a plus lieu d'introduire les $x_{i,j}$, qui sont tous rempacés par les $a_{i,j}$.

TROISIÈME SECTION : UN PEU DE THÉORIE DES MODÈLES

Il s'agit à présent de déterminer quand une expression de $\mathscr{O}$ est « vraie », ou tautologique, et de comparer cette situation au cas ou une expression est syntaxiquement valide. Nous commençons par rappeler comment on procède pour la logique ordinaire, le calcul propositionnel $\mathscr{L}_0$.

Cette approche sémantique permettra de mieux saisir la nature de notre formalisation, et en particulier de montrer que $\mathscr{O}$ n'est pas contradictoire.

Valuation et interprétation

Une *valuation* est une application qui associe à chaque variable la valeur $\mathfrak{V}$ ou $\mathfrak{F}$ (valeurs de vérité naïvement interprétées comme « le vrai » et « le faux »). On leur applique les opérations logiques ordinaires "et", "ou", "non", "si", etc. Par exemple, $\mathfrak{V}$ et $\mathfrak{F} = \mathfrak{F}$.

L'*interprétation* d'une expression α relativement à une valuation $\mathbf{f}$, notée $\mathfrak{Int}_{\mathbf{f}}(\alpha)$, est définie récursivement comme suit :

(*i*) Si α est une variable, $\mathfrak{Int}_{\mathbf{f}}(\alpha) = \mathbf{f}(\alpha)$.

(*ii*) Si α est de la forme $\sim\beta$, $\mathfrak{Int}_{\mathbf{f}}(\alpha)$ = non $\mathfrak{Int}_{\mathbf{f}}(\beta)$.

(*iii*) Si α est de la forme $\gamma\wedge\beta$, $\mathfrak{Int}_{\mathbf{f}}(\alpha) = \mathfrak{Int}_{\mathbf{f}}(\gamma)$ et $\mathfrak{Int}_{\mathbf{f}}(\beta)$.

etc pour tous les connecteurs de $\mathscr{L}_0$.

Une valuation est un *modèle* d'une expression si l'interprétation de cette expression relativement à cette valuation est $\mathfrak{V}$.

Finalement, on dira que α est une tautologie, ou est vraie ("$\models\alpha$"), si toute valuation est un modèle de α. Nous répétons que pour $\mathscr{L}_0$, $\models\alpha$ est équivalent à $\vdash\alpha$.

Nous allons étendre la théorie des modèles à $\mathcal{O}$, puis montrer que ce langage est lui aussi syntaxiquement complet — c'est-à-dire que dans notre logique autoréférente, tout ce qui est vrai est vérifiable par la méthode décrite dans la deuxième section, et vice-versa.

Nous conservons la définition d'une valuation et d'un modèle, à une seule modification près : la valuation doit associer l'une des valeurs $\mathfrak{V}$ ou $\mathfrak{F}$ à toute variable libre et à toute réalisation de toute variable liée. Nous pouvons néanmoins conserver l'ancienne définition de la notion de valuation chaque fois que nous avons affaire à une forme normale ; comme nous nous ramènerons toujours à cette situation, il n'est pas nécesaire d'établir une distinction entre les valuations de $\mathcal{L}_0$ et celles de $\mathcal{O}$.

Nous disons que deux valuations sont *semblables* pour une expression si elles diffèrent au plus sur des variables qui sont liées dans cette expression. Si l'expression n'est pas de forme normale, cette définition doit se formuler en demandant que les valuations ne diffèrent que sur des réalisations liées.

Contrairement au cas de $\mathcal{L}_0$, l'interprétation d'une expression de $\mathcal{O}$ relativement à une valuation peut prendre sa valeur parmi *trois* éléments : $\mathfrak{V}$, $\mathfrak{F}$ ou $\mathfrak{I}$ — 'i' comme *incohérent*. La définition récursive de $\mathfrak{Int}_{\mathfrak{f}}(\alpha)$ s'enrichit d'une nouvelle clause :

Si α est de la forme $\text{Я}x\alpha^\circ$, alors :

si $\mathfrak{Int}_{\mathfrak{f}}(\alpha^\circ) = \mathfrak{f}(x)$, $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{f}(x)$;
sinon, $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{I}$.

Quant aux autres opérateurs, leur définition reste la même, excepté que le résultat d'une opération est $\mathfrak{I}$ aussitôt que l'un des opérandes vaut $\mathfrak{I}$.

