Déflationnisme et conservativité : quelqu'un a-t-il changé de sujet ?

Henri Galinon

Déflationnisme et conservativité : quelqu'un a-t-il changé de sujet ?

Henri Galinon
Université Paris 1, CNRS, ENS (France)

Résumé : Nous clarifions et critiquons un argument influent opposé au déflationnisme par [Shapiro 1998b] et [Ketland 1999], fondé sur la non-conservativité des extensions des théories formalisées par des principes aléthiques universellement admis. À cette interprétation anti-déflationniste des phénomènes de non-conservativité, nous en opposons une autre, compatible à la fois avec les faits logiques et les thèses déflationnistes.

Abstract: [Shapiro 1998b] and [Ketland 1999] have argued against deflationary views of truth on the ground that an adequate truth-theoretic extension of a theory is a non-conservative extension. We clarify the argument and offer an alternative interpretation of the observed non-conservativeness phenomenon, compatible both with the logical facts and the deflationist's thesis.

1 Introduction

Le déflationnisme [1] en matière de vérité est porté par une double intuition. D'une part, l'intuition de ce que l'on appelle parfois la *transparence* de la notion de vérité, à savoir l'idée qu'un énoncé et l'énoncé qui lui attribue la vérité expriment une même proposition — une idée que l'on trouve déjà en substance chez Frege [2]. D'autre part, l'intuition de l'*indispensabilité* de la notion

Philosophia Scientiæ, 16 (3), 2012, 1–19.

1. Nous remercions François Rivenc, Leon Horsten et un relecteur anonyme pour leurs commentaires et leurs suggestions.

2. Par exemple dans le passage suivant :

> Il vaut aussi de remarquer que la proposition : « Je sens une odeur de violette » a même contenu que la proposition : « Il est vrai que je sens une odeur de violette. » Il semblerait que rien n'est ajouté à la pensée quand je lui attribue la propriété d'être vraie. [...] La dénotation du mot « vrai » semble unique en son genre. Serait-ce que nous ayons affaire à

de vérité à certaines fins expressives. En effet, si nous ne pouvons pas *dire* une infinité d'énoncés, nous pouvons parfois *décrire* ou *désigner* cet ensemble infini d'énoncés (par exemple, par la description définie « les théorèmes de l'arithmétique ») et, en attribuant la vérité aux énoncés de cet ensemble (« Tous les théorèmes de l'arithmétique sont vrais »), dire en substance, par une voie détournée (la montée sémantique), ce que notre finitude nous interdit de dire directement : en parlant des énoncés, nous continuons à « parler du monde [3] ». La thèse déflationniste est alors que la notion de vérité n'est pas une *notion explicative* — une notion qui ne joue pas de rôle « substantiel » dans nos explications — mais *seulement* un outil « expressif », dont le rôle *ne* serait *que* de permettre d'exprimer certaines généralisations. Si cette thèse est correcte, les principes fondamentaux qui gouvernent notre compréhension de la notion de vérité, quels que soient ces principes, sont mobilisables dans le discours scientifique légitime à la façon d'auxiliaires indispensables pour l'expression de certains faits, hypothèses, ou autres, mais ils n'en constituent jamais la clé de voûte explicative et ne nous apprennent rien sur « le monde ». Or cette thèse soulève immédiatement une question logique concernant la cohérence interne du déflationnisme. On peut la formuler ainsi : une *théorie de la vérité* est-elle seulement possible, qui soit capable à la fois de rendre compte des usages de la notion qu'un déflationniste juge essentiels, en particulier, donc, de son usage comme moyen d'expression de certaines généralisations, *et* qui soit compatible avec la thèse déflationniste selon laquelle la notion de vérité n'est pas une notion explicative ?

Il est clair que toute tentative de réponse à cette question devra en même temps formuler deux propositions visant à en préciser les termes. Il faudra d'abord proposer une explication, au sens de Carnap, de la distinction entre « notion explicative » et « notion purement expressive » et, d'autre part, préciser ce que sont ces « usages » de la notion de vérité dont une théorie de la vérité doit rendre compte. Par conséquent, l'évaluation philosophique de la cohérence interne du déflatonnisme s'articule autour, non pas d'un problème unique, mais de trois problèmes :

Le Problème de la stabilité : La thèse de l'absence de rôle explicatif du prédicat de vérité est-elle compatible avec la thèse concernant son rôle expressif ?

