Sur le théorème fondamental de Herbrand

Mise en contexte, preuve, application à la logique classique et modale

0. Introduction

0.1. Herbrand et la première crise des fondements

Au commencement de ses recherches logiques, Herbrand a pour guide premier les Principia Mathematica. Ses travaux sont très tôt axés sur la construction d'un algorithme de preuve des théorèmes ; tous les problèmes posés par la première crise des fondements l'intéressent ainsi au premier chef et parmi eux, celui de la réduction. Pour sa formulation rigoureuse, l'algorithmique projetée par Herbrand a besoin d'un langage simplifié, d'une « sténographie »[1] de l'ensemble des mathématiques.

Or Russell et Whitehead ont fourni ce langage. Ils ont su montrer que si l'on excepte la ponctuation, la totalité des propositions mathématiques est réductible à trois signes fondamentaux : la négation (notre ¬) et la disjonction (notre ∪) pour fonder la logique propositionnelle ; et l'existentiel (notre ∃) pour fonder la logique des prédicats.

	Notation d'Herbrand	Forme moderne
Disjonction	$\vee$	$\vee$
Négation	$\backsim$	$\neg$
Existence	$E\,x$ (parfois $\exists x$)	$\exists x$
Prédication	$P\,x$	$P\,x$

On voit facilement que les opérations logiques les plus simples peuvent être reformulées à partir de ces quelques relations fondamentales :

	Notation d'Herbrand	Forme moderne	Décomposition
Conjonction	$P \times Q$	$P \wedge Q$	$\neg\,((\neg\,P) \vee (\neg\ Q))$
Conditionnel	$P \supset Q$	$P \rightarrow Q$	$(\neg P) \vee Q$
Universel	$(x),\ P\,x$	$\forall x,\ P\,x$	$\neg\,(\exists x,\ \neg\,P\,x)$
Equivalence	$P \equiv Q$	$P \equiv Q$	$(P \rightarrow Q) \wedge (Q \rightarrow P)$

Les Principia donnent aussi 5 règles fondamentales de raisonnement (que Herbrand reprend en les modifiant à la marge dans [1930] et [1931c]).

Règle I	Substituabilité des propositions dans les relations élémentaires
Règle II	Si P est vraie, et si $P \rightarrow Q$, alors Q est vraie
Règle III	Si x est une variable, si $P\,x$ est vraie, alors $\forall x,\ P\,x$ est vraie
Règle IV	Soit x une variable, y un objet en général. Si $P\,x\,y$ devient vraie quand on remplace x par y, alors $\exists x, P\,x\,y$ est vraie
Règle V	Sont équivalentes pour la valeur de vérité les propositions $\exists x(P\,x \vee Q)$ et $(\exists x, P\,x) \vee Q$

[1] [1930a]

La réduction effectuée par Russell présente pour Herbrand deux avantages majeurs :

1. Elle écarte tout présupposé sur l'existence des objets étudiés. Les formes logiques isolées par Russell ne sont pas censées renvoyer à une quelconque réalité. Elles ne sont pas davantage des structures innées et universelles de la pensée. La logique des Principa n'est fondée que sur « une certitude expérimentale »[2] ; ses auteurs ont pris toute la littérature mathématique existante, ont cherché à la réduire à l'extrême, et ont fini par buter sur les trois symboles fondamentaux. Les recherches futures exigeront peut-être du logicien qu'il rajoute de nouveaux symboles premiers ; mais il s'agit là d'un problème contingent. L'essentiel pour Herbrand est de disposer d'une structure logique déliée de la réalité.

2. Ce faisant, il est possible d'étudier chaque théorie segmentée en faisant abstraction de ses objets (les entiers pour l'arithmétique, les figures pour la géométrie) pour ne retenir que la structure de ses raisonnements.

0.2. Pour une théorie de la preuve

0.2.1. Une « Beweistheorie »

Cette opération visant à dévider chaque théorie mathématique de ses objets pour ne garder que son squelette démonstratif, Herbrand en trouve le paradigme dans les travaux de celui qui sera son grand modèle intellectuel : David Hilbert. Hilbert appelle cette opération d'abstraction la « métamathématique », une discipline qui « a pour objet d'étude, non pas les objets dont s'occupent habituellement les mathématiciens, mais les phrases mêmes qu'ils peuvent prononcer sur ces objets. Elle prend en considération les propositions que l'on peut énoncer dans telle théorie, et cherche leurs propriétés caractéristiques : c'est en quelque sorte une mathématique du langage. »[3]

Héritier des recherches fondatrices de Peano, le métamathématicien considère que chaque champ des mathématiques peut être entièrement réécrit avec la logique de premier ordre, dans un langage ne comprenant que des variables, plus un nombre fini (de préférence très limité) d'objets fondamentaux et de prédicats fondamentaux (qui ne sont pas réductibles par l'analyse logique). Par exemple[4], l'arithmétique est virtuellement reformulable avec un seul objet fondamental (le zéro) et un seul prédicat (l'addition A qui permet d'obtenir $n+1$ à partir de l'entier n) ; tous les autres objets, toutes les autres propositions ne sont usitées que par principe d'économie ; ils n'ont pas de valeur démonstrative propre ; on pourrait par exemple réexprimer l'entier 5 sous la forme $A^5 0$; de même, en théorie, chaque théorème de l'arithmétique, même le plus élaboré, est reformulable par des combinaisons n'impliquant que le prédicat A.

Prenons[5] par exemple le théorème de Zeckendorf : « Tout entier naturel peut s'écrire comme la somme de termes non-consécutifs d'une suite de Fibonacci », formellement :

$$\forall N \in \mathbb{N},\ \exists!\,(c_0,\dots,c_k) \in \mathbb{N}^k,\ c_0 \geqslant 2,\ c_{i+1} > c_i + 1,\ t.q.$$

$$N = \sum_{i=0}^{k} \mathcal{F}_{c_i}$$

Où $\mathcal{F}_n$ donne le n-ième terme de la suite de Fibonacci.