Enfin, une expression α de $\mathcal{O}$ est dite vraie, ou tautologique, si les deux conditions suivantes $\mathcal{A}$ et $\mathcal{C}$ sont satisfaites :

$\mathcal{A}$: Il n'y a aucune valuation $\mathfrak{f}$ pour laquelle $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{F}$.
$\mathcal{C}$: Toute valuation est semblable à un modèle.

On note alors $\models\alpha$. De même que lorsque nous avons défini $\vdash\alpha$, il est nécessaire de constater que si α est dans $\mathcal{O}$ *et* dans $\mathcal{L}_0$, les définitions de $\models\alpha$ relativement à $\mathcal{O}$ et à $\mathcal{L}_0$ coïncident.

On peut formuler $\mathcal{A}$ et $\mathcal{C}$ de façon équivalente par :

$\mathcal{A}'$: Pout toute valuation $\mathfrak{f}$, $\mathfrak{Int}_{\mathfrak{f}}(\alpha) \neq \mathfrak{F}$.
$\mathcal{C}'$: Toute valuation est semblable à une valuation $\mathfrak{f}$ pour laquelle $\mathfrak{Int}_{\mathfrak{f}}(\alpha) \neq \mathfrak{I}$.

Ces deux couples de conditions sont équivalents en tant que conjonctions, et $\mathcal{A}'$ est clairement équivalente à $\mathcal{A}$. Par contre, bien que $\mathcal{C}$ implique $\mathcal{C}'$, $\mathcal{C}$ n'est pas individuellement équivalente à $\mathcal{C}'$; en effet, '$p\wedge\sim p$' satisfait $\mathcal{C}'$ mais non pas $\mathcal{C}$.

Complétude syntaxique

Nous allons montrer que $\mathcal{O}$ est syntaxiquement complet, c'est-à-dire que $\vdash\alpha$[3] et $\models\alpha$ sont équivalents ; en fait nous montrerons un peu plus que cela :

1°) *Si* C(α) *est un théorème de* $\mathcal{L}_0$, *alors* α *satisfait* $\mathcal{C}'$.

Considérons une valuation quelconque $\mathfrak{g}$. C(α) est un théorème de $\mathcal{L}_0$, donc c'est une tautologie et $\mathfrak{Int}_{\mathfrak{g}}(C(\alpha)) = \mathfrak{V}$. Il existe donc un v tel que $\mathfrak{Int}_{\mathfrak{g}}(\Theta_{v}(\alpha)) = \mathfrak{V}$. Notons $\mathfrak{f}$ la valuation obtenue en inversant $\mathfrak{g}(x_{i,j})$

[3]Dans ce paragraphe et sauf indication contraire, α est de forme normale.

chaque fois que $v_{i,j}$ = '~'. $\mathfrak{f}$ est semblable à $\mathfrak{g}$; de plus $\mathfrak{Int}_{\mathfrak{f}}(\Theta_{v^\circ}(\alpha)) = \mathfrak{V}$, et donc $\mathfrak{Int}_{\mathfrak{f}}(\alpha) \neq \mathfrak{I}$. Ainsi α satisfait bien $\mathcal{C}'$.

2°) *Si α satisfait $\mathcal{C}'$, alors* C(α) *est un théorème de* $\mathscr{L}_0$.

Considérons une valuation quelconque $\mathfrak{g}$. Il s'agit de montrer que $\mathfrak{Int}_{\mathfrak{g}}(C(\alpha)) = \mathfrak{V}$. Comme α satisfait $\mathcal{C}'$, il y a une valuation $\mathfrak{f}$ semblable à $\mathfrak{g}$ pour laquelle $\mathfrak{Int}_{\mathfrak{f}}(\alpha) \neq \mathfrak{I}$. Il en suit que $\mathfrak{Int}_{\mathfrak{f}}(\Theta_{v^\circ}(\alpha)) = \mathfrak{V}$. En notant v l'application qui associe '~' aux seules variables x pour lesquelles $\mathfrak{f}(x) \neq \mathfrak{g}(x)$ (et '~~' sinon)[4], $\mathfrak{Int}_{\mathfrak{g}}(\Theta_{v}(\alpha)) = \mathfrak{V}$. Ainsi, $\mathfrak{Int}_{\mathfrak{g}}(C(\alpha)) = \mathfrak{V}$ et C(α) est bien un théorème de $\mathscr{L}_0$.