C'est le problème principal, celui de la cohérence interne du déflationnisme.

Le Problème de la frontière : Proposer une explication de la distinction entre « notion explicative » et « notion purement expressive ».

---

quelque chose qui ne peut nullement être appelé propriété dans le sens usuel ? [Frege 1971, « La pensée », 174].

3. Le développement classique de ce point de vue se trouve dans [Quine 1970, chap. 2].

C'est la première condition d'une réponse au Problème de la stabilité. Une bonne réponse au Problème de la frontière doit respecter les intuitions déflationnistes fondamentales si elle doit être utilisée pour en étudier la cohérence interne.

Le Problème de l'adéquation : De quels usages de la notion de vérité une théorie de la vérité doit-elle rendre compte ?

On peut voir ce problème comme celui de la recherche d'un critère d'adéquation des théories de la vérité qui vienne compléter le critère classique de Tarski. Le critère d'adéquation de Tarski assure que le prédicat de vérité théorisé saisit la notion de vérité comme correspondance, en assurant l'assertabilité des équivalences-T [4]. La seconde condition d'adéquation doit garantir un certain nombre d'autres usages de la notion de vérité jugés essentiels. Le Problème de l'usage est de dire ce qu'ils sont.

Dans la suite de cet article, je me propose de reconstruire, puis de critiquer, dans le cadre d'analyse que je viens de proposer, un argument influent opposé au déflationnisme par Stewart Shapiro et Jeffrey Ketland, « l'argument de la conservativité [5] ». Le passage suivant de [Shapiro 1998b] donne un aperçu de la façon dont les phénomènes d'incomplétude arithmétique sont mobilisés dans l'argument de la conservativité pour réfuter le déflationnisme :

> Retournons à notre théorie arithmétique A et à son énoncé de Gödel G (ou Coh) [6]. Supposons qu'un professeur de logique asserte que G est vrai, et qu'un étudiant désemparé demande une explication. L'étudiant croit l'affirmation du professeur selon laquelle G est vrai, mais il veut qu'on lui montre pourquoi cet énoncé est vrai. L'étudiant veut quelque chose comme une preuve convaincante ou une preuve explicative. La réponse naturelle est de faire remarquer que tous les axiomes de A sont vrais et que les règles d'inférence préservent la vérité. Il suit que « 0=1 » n'est pas un théorème

4. C'est-à-dire de la forme :

$$Vr(s) \leftrightarrow p.$$

où il faut remplacer « p » par un énoncé et « s » par un nom de cet énoncé.

5. Cet argument a été développé indépendamment par les deux auteurs à peu près à la même époque dans [Shapiro 1998b] et [Ketland 1999], dont c'est le thème principal. L'argument a donné lieu à de nombreuses discussions dans des articles ultérieurs, par exemple [Tennant 2002], [Halbach & Horsten 2002]. La réponse qu'a présentée [Field 1999] est le point de départ de notre propre réponse (voir la dernière note du présent article). Comme l'a noté à juste titre un relecteur anonyme, la question traitée ici est liée au problème plus général de l'interprétation épistémologique des théorèmes d'incomplétude de Gödel. Pour des raisons de place, nous nous en tiendrons à la discussion de l'argument de Shapiro et Ketland.

6. N.d.T. : « Coh » est un énoncé du langage de A qui exprime la cohérence de A. Nous reviendrons sur ce point dans la suite.

> et donc que A est cohérente. L'énoncé de Gödel est équivalent à la cohérence de A. Il me semble que cette version informelle de la dérivation de Coh et G est une bonne *explication* s'il en est. L'argument montre pourquoi G est vrai. [...] Notre étudiant veut savoir pourquoi G est vrai — ou pourquoi G est une conséquence — et le passage par la notion de vérité fournit cette explication. [Shapiro 1998b, 505, ma traduction]

Cette brève mise en scène de la réponse du Professeur à l'étudiant ingénu contient le noyau des réponses de Shapiro et Ketland aux Problèmes de la frontière et de l'adéquation. Parce que le raisonnement précédent est naturel, une théorie de la vérité adéquate doit permettre d'en rendre compte. Or l'énoncé de la cohérence de A est un énoncé du langage de A qui n'était pas dérivable de A seulement ; on a donc fait un usage explicatif de la notion de vérité. J'appellerai le raisonnement précédent la *preuve de la cohérence par la vérité*.