La démonstration (de l'existence d'une telle décomposition) se fait facilement par récurrence :

Initialisation : P_1 est vraie puisque $1 = \mathcal{F}_1$

Hérédité : Supposons que P_N soit vraie : on peut écrire une décomposition :

$$N = \mathcal{F}_{i_1} + \dots + \mathcal{F}_{i_k}$$

On pose alors facilement :

$$N + 1 = \mathcal{F}_2 + \mathcal{F}_{i_1} + \dots + \mathcal{F}_{i_k}$$

Pour un résultat théorique aussi élémentaire, il est facile de concevoir tous les objets usités (par exemple les nombres de Fibonacci $\mathcal{F}_n$) comme des abréviations d'une écriture rigoureuse qui ne recourrait que le zéro et la relation A.

Dans ce cadre conceptuel, les axiomes d'une théorie ne sont plus tant des propositions premières non-prouvées que des « matrices »[6] formelles, des « structures vides »[7] censées permettre de générer une infinité d'autres propositions. Exemple d'Herbrand : « l'axiome d'induction totale » (qui est au fondement du raisonnement par récurrence) :

$$(P0 \vee (\forall n,\ Pn \to P(n+1))) \to \forall n,\ Pn$$

[2] [1930a, 247]
[3] [1930a, 243]
[4] [1930a, 244]
[5] L'exemple est de nous, ce résultat étant postérieur à la mort de Herbrand
[6] Ibid.
[7] Ibid.

On voit qu'il ne s'agit pas tant là d'un axiome en soit, mais de la « matrice d'une infinité d'axiomes », i.e. toutes les propriétés qui seront démontrées à partir de ce type de raisonnement.

0.2.3. Vers un algorithme de décidabilité

Une fois ce dévidage, cette réduction de chaque théorie effectués, la création d'un algorithme qui décide de la valeur d'une preuve mathématique devient beaucoup plus simple. Le métamathématicien n'a plus à travailler sur des objets en particulier, mais sur des objets indéterminés, des objets vides, les « variables »[8]. Quant aux propositions et théorèmes de chaque discipline, ils sont déduits des objets et des prédicats premiers « par des méthodes purement logiques, par des règles de raisonnement universelles, indépendantes de la théorie envisagée »[9].

La « Beweistheorie » se déploie alors ainsi ; Soit une théorie mathématique donnée. Soit une propriété P. La démarche hilbertienne consiste :

1. A construire une méthode de preuve, universellement applicable, permettant de savoir si telle proposition possède la propriété P ;

2. A démontrer que tous les axiomes de la théorie possèdent la propriété P ;

3. A montrer que toutes les propositions sont déductibles des axiomes (à partir des seules règles de Russell et Whitehead).

Une théorie est alors dite « contradictoire » s'il est possible d'y démontrer à la fois P et $\neg P$. Elle est « complète » s'il est possible de démontrer soit P soit $\neg P$

Ce qui intéresse le métamathématicien n'est ainsi plus tant la valeur sémantique des assertions dans une théorie que la validité de leurs structures, le fait qu'une proposition soit ou non une « identité », terme auquel Herbrand donne une définition strictement formelle : « Est une identité propositionnelle [...] toute proposition dont la valeur logique est le vrai, quelles que soient les valeurs logiques et les lettres qui y figurent,»[10].

Pour Herbrand, la véritable question de la « décidabilité » (au sens « restreint »[11] du terme) ne porte pas sur la valeur de vérité d'une proposition, mais sur sa nature d'identité, sa validité indépendamment de l'interprétation donnée aux termes (ce que nous appelons la « satisfaisabilité »). C'est là « l'Entscheidungsproblem » tel que le conçoivent les Hilbertiens à l'orée des années 1930.

0.2.4. Un principe de neutralité

Herbrand insiste sur la neutralité épistémologique qu'implique le programme hilbertien et l'idée d'algorithme de décidabilité : « La métamathématique cherche seulement à examiner les théories déjà existantes et étudie les caractères des propositions qui y sont vraies ; elle ne prend pas part aux discussions que celles-ci soulèvent ; elle ne cherche pas à les départager ; elle se borne à signaler qu'en raisonnant de telle manière, les résultats obtenus posséderont telles propriétés [...]Toutes les théories ont à ses yeux égal droit de cité »[12]

Plus profondément encore, le programme d'Hilbert se fonde sur une complète neutralité sur les choses en soi : « À aucun moment la métamathématique ne cherchera à savoir si une théorie donnée décrit convenablement les propriétés de tel objet, si elle correspond à quelque chose de réel ou non ; elle ne le pourrait d'ailleurs pas. [...] Le rôle des mathématiques est peut-être uniquement de nous fournir des raisonnements et des formes, et non pas de chercher quels sont ceux qui s'appliquent à tel ou tel objet. Pas plus que le mathématicien qui étudie l'équation de propagation des ondes n'a à se demander si dans la nature les ondes satisfont effectivement à cette équation, pas plus en étudiant la théorie des ensembles ou l'arithmétique il ne doit se demander si les ensembles ou les nombres auxquels il pense intuitivement satisfont bien aux hypothèses de la théorie qu'il considère. Il doit se borner à développer les conséquences de ces hypothèses et à les présenter de la manière la plus suggestive ; le reste est le rôle du physicien, ou du philosophe »[13].

Pour Herbrand, l'objectivité mathématique n'est pas découverte par l'homme dans les objets du monde extérieur ; elle est « créée »[14].

0.3. Herbrand et la « seconde » crise des fondements

0.3.1. Le moment intuitionniste

Ce principe de neutralité est aussi une réponse aux débats contemporains.