3°) *Si* A(α) *est un théorème de* $\mathscr{L}_0$, *alors α satisfait $\mathcal{A}'$.*

Supposons par l'absurde que $\mathfrak{f}$ soit une valuation pour laquelle $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{F}$. En partiulier, $\mathfrak{Int}_{\mathfrak{f}}(\alpha) \neq \mathfrak{I}$, et donc $\mathfrak{Int}_{\mathfrak{f}}(\Theta_{v^\circ}(\alpha)) = \mathfrak{V}$. Comme $\mathfrak{Int}_{\mathfrak{f}}(A(\alpha)) = \mathfrak{V}$, $\mathfrak{Int}_{\mathfrak{f}}(\alpha[\alpha_{0,1}|a_{0,1}]\ldots[\alpha_{0,n_0}|a_{0,n_0}]$[5]$) = \mathfrak{V}$. Mais comme $\mathfrak{Int}_{\mathfrak{f}}(\Theta_{v^\circ}(\alpha)) = \mathfrak{V}$ implique $\mathfrak{Int}_{\mathfrak{f}}(\alpha_{i,j}) = \mathfrak{Int}_{\mathfrak{f}}(a_{i,j})$ pour tous les $a_{i,j}$, et $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{Int}_{\mathfrak{f}}(\alpha[\alpha_{0,1}|a_{0,1}]\ldots[\alpha_{0,n_0}|a_{0,n_0}]) = \mathfrak{V}$, ce qui est absurde.

4°) *Si α satisfait $\mathcal{A}'$, alors* A(α) *est un théorème de* $\mathscr{L}_0$.

Considérons une valuation quelconque $\mathfrak{f}$. Il s'agit de montrer que $\mathfrak{Int}_{\mathfrak{f}}(A(\alpha)) = \mathfrak{V}$. Il suffit de montrer que cette interprétation n'est pas $\mathfrak{F}$, car A(α) est dans $\mathscr{L}_0$. Supposons donc par l'absurde que $\mathfrak{Int}_{\mathfrak{f}}(A(\alpha)) = \mathfrak{F}$. Alors $\mathfrak{Int}_{\mathfrak{f}}(\Theta_{v^\circ}(\alpha)) = \mathfrak{V}$ et $\mathfrak{Int}_{\mathfrak{f}}(\alpha[\alpha_{0,1}|a_{0,1}]\ldots[\alpha_{0,n_0}|a_{0,n_0}]) = \mathfrak{F}$. Ainsi, $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{F}$, ce qui contredit $\mathcal{A}'$. Donc A(α) est un théorème de $\mathscr{L}_0$.

C'est un corollaire immédiat des quatre résultats ci-dessus que $\models\alpha$ est équivalent à $\vdash\alpha$, et donc $\mathscr{O}$ est syntaxiquement complet.

De plus, $\mathscr{A}$ est exactement la classe des expressions satisfaisant $\mathcal{A}'$, et $\mathscr{C}$ est exactement la classe des expressions satisfaisant $\mathcal{C}'$.

---

[4] Il vaut la peine de noter que si ce v est bien conforme à la définition de la section précédante, c'est uniquement par ce que les deux valuations sont semblables.

[5] Voir la dernière remarque de la deuxième section.

Non-contradiction

Il est grand temps de montrer que $\mathcal{O}$ n'est pas contradictoire. Cela nous sera d'autant facilité que nous n'avons plus à distinguer $\vdash\alpha$ de $\models\alpha$.

Supposons donc par l'absurde que $\models\alpha$ et $\models\sim\alpha$. Par $\mathcal{C}$, il existe un modèle $\mathfrak{f}$ de α. Alors $\mathfrak{Int}_{\mathfrak{f}}(\sim\alpha) = \mathfrak{F}$, et par $\mathcal{A}$, $\sim\alpha$ n'est pas une tautologie, ce qui contredit l'hypothèse.