Au moment où Herbrand arrive à Göttingen grâce à une bourse de la fondation Rockefeller y sévit ce qu'on a pu appeler la seconde crise des fondements[15], ouverte par l'émergence de l'intuitionnisme de Brouwer, qui refuse avec

[8] « Certaines lettres que nous appellerons 'variables', seront considérées comme des indéterminées, et représenteront un objet quelconque d'une certaine catégorie d'objets (elles jouent le même rôle que les variables en algèbre) » [1930a, 244]
[9] [1930a, 246]
[10] [1928]
[11] [1931 (1968, p. 178)]
[12] [1930a]
[13] Ibid.
[14] Chevalley, J., in [1968]
[15] [Dubucs, J., Egre, P., 2006]

fracas la validité universelle des principes logiques classiques et en général l'approche linguistique de la logique (l'idée qu'elle devrait s'appuyer sur une théorie de la signification).

Dans l'esprit de Brouwer, l'intuitionnisme devait s'opposer aussi bien au programme ontologique (celui des partisans du platonisme mathématique, héritiers du Frege des *Grundlagen der Arithmetik*) qu'au programme symbolique et métamathématique de Hilbert. A cette double opposition correspondait un double refus ; l'objet mathématique ne devait pas être pensé comme une chose en soi, ou comme une vulgaire convention symbolique, mais comme un acte.

Cet acte producteur a deux fondements chez Brouwer : 1. La reconnaissance de la perception du changement temporel ; 2. La possibilité d'engendrer de nouvelles entités mathématiques (surtout à l'aide de suites infinies). Exemple paradigmatique : les entiers naturels sont, selon Brouwer, générés par la conscience de la succession du temps : ce changement étant « la dissolution d'un moment de vie en deux choses distinctes, l'une cédant la place à l'autre mais étant retenue dans la mémoire. Si la dualité [two-ity] ainsi créée est privée de toutes ses qualités, il reste cependant la forme vide du substrat commun à toutes les dualités. L'intuition mathématique fondamentale réside dans ce substrat commun, cette forme vide »[16].

Or à l'époque toute l'énergie de Hilbert est employée, à Göttingen, à contrer l'intuitionnisme et à structurer le programme finitiste pour répondre à ses objections.

0.3.2. L'argument formel

Prima facie, l'argument nodal que Herbrand oppose à Brouwer est de pure forme. Exemple de Herbrand : prenons ce que nous appelons (depuis sa démonstration en 1995) le théorème de Fermat-Wiles : « Il n'existe pas de nombres entiers strictement positifs x, y et z et $n > 2$ tels que : $x^n + y^n = z^n$ ». Imaginons, dit Herbrand, que ce théorème soit démontrable en utilisant des concepts réprouvés par les intuitionnistes comme Brouwer (typiquement $\mathbb{N}$, l'ensemble des entiers naturels). Si le métamathématicien conclut après analyse que la théorie usitée pour la démonstration est non-contradictoire et complète, il n'a aucune raison d'objecter quoi que ce soit. On peut à la limite concéder à Brouwer que cette preuve ne renvoie à aucune « chose en soi » ; mais en aucun cas on ne peut affirmer que la démonstration n'est pas valable : « Il faut conclure que l'on a le droit de se servir de ces notions interdites puisque tout résultat démontré en les utilisant comme intermédiaires ne peut être faux. Seulement, ces notions devront être considérées par Brouwer comme des éléments sans signification réelle, des éléments idéaux, comme dit Hilbert. »[17].

L'argument intuitionniste est ainsi réduit par les Hilbertiens à sa dimension purement ontologique. Brouwer peut objecter à telle théorie qu'elle n'a pas de « Bedeutung » ; mais cela n'a aucune conséquence pour la théorie de la preuve (et pour toute la logique formelle sur laquelle elle est fondée).

0.3.3. L'argument ontologique

En réalité, cet argument en apparence de pure forme cache une objection plus profonde. Herbrand retourne son principe de neutralité contre la psychologie de Brouwer. Réfuter une théorie des nombres sous prétexte qu'elle ne correspond pas à cette « arithmétique intérieure », à cette succession temporelle interne sur laquelle Brouwer voulait tout refonder, ce serait aussi absurde que de mettre à bas les lois de l'analyse sous prétexte que tel modèle mathématique a échoué à prédire tel phénomène de physique statistique. Critique du psychologisme de Brouwer, Herbrand considère « qu'il y a le même rapport entre une logique formelle et un mode de pensée qu'entre une équation mathématique et un phénomène physique »[18]. Notre arithmétique peut très bien ne pas correspondre à l'arithmétique intérieure de Brouwer ; cela n'enlève rien à sa cohérence interne et à son caractère opératoire.

L'intuitionnisme est ainsi dénoncé comme un dogmatisme, au même titre que le platonisme. Les deux écoles partagent[19] l'idée que l'objet, la chose en soi, posséderait une structure mathématique exogène dont le discours logique devrait rendre compte : « Herbrand décèle ainsi dans les deux doctrines un dogme commun, qu'il récuse sous toutes ses versions : l'idée des mathématiques comme science descriptive, description d'objets indépendants dans un cas, de vécus mentaux dans l'autre. Aucune des deux conceptions ne rend compte de l'objectivité des mathématiques. Cette objectivité ne peut être fondée sur aucune réalité donnée, mentale ou pas »[20] ; elle doit être construite, générée algorithmiquement.

0.3.4. L'argument opératoire

Dernier argument contre Brouwer : son dogmatisme risque de restreindre le champ des outils mathématiques.

Herbrand commence par remarquer[21] que l'histoire des mathématiques a prouvé a bien des reprises que l'intuition est un mauvais guide pour décider de la valeur d'un théorème. Exemple d'Herbrand lui-même ; le célèbre cinquième postulat d'Euclide.