Tables de vérité

L'approche sémantique que nous avons exposée permet de décider si une expression est autoréférente par la méthode particulièrement agréable des « tables de vérité ». Voici la marche à suivre :

D'abord, on écrit sous chaque variable (ou réalisation de variable) les valeurs de vérité que lui attribuent toutes les valuations possibles. On détermine ensuite successivement la valeur (*i.e.* l'interprétation $\mathfrak{V}$, $\mathfrak{F}$ ou $\mathfrak{I}$) de chaque sous-expression selon la méthode décrite au paragraphe *valuation et interprétation*. Enfin, on teste les conditions $\mathcal{A}$ et $\mathcal{C}$. La condition $\mathcal{A}$ se vérifie immédiatement ; pour qu'il en aille de même avec $\mathcal{C}$, il suffit de faire varier en premier les variables (ou réalisations) liées, de cette sorte les valuations semblables seront successives. Voici pour exemple l'analyse de '$q\Leftarrow$Я$(p\sim p\vee q)$' :

q	$\Leftarrow$	Я	p	$\sim p$	$\vee$	q
$\mathfrak{V}$	$\mathfrak{V}$	$\mathfrak{V}$	$\mathfrak{V}$	$\mathfrak{F}$	$\mathfrak{V}$	$\mathfrak{V}$
$\mathfrak{V}$	$\mathfrak{I}$	$\mathfrak{I}$	$\mathfrak{F}$	$\mathfrak{V}$	$\mathfrak{V}$	$\mathfrak{V}$
$\mathfrak{F}$	$\mathfrak{I}$	$\mathfrak{I}$	$\mathfrak{V}$	$\mathfrak{F}$	$\mathfrak{F}$	$\mathfrak{F}$
$\mathfrak{F}$	$\mathfrak{I}$	$\mathfrak{I}$	$\mathfrak{F}$	$\mathfrak{V}$	$\mathfrak{V}$	$\mathfrak{F}$

$\mathcal{A}$ est satisfait, mais non $\mathcal{C}$, puisque la dernière ligne n'est semblable qu'à celle qui la précède.

QUATRIÈME SECTION : LA DÉRIVABILITÉ

Nous en arrivons à la partie la plus difficile et incertaine de notre recherche : la mise en évidence de règles de déduction pour $\mathcal{O}$. Il n'y a aucune raison *a priori* pour que l'on puisse exhiber un ensemble de règles qui permettent de déduire exactement les expressions vraies (ou syntaxiquement valides, puisque nous avons montré que ces notions se confondent pour notre langage). Pourtant, c'est bien des règles de déduction qu'il nous faut pour pouvoir manipuler notre langage $\mathcal{O}$, pour pouvoir *raisonner*. La première caractérisation des propositions « vraies » dans un sens ou un autre que nous ayons rencontré passait par une notion de validité syntaxique ; c'est la caractérisation la plus simple, mais non pas la plus parlante. La seconde caractérisation, faisant appel à la sémantique, est beaucoup plus explicite : elle nous informe en quelque sorte sur *ce que cela signifie* qu'une expression autoréférente soit vraie. Elle nous fournit également un algorithme de décision assez proche de l'analyse que l'on fait en pratique des phrases autoréférentes. Nous désirons cependant ajouter à notre assortiment une analyse purement déductionnelle.

Règles de déduction

Habituellement, on axiomatise la logique ordinaire $\mathcal{L}_0$ en énumérant un certain nombre d'axiomes (par exemple le "tiers exclu" : '$p\vee{\sim}p$') et en énonçant deux règles de déductions : la substitution (si $\vdash\alpha$ alors $\vdash\alpha(x|\beta)$) et le *modus ponens* (si $\vdash\alpha$ et $\vdash\alpha\Rightarrow\beta$, alors $\vdash\beta$). Nous allons procéder de façon analogue pour $\mathcal{O}$.

Nous notons $\Vdash\alpha$ le fait qu'une expression α soit *démontrable* ; en donnant les règles régissant l'emploi du symbole '$\Vdash$', nous définirons donc ce que "démontrable" veut dire. Comme nous désirons conserver les axiomes de $\mathcal{L}_0$ et ses règles de déduction internes, nous posons de façon résumée :

Si α est dans $\mathscr{L}_0$ et $\vdash\!\!-\alpha$, alors $\Vdash\cdots\alpha$.

Passons à la règle de substitution. Pour rester fidèle à notre conception de la vérité des propositions, nous n'autoriserons dans $\mathscr{O}$ que les substitutions d'une expression cohérente à une variable libre :

Si x est libre dans α et $\Vdash\cdots\alpha$, alors $\Vdash\cdots\alpha(x|\beta)$, pour autant que β soit cohérente (c'est-à-dire $\vdash\!\!-\mathrm{C}(\beta)$).