[16] [Brouwer, L.E.J., 1907]
[17] [1930a, 252]
[18] [Chevalley, J., in 1968]
[19] La construction d'un débat articulé entre « platonisme » et « intuitionnisme » est une construction épistémologique postérieure à la mort de Herbrand ; elle a notamment été avancée par [Fraenkel, A.A. 1934] et [Bernays, P., 1934]
[20] [Dubucs, J., Egre, P., 2006]
[21] [1930a]

Cet axiome est resté un fondement de la géométrie pendant deux millénaires, alors même qu'il existait dès l'Antiquité de sérieux doutes sur sa démonstrabilité. Il a fallu attendre les années 1830, et les premières recherches sur les géométries non-euclidiennes (en l'occurrence, les travaux de Lobachevsky et Bolyai sur la géométrie hyperbolique) pour apporter une preuve définitive : tous les postulats d'Euclide peuvent être intégrés dans les géométries non-euclidiennes sauf le 5ème, dont l'introduction peut conduire à une contradiction[22].

On le voit dans cet exemple : l'objectivité mathématique n'est pas description d'une intériorité ; elle est une construction. Herbrand remarque qu'à la limite, le programme de Hilbert, qui reconstruit toute la mathématique à partir des symboles très simples fournis par Russell et Whitehead est beaucoup plus « intuitionniste » que les présupposés de Brouwer : « [Hilbert] s'est imposé de n'employer que des modes de raisonnement si immédiats qu'ils entraînent avec eux la conviction dans tous les cas où on aura à les employer »[23].

Ce mouvement de construction formelle de l'objectivité mathématique, loin d'enfermer le chercheur dans le solipsisme, fonde une objectivité supérieure : « L'objectivité ne s'atteint que dans la symbolique pure, c'est-à-dire en vidant complètement les symboles de toute signification »[24]. Mais ce dévidage ne saurait être compris comme un enfermement, bien au contraire : « L'acte mathématique [n'est] pas une sorte d'acte gratuit. Sans doute est-il possible de faire des mathématiques avec n'importe quels axiomes et n'importe quelles règles de raisonnement, mais en réalité [...] la rigueur a en quelque sorte deux faces complémentaires : si elle est d'abord exigence d'un formalisme, respect des 'règles du jeu', elle est aussi, dans le sens que lui donnait Léonard de Vinci, tentative de description toujours plus parfaite d'un donné »[25].

Ce travail de dévidage des symboles et de construction formelle de l'objectivité est au fondement de la définition herbrandienne des « champs », dont l'exemple princeps est « le domaine de Herbrand » $\mathcal{D}_{H,\infty}$ qui va être maintenant construit.

1. Enoncé et preuve du théorème de Herbrand

1.0. Transformations initiales

1.0.1. Contexte

L'Entscheidungsproblem qui dominait les recherches logiques du temps de Herbrand a depuis été résolu par la négative. Pour la logique propositionnelle, il est possible de construire un algorithme avec un nombre fini d'opérations permettant de savoir si oui une non une formule est une identité (si elle est démontrable). Inversement, pour les logiques d'ordre supérieur, et notamment pour la logique des prédicats, une telle opération n'est pas systématiquement possible ; les formules de la logique de premier ordre sont dites (par assimilation au vocabulaire de Hilbert) « indécidables ».

Cela dit, cette indécidabilité sera démontrée bien après la mort de Herbrand (notamment par Tarski et Turing). Lorsqu'Herbrand publie sa thèse en 1930, il est convaincu avec Hilbert que l'Entscheidungsproblem peut être résolu par la positive. Et le théorème d'Herbrand est censé être une première étape dans la construction d'un algorithme de décision pour les formules de la logique de premier ordre.

En l'occurrence, il s'agit de réduire une formule de logique des prédicats X en une conjonction finie de formules de logique propositionnelle $X_1, X_2, \ldots, X_d$; X est alors insatisfaisable, si et seulement si, un certain nombre de X_i sont logiquement contradictoires.

1.0.2. Formules prénexes

La preuve du théorème implique une première transformation en formule « prénexe ».

Définition 0.1. Une formule de logique de premier ordre est dite « prénexe » si tous les quantificateurs se situent à gauche de l'expression.

Exemple 0.1.

	Transformation prénexe
$Gx \to (\forall x Fx)$	$\forall x(Gx \to Fx)$
$(\exists x Fx) \to Gx$	$\forall x(Fx \to Gx)$
$Gx \vee (\exists x Fx)$	$\exists x(Gx \vee Fx)$

1.0.3. Skolémisation

Seconde transformation nécessaire à la preuve : la « skolémisation », qui transforme toute formule prénexe en formule universelle.

[22] L'indépendance du 5ème postulat par rapport aux autres axiomes a été démontrée formellement en 1868 par Eugenio Beltrami
[23] [1930a]
[24] [Chevalley, J., in (1968)]
[25] [ibid.]

Définition 0.2. – Une formule prénexe de logique de premier ordre est dite « universelle » si elle ne comporte que le quantificateur universel $\forall$.

Notons $\mathcal{G}$ une clause prénexe de la logique des prédicats. Notons $y_1, y_2, \ldots, y_m$ les variables de $\mathcal{G}$ quantifiées existentiellement (introduites à gauche par $\exists$) et de même $x_1, x_2, \ldots, x_n$ les variables quantifiées universellement (introduites par $\forall$). La skolémisation consiste simplement à remplacer chaque y_i par une fonction des variables instanciées universellement telle que : $y_i = f(x_1, x_2, \ldots, x_n)$ (ou par défaut, si $n = 0$, par une constante).

En notant $\mathcal{G}'$ la clause $\mathcal{G}$ privée de ses quantifications, on a la forme :

$$\forall x_1 \forall x_2 \ldots \forall x_n \exists y_1 \exists y_2 \ldots \exists y_m \mathcal{G}'(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$$

Si pour tous $x_1, x_2, \ldots, x_n$, il existe au moins un y_1 telle que la proposition $\mathcal{G}'(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$ est vraie, alors on peut introduire une fonction $y_1 = f(x_1, x_2, \ldots, x_n)$ qui sera telle que $\mathcal{G}'(x_1, x_2, \ldots, x_n, f(x_1, x_2, \ldots, x_n), y_2, \ldots, y_m)$ est vraie.

Exemple 0.2. (tiré des fameuses « descriptions définies de Russell ») :

La proposition : « On appelle roi celui qui dirige un royaume » peut être transcite :

$$\forall x \exists y (Roi(x) \rightarrow (Royaume(y) \wedge Regne(x, y)))$$

Sa forme skolémisée sera, en posant $y = f(x)$

$$\forall x (Roi(x) \rightarrow Royaume(f(x)) \wedge Regne(x, f(x)))$$

1.0.4. Herbrandisation

En réalité, Herbrand (fidèle en cela aux Principia Mathematica qui posent d'abord le quantificateur existentiel $\exists$ pour en dériver l'universel $\forall$ par la formule $\forall x P x \equiv \neg (\exists x \neg P x)$) applique la skolémisation, mais aussi l'opération complémentaire, qui consiste à faire disparaître tous les quantificateurs universels. Cette opération a été nommée en son honneur la « herbrandisation ».

Elle consiste : 1. A remplacer les variables libres par des constantes ; 2. A supprimer tous les quantificateurs universels (ou les négations de quantificateurs existentiels) ; 3. Et à remplacer les variables afférentes par une fonction des variables quantifiées existentiellement.

1.0.5. Cas particulier de la logique modale

C'est cette seconde étape – la skolémisation – qui constitue le problème majeur pour une transcription du théorème de Herbrand en logique modale[26]. La forme skolémisée d'une expression de logique modale, pratiquée avec la méthode que nous venons d'exposer, sera dans la plupart des cas, ambiguë.

Prenons un exemple simple : $\Box \forall x, P x$. La skolémisation (l'herbrandisation ici) est triviale ; il suffit de remplacer x par une fonction des autres variables, et comme il n'y a aucune autre variable, ce sera une constante a. Mais justement, la formule $\Box P a$ est ambiguë par le statut même de la modalité. Dans la logique modale la plus courante, la logique modale K, a doit s'interpréter dans la théorie des « mondes possibles » de Kripke. Prenons un monde p et deux mondes q_1 et q_2 accessibles à partir de p. Mettons que $P a$ soit vraie dans p. $\Box P a$ peut alors être interprétée de deux manières : 1. L'objet désigné par a a la propriété $\Box P$ dans les mondes accessibles q_1 et q_2 ; 2. La formule $P a$ est vraie dans tous les mondes accessibles depuis p, ici, dans q_1 pour l'objet que a désigne dans le monde q_1 et dans q_2 pour l'objet que a désigne en q_2.

En somme $\Box P a$ peut être la formule skolémisée-herbrandisée de deux formes :

1. $\Box \forall x, P x$ (dans le cas où a a un sens restreint)
2. $\forall x, \Box P x$ (dans le cas où a a un sens large)

La solution de M. Fitting[27] consiste à fixer le sens des termes en choisissant la première expression (dont découle toute la logique modale K, puisque Kripke ne retient que le symbole de nécessité, à l'exclusion des symboles de possibilité).

On introduit de plus une notation spécifique pour la skolémisation-herbrandisation $\langle \lambda x.\phi \rangle (t)$, la substitution de t à x dans une formule donnée, équivalent modal de la transformation de $\exists x \forall y R(x, y)$ en $\exists x R(x, f(x))$.

Une fois l'herbrandisation terminée, il ne reste que des quantificateurs universels ; l'expansion de Herbrand d'une formule modale consiste alors à transformer ces universaux en conjonction d'instances de la formule sur les éléments du « domaine de Herbrand » (tel qu'on le construira plus bas). La démonstration en logique modale se ramène alors au cas de la logique générale (à cette nuance près que la logique modale permet de former un plus grand nombre d'instances de base au sens de la définition 1.3., mais cela ne modifie pas la preuve du théorème 1.1.).

[26] [Fitting, M., 1996]
[27] [Fitting, M, 1996, p.3-4]

1.1. Constructions et démonstration

1.1.1. Domaine de Herbrand

Soit un ensemble de formules de logique de premier ordre, noté E.

Définition 1.1. – Le domaine (ou univers) de Herbrand de niveau 0, noté $\mathcal{D}_{H,0}$ est définie comme :

1. L'ensemble des constantes $a_1,\dots,a_n$ apparaissant dans les formules de E (ou à défaut, si $n=0$, une constante créée arbitrairement, a_0)

2. Tous les symboles des m fonctions apparaissant dans les formules de E, notés $f_1,\dots,f_m$ instanciées pour chaque constante $a_1,\dots,a_n$.

On construit par récurrence les ensembles de niveau supérieur :

$$\forall k,\ \mathcal{D}_{H,k+1}=\mathcal{D}_{H,k}\cup\left\{\forall i\in[\![1,n]\!],\forall j\in[\![1,m]\!],f_j^{k+2}(a_i)\right\}$$

On appelle alors « domaine de Herbrand » (ou « univers de Herbrand ») au sens strict, le domaine de niveau infini $\mathcal{D}_{H,\infty}$.

Exemple 1.1.

$$E=\{P(a),\ P(x)\wedge Q(f(x))\}$$
$$\mathcal{D}_{H,0}=\{a,f(a)\}$$
$$\mathcal{D}_{H,1}=\mathcal{D}_{H,0}\cup\{f(f(a))\}=\{a,f(a),f(f(a))\}$$
$$\mathcal{D}_{H,2}=\mathcal{D}_{H,1}\cup\left\{f^3(a)\right\}=\left\{a,f(a),f^2(a),f^3(a)\right\}$$

Base de Herbrand

Définition 1.2. – En injectant tous les éléments du domaine de Herbrand dans les prédicats de E en fonction de leur arité, on obtient la « base de Herbrand », notée $\mathcal{B}_H$:

Exemple 1.2. En reprenant l'exemple 1.1., on aura pour base :

$$\mathcal{B}_H=\{P(a),P(f(a)),P(f(f(a))),\dots,Q(a),Q(f(a)),Q(f(f(a))),\dots\}$$

Définition 1.3. – Une « instance de base » d'une formule F est obtenue en remplaçant dans cette formule les variables libres par les éléments du domaine de Herbrand.