Enfin, voici le *modus ponens* :

Si $\Vdash\cdots\alpha$ et $\Vdash\cdots\alpha\Rightarrow\beta$, alors $\Vdash\cdots\beta$.

Conservation de la vérité

De même que nous avons comparé la vérité sémantique à la validité syntaxique, nous allons comparer la démontrabilité avec les deux formes de validité que nous conaissons déjà.

Il est clair par nos définitions que les axiomes proposés sont vrais. Passons donc à la première règle de déduction.

Considérons deux expressions α et β de $\mathscr{O}$ avec x libre dans α, $\vdash\alpha$ et $\vdash\!\!-\mathrm{C}(\beta)$. Nous pouvons supposer que toutes les expressions que nous manipulons sont de forme normale. Il s'agit de montrer que $\alpha(x|\beta)$ satisfait $\boldsymbol{A'}$ et $\boldsymbol{C'}$. Soit $\mathfrak{f}$ une valuation quelconque. Posons $\mathfrak{g}(x) = \mathfrak{Int}_{\mathfrak{f}}(\beta)$ et $\mathfrak{g}(y) = \mathfrak{f}(y)$ pour les autres variables y. De cette façon, $\mathfrak{Int}_{\mathfrak{f}}(\alpha(x|\beta)) = \mathfrak{Int}_{\mathfrak{g}}(\alpha) \neq \mathfrak{F}$ et $\alpha(x|\beta)$ satisfait $\boldsymbol{A'}$.Or $\mathfrak{f}$ est semblable pour α (et donc pour $\alpha(x|\beta)$) à une valuation $\mathfrak{h}$ pour laquelle $\mathfrak{Int}_{\mathfrak{h}}(\alpha) = \mathfrak{V}$. Comme $\models\mathrm{C}(\beta)$, il existe une valuation $\mathfrak{k}$ semblable à $\mathfrak{h}$ pour β (et donc semblable à $\mathfrak{f}$ pour $\alpha(x|\beta)$) pour laquelle $\mathfrak{Int}_{\mathfrak{k}}(\beta) \neq \mathfrak{I}$. Il suit que $\mathfrak{Int}_{\mathfrak{k}}(\alpha(x|\beta)) \neq \mathfrak{I}$, et donc $\alpha(x|\beta)$ satisfait $\boldsymbol{C'}$. Nous avons donc prouvé que $\Vdash\alpha(x|\beta)$.

Pour examiner le cas du *modus ponens*, considérons deux expressions α et β de $\mathscr{O}$ avec, $\vdash\alpha$ et $\vdash\alpha\Rightarrow\beta$. Il s'agit cette fois de montrer que β satisfait $\mathcal{A}'$ et $\mathcal{C}'$. Comme $\alpha\Rightarrow\beta$ satisfait $\mathcal{C}'$ et que deux valuations semblables pour $\alpha\Rightarrow\beta$ sont semblables, pour β, à deux valuations qui ne diffèrent que sur des variables n'apparaissant pas dans β (car $\alpha\Rightarrow\beta$ est supposée de forme normale), β satisfait $\mathcal{C}'$. Supposons par l'absurde que β ne satisfasse pas $\mathcal{A}'$, et que $\mathfrak{Int}_{\mathfrak{f}}(\beta) = \mathfrak{F}$. Comme α satisfait $\mathcal{C}$, on peut supposer sans perte de généralité, par le même argument de normalité que ci-dessus, que $\mathfrak{Int}_{\mathfrak{f}}(\alpha) = \mathfrak{V}$. Mais alors $\mathfrak{Int}_{\mathfrak{f}}(\alpha\Rightarrow\beta) = \mathfrak{F}$, ce qui contredit l'hypothèse et montre donc que $\dashv\beta$.

Ainsi, $\vdash\cdots\alpha$ implique $\vdash\alpha$.

Les deux règles que nous avons mises en avant sont donc cohérentes ; mais c'est une tout autre question de savoir si elles suffisent à engendrer tout $\mathscr{C}\cap\mathscr{O}\mathscr{V}$. Définir un ensemble de règles complet peut une tâche assez ardue, et j'espère avoir l'occasion d'y revenir.