Exemple 1.3. L'instance de base de la formule $Q(x)\wedge R(x,y)$ en $Q(a)\wedge R(a,f^2(a))$ en est un exemple parmi beaucoup d'autres.

Interprétation de Herbrand

Définition 1.4. – On note Σ la signature de E, $\mathcal{B}_0$ un sous-ensemble (choisi librement) de $\mathcal{B}_H$, et $d_1,\dots d_l$ les éléments de $\mathcal{D}_{H,\infty}$. Une interprétation dans $\mathcal{D}_{H,\infty}$ est une « interprétation de Herbrand » si elle attribue la signification suivante aux symboles de la signature :

	$\mathcal{H}$-Interprétation
Constante de la signature a_i	a_i
Symbole de fonction f_j	$\left\{f_j(d_1,\dots d_l)\,\middle|\,d_1,\dots d_l\in\mathcal{D}_{H,\infty}\right\}$
Symbole de relation P	$\{d_1,\dots d_l\,\vert\,(d_1,\dots d_l\in\mathcal{D}_{H,\infty})\wedge(P(d_1,\dots d_l)\in\mathcal{B}_0)\}$
Variable propositionnelle x	1 si et seulement si $x\in\mathcal{B}_0$

C'est là une formalisation contemporaine ; Herbrand lui-même définit cette interprétation de manière plus directe : « Nous dirons que nous avons un système de 'valeurs logiques' dans les champs [le domaine de Herbrand] quand nous aurons donné une valeur logique à toute proposition obtenue en remplaçant les arguments d'une proposition élément [de E] par les éléments des champs. On en déduit la valeur logique de propositions sans variables apparentes ni fonctions descriptives, ne contenant comme variables réelles que les éléments de champs »[28]. En somme, on prend tous les termes de $\mathcal{B}_H$ et on leur attribue librement n'importe quelle valeur sémantique $\{0,1\}$.

Lemme 1.1. Soit r une affectation quelconque des termes de $\mathcal{D}_{H,\infty}$ dans une formule quelconque T. Alors la $\mathcal{H}$-interprétation de T sous cette affectation, notée Tr, est la valeur syntaxique de Tr ; on note :

$$[\![T]\!]_{\mathcal{H},r}=Tr$$

[28][1930, (1968, p. 127)]

Exemple 1.4. Soit la clause $P(x) \wedge \neg P(x)$; le domaine de Herbrand est réduit à une constante générée par défaut a_0 ; l'univers de Herbrand est réduit à $P(a_0)$; il a une seule affectation possible des termes de $\mathcal{D}_{H,\infty}$ et quelle que soit l'interprétation de Herbrand (que $P(a_0)$ valle 1 ou 0), la clause $P(x) \wedge \neg P(x)$ rend la valeur 0.

Lemme 1.2. – Soit une interprétation | quelconque sur une base de Herbrand ; alors une existe une $\mathcal{H}$-interprétation j telle que pour toute formule et pour toute permutation des termes, ces deux interprétations coïncident.

Preuve du lemme 1.2. – On choisit j' telle que l'ensemble des formules closes vraies dans j le soient aussi dans j'. On a alors pour toute formule A les valorisations suivantes :

$$[\![A]\!]_{j',r} = [\![Ar]\!]_{j'} = [\![Ar]\!]_j$$

La première égalité est l'application simple du lemme 1.1. La seconde découle de notre choix initial pour j', que nous avons construite telle que toutes les instances closes aient par elle la même valeur que par j. On déduit $[\![\,]\!]_j = [\![\,]\!]_j$ par récurrence sous les composantes internes de A.

Lemme 1.3. – (Corrélat direct du lemme 1.1.) – Un ensemble de formules faux sous toutes les $\mathcal{H}$-interprétations l'est aussi sous toutes les interprétations en général.

Enoncé du théorème de Herbrand

Définition 1.5. – Un ensemble de formules E est dit « insatisfaisable » si et seulement si la conjonction de toutes ces formules rend la valeur logique 0 quelle que soit l'interprétation choisie.

Définition 1.6. – (Définition de la « satisfaisabilité » par la négation de la définition 1.5.)

Propriété 1.1. – (déduite du lemme 1.3.) Une formule ε est satisfaite par une $\mathcal{H}$-interprétation $\mathcal{I}$ si et seulement si toute instance de base est satisfaite par $\mathcal{I}$.

Propriété 1.2. – (négation de la propriété 1.1.) Une formule ε est falsifiée par une $\mathcal{H}$-interprétation $\mathcal{I}$ si et seulement il existe au moins une instance de base qui soit fausse par $\mathcal{I}$.

Propriété 1.3. – (généralisation de la propriété 1.2.) Un ensemble de formules E est insatisfaisable si et seulement si, par toutes les $\mathcal{H}$-interprétations, il y a au moins une instance de base ou une clause qui soit fausse.

Théorème 1.1. (Théorème de Herbrand) – Soit E' une proposition prénexe skolémisée $\forall x_1 \ldots \forall x_n \mathcal{E}(x_1, \ldots, x_n)$ où $\mathcal{E}$ ne contient aucun quantificateur. Alors E' est insatisfaisable si et seulement il existe un ensemble fini insatisfaisable d'instances de bases des formules de $\mathcal{E}$ sur des éléments de la signature de $\mathcal{E}$.

Preuve du théorème

Preuve du théorème 1.1.

Sens réciproque – S'il existe un ensemble fini d'instances de base de $\mathcal{E}(x_1, \ldots, x_n)$ sur des éléments de la signature de $\mathcal{E}$ qui rend 0 quelle que soit l'interprétation choisie, alors $\forall x_1 \ldots \forall x_n \mathcal{E}(x_1, \ldots, x_n)$ rendra la valeur 0 quelle que soit l'interprétation choisie.

Sens direct – $\forall x_1 \ldots \forall x_n \mathcal{E}(x_1, \ldots, x_n)$ est supposée ici insatisfaisable, rendant la valeur 0 quelle que soit l'interprétation choisie. La définition 1.5., corrélat du lemme 1.2.„ permet de réduire l'analyse, non plus à toute interprétation possible, mais à toute $\mathcal{H}$-interprétation sur le domaine de Herbrand de E', i.e. à une application de la base de Herbrand de E' dans $\{0,1\}$. La forme $\forall x_1 \ldots \forall x_n \mathcal{E}(x_1, \ldots, x_n)$ peut alors être réduite à une conjonction d'instances de base de $\mathcal{E}$ avec les éléments du domaine de Herbrand $\mathcal{E}((d_1, \ldots, d_l)\, r_1) \wedge \mathcal{E}((d_1, \ldots, d_l)\, r_2) \wedge \ldots$ Si $\forall x_1 \ldots \forall x_n \mathcal{E}(x_1, \ldots, x_n)$ est insatisfaisable sous toute interprétation, alors la conjonction d'instances de base que nous avons créée est fausse pour toute $\mathcal{H}$-interprétation, ce qui veut qu'une instance de base $\mathcal{E}((d_1, \ldots, d_l)\, r_i)$ ou qu'une conjonction finie de plusieurs instances de base est fausse.

Exemples d'application

Exemple 1.5. (Formule décidable insatisfaisable)

Formule	$E' = \forall x\,(P(x), Q(x), \neg P(a) \vee \neg Q(b))$
Formule privée des universaux	$\mathcal{E}(x) = P(x), Q(x), \neg P(a) \vee \neg Q(b)$
Signature	$\Sigma_{E'} = \{a, b, P, Q\}$
Domaine de Herbrand	$\mathcal{D}_{H,\infty_{E'}} = \{a, b\}$
Base de Herbrand	$\mathcal{B}_H = \{P(a), Q(a), P(b), Q(b)\}$
Enumération des instances de base de $\mathcal{E}(x)$ avec les éléments de $\mathcal{D}_{H,\infty_{E'}}$	• $P(a), Q(a), \neg P(a) \vee \neg Q(b)$ n'est pas contradictoire • $P(b), Q(a), \neg P(a) \vee \neg Q(b)$ n'est pas contradictoire • $P(a), Q(b), \neg P(a) \vee \neg Q(b)$ est une contradiction • $P(b), Q(b), \neg P(a) \vee \neg Q(b)$ n'est pas contradictoire
Conclusion	Par le théorème de Herbrand, E' est insatisfaisable

Parfois, il faut, pour enclencher le théorème de Herbrand, former un ensemble de plusieurs instances de base, comme dans l'exemple suivant :

Exemple 1.6. (Formule décidable insatisfaisable – énumération non finie)

Formule	$E' = \forall x \left(P(x) \vee \neg P\left(f(x)\right), \neg P(a), P\left(f^2(a)\right)\right)$
Formule privée des universaux	$\mathcal{E}(x) = P(x) \vee \neg P\left(f(x)\right), \neg P(a), P\left(f^2(a)\right)$
Signature	$\Sigma_{E'} = \{a, P, f\}$
Domaine de Herbrand	$\mathcal{D}_{H,\infty_{E'}} = \{a, f(a), f^2(a), \ldots\}$
Base de Herbrand	$\mathcal{B}_H = \{P(a), P\left(f(a)\right), \ldots\}$
Enumération des instances de base de $\mathcal{E}(x)$ avec les éléments de $\mathcal{D}_{H,\infty_{E'}}$	• $P(a) \vee \neg P\left(f(a)\right), \neg P(a), P\left(f^2(a)\right)$ • $P\left(f(a)\right) \vee \neg P\left(f(a)\right), \neg P(a), P\left(f^2(a)\right)$ En soi, ces deux premières instances ne sont pas contradictoires, mais leur conjonction l'est
Conclusion	Par le théorème de Herbrand, E' est insatisfaisable

Exemple 1.7. (Formule décidable satisfaisable)

Formule	$E' = \forall x \left(P(x) \vee Q(x), \neg P(a), \neg Q(b)\right)$
Formule privée des universaux	$\mathcal{E}(x) = P(x) \vee Q(x), \neg P(a), \neg Q(b)$
Signature	$\Sigma_{E'} = \{a, b, P, Q\}$
Domaine de Herbrand	$\mathcal{D}_{H,\infty_{E'}} = \{a, b\}$
Base de Herbrand	$\mathcal{B}_H = \{P(a), Q(a), P(b), Q(b)\}$
Enumération des instances de base de $\mathcal{E}(x)$ avec les éléments de $\mathcal{D}_{H,\infty_{E'}}$	• $P(a) \vee Q(a), \neg P(a), \neg Q(b)$ n'est pas contradictoire • $P(a) \vee Q(b), \neg P(a), \neg Q(b)$ n'est pas contradictoire • $P(b) \vee Q(a), \neg P(a), \neg Q(b)$ n'est pas contradictoire • $P(a) \vee Q(b), \neg P(a), \neg Q(b)$ n'est pas contradictoire
Conclusion	Par le théorème de Herbrand, E' est satisfaisable

On utilise ici la négation du théorème de Herbrand. Mais on remarque immédiatement que cet usage n'est valable que si l'énumération est finie. On peut facilement bien imaginer une formule dont l'énumération des instances de base serait infinie, et où un logiciel ne trouverait aucune contradiction dans les 100.000 premières instances et toutes leurs combinaisons par intersections. Cette formule serait alors indécidable.

Cette indécidabilité était conçue par Herbrand comme purement provisoire. L'algorithme construit ici ne permettait pas de décider pour toute formule de la logique des prédicats. Mais dans l'esprit d'Herbrand, les recherches futures, axées sur la méthode de Hilbert, étaient censées fournir un algorithme complet de décidabilité. On sait aujourd'hui grâce à Gödel que ce n'est pas possible, et que paradoxalement le théorème de Herbrand, plus général que le pensait son auteur, prouve cette indécidabilité, fournissant ainsi une démonstration alternative au célèbre premier théorème d'incomplétude.

Du théorème de Herbrand découlent ainsi deux corrélats :

Proposition 1.1. – Il n'existe pas d'algorithme décidant de la validité des formules du premier ordre.

Proposition 1.2. – L'ensemble des formules valides est récursivement énumérable[29].

Conclusion – Le premier théorème d'incomplétude comme rupture symbolique

Ce que Herbrand prenait donc pour une simple limite contingente de son théorème, limite censée être bientôt dépassée par la recherche, était en fait la preuve d'une limitation inhérente à toute théorie cohérente, ce que montrera K. Gödel l'année même de la mort de Herbrand[30]. Herbrand prend acte de la découverte majeure dans un de ses derniers textes : « Il est impossible de démontrer la non-contradiction d'une théorie par des raisonnements formalisables dans cette théorie, quand cette théorie contient l'arithmétique »[31]. Il donne sa propre version du théorème d'incomplétude en construisant une proposition de la forme $P\,x\,y\,z$ signifiant « La démonstration numéro x démontre la y-ième proposition pour la valeur z de sa variable ». Si on suppose que la théorie est non-contradictoire et cohérente, si on note β le numéro de la proposition suivante : $\forall x, \neg P\,x\,y\,y$: « Pour tout x, aucune démonstration ne peut prouver la proposition y pour la valeur y de sa variable » ; supposons alors que $P\,x\,\beta\,\beta$ soit vraie ; on en conclut immédiatement que $\forall x, \neg P\,x\,\beta\,\beta$; on pourrait alors montrer dans la théorie la proposition $\neg P\,x\,\beta\,\beta$ et on arriverait à une contradiction.

Généralement présenté comme la ruine du projet finitiste de Hilbert, la preuve du premier théorème d'incomplétude peut au contraire être lue comme une confirmation de l'extrême finesse des intuitions de J. Herbrand ; au-delà de

[29] On ne prouvera pas ces deux points ; la démonstration date des années 1940 ; on en trouvera un exposé détaillé dans [Goubault-Larrecq & Mackie, 1997, p.204-205]

[30] [Gödel, K., 1931]

[31] [1931c]

Gödel, il faudra attendre les travaux les plus éminents de l'histoire de la mathématique informatique (Turing, Tarski) pour obtenir une preuve rigoureuse des nombreux corrélats du théorème de Herbrand.

Bibliographie

Sources – Ecrits de Herbrand

[1928] « Sur la théorie de la démonstration », *CRAS*, vol. 186, p. 1274-1276

[1929] « Non-contradiction des axiomes arithmétiques », *CRAS*, vol. 188, p. 303-304

[1929a] « Sur quelques propriétés des propositions vraies et leurs applications », *CRAS*, vol. 188, p. 1076-1078

[1929b] « Sur le problème fondamental des mathématiques », *CRAS*, vol. 189, octobre 1929, p. 554-556

[1930] *Recherches sur la théorie de la démonstration* : Thèses présentées à la faculté des sciences de Paris, Paris, 128 p.

[1930a] « Les bases de la logique hilbertienne », *Revue de métaphysique et de morale*, vol. 37, no 3, 1930, p. 243-255

[1931] « Sur le problème fondamental de la logique mathématique », *Sprawozdania z posiedzenow Towarzystwa Naukowego Warzsawskiego, Wydzial III Nauk-Matematyczno-fizycznych*, 24

[1931a] Note non signée sur *Herbrand 1930, Annales de l'université Paris VI*, 186-189

[1931b] Notice pour Jacques Hadamard (inédit in [1968])

[1931c] « Sur la non-contradiction de l'Arithmétique », *Journal für die reine und angewandte Mathematik*, vol. 166, no 1, p. 1-8

[1968] *Ecrits logiques* (avec une note de Claude Chevalley, « Sur la pensée de Jacques Herbrand ») ; sauf mention contraire, les textes de Herbrand sont tous cités dans cette édition.

Sources – Alia

Bernays, P., Hilbert, D., [1939] Grundlagen der Mathematik

Bernays, P., [1936] « Sur le platonisme dans les mathématiques », L'enseignement mathématique, 34, p. 52-69

Brouwer, L.E.J. [1907] Over de grondslagen der wiskunde

Chevalley [1936] « Sur la pensée de Jacques Herbrand », L'enseignement mathématique, 34, p. 97-102 (cité dans [1968]).

Fraenkel, A.A., [1936] « Sur la notion d'existence dans les mathématiques », L'enseignement mathématique, 34, p. 52-69

Gödel, K., [1931], « Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme », *Monatshefte für Mathematik und Physik*, 38

Hilbert, D., [1922] « Die logische Grundlagen der Mathematik », Mathematische Annalen 88

Peano, G., [1889], Arithmetices principia: nova methodo exposita,

Russell, B., Whitehead, A.N., [1910], Principia Mathematica, Cambridge University Press

Littérature secondaire

Dubucs, J., Egre, P., [2006] « Jacques Herbrand » in Bitbol, M., Gayon, J., L'Epistémologie française 1830-1970, P.U.F.

Goubault-Larrecq, J., Mackie, I. [1997]. Proof Theory and Automated Deduction, volume 6 of Applied Logic Series, Kluwer Academic Publishers

Goubault-Larrecq, J., Mackie, I., [2003] « Démonstration automatique des théorèmes / Logique classique du premier ordre », Notes de cours.

Fitting, Melvin. [1996]. A Modal Herbrand Theorem. Fundam. Inform.. 28. 101-122. 10.3233/FI-1996-281206.

Paulin, C., [2012], Eléments de logique pour l'informatique, notes de cours, 54 p.

Sieg, Wilfried [1991] Herbrand analyses. Arch. Math. Logic 30, 409–441

Sieg, Wilfried [2005] Only two letters: The correspondence between Herbrand and Gödel, Bulletin of Symbolic Logic 11 (2):172-184.