Le plan de l'article sera le suivant. Dans une première partie, je propose une reconstruction rationnelle de l'argument général développé par Shapiro et Ketland. Puis j'attaquerai l'argument sur ce que j'appellerai la *thèse de la conservativité*. Je soutiens que les phénomènes de non-conservativité observés à propos des théories de la vérité sont non seulement compatibles avec les thèses déflationnistes, mais en parfaite harmonie avec elles.

2 L'argument de la conservativité

La forme générale de l'argument est la suivante [7] :

Argument Principal :

1. Une théorie adéquate de la vérité doit être réflexive.
2. Une théorie déflationniste de la vérité doit être conservative.
3. Une théorie réflexive de la vérité n'est pas conservative.
4. Donc les théories déflationnistes de la vérité ne sont pas adéquates.

Les deux premières prémisses de l'argument sont des thèses philosophiques. J'appellerai la première la thèse de la réflexion et la seconde la thèse de la conservativité . La troisième prémisse est un fait de logique. Je prends les trois points dans cet ordre.

7. Je remercie Jeffrey Ketland, dont les remarques m'ont aidé à clarifier le sens exact qu'il entend donner à l'argument de la conservativité.

2.1 La thèse de la réflexion

La thèse de la réflexion est la réponse partielle de Shapiro et Ketland au Problème de l'adéquation, à la question de savoir ce dont une théorie de la vérité *doit* rendre compte. Une des raisons pour lesquelles Tarski jugeait insuffisante la théorie minimale de la vérité, cette théorie qui se réduit aux équivalences-T, était qu'il n'est pas possible dans l'extension aléthique minimale d'une théorie de dériver la moindre généralisation concernant la vérité [8]. En particulier, dans l'extension aléthique minimale d'une théorie, on ne peut *prouver* que les *instances* du principe de non-contradiction, mais non la loi générale elle-même. Cette remarque pourrait naturellement donner lieu à une nouvelle condition d'adéquation pour les théories de la vérité, sous la forme d'une « thèse de la généralisation », quelque chose comme :

Thèse de la généralisation :

> Une théorie de la vérité doit permettre de prouver les généralisations dont les instances sont toutes des conséquences de la condition d'adéquation de Tarski, c'est-à-dire des équivalences-T.

Mais ce n'est pas la thèse que retiennent Shapiro et Ketland. La condition d'adéquation qu'ils proposent est plus forte. L'idée est qu'une théorie de la vérité doit encore être *réflexive* :

Thèse de la réflexion :

> Dans une extension aléthique adéquate d'une théorie A, il doit être possible de prouver : « Tous les théorèmes de A sont vrais [9]. »

Pourquoi l'extension d'une théorie A par une théorie du concept de vérité (et la syntaxe) *devrait*-elle en droit permettre de dériver la vérité de la théorie de départ ? Quel est le sens de cette exigence ? L'argument principal en faveur de la thèse de la réflexion est un argument de *fidélité à l'usage ordinaire*.

L'argument de la réflexion :

> C'est un fait, pourrait-on soutenir, que si un sujet accepte une théorie A et possède un concept de vérité, alors il est en position de conclure que tous les théorèmes de A sont vrais [10]. C'est ce qu'est supposé illustrer (entre autres) le bref scénario de Shapiro présenté plus haut : un sujet accepte la théorie A, possède un concept de vérité, et il est capable d'inférer que tous les axiomes, puis tous les théorèmes de A sont vrais. L'idée est donc que si ce scénario représente correctement un usage ordinaire de la notion de vérité, la fidélité aux attributions ordinaires de

8. J'appelle extension aléthique minimale d'une théorie A la théorie obtenue en étendant A par les équivalences-T (pour le langage de A) et les axiomes spécifiques standard permettant de décrire la syntaxe du A.

9. Les énoncés de ce type sont connus dans la littérature logique comme des *Principes de réflexion*. Nous préciserons parfois « Principe de réflexion sur A » pour indiquer sur quelle théorie il s'agit de « réfléchir ».

10. On suppose que A elle-même contient sa propre syntaxe.

> vérité commande que d'une théorie A et d'une théorie de la vérité pour le langage de A nous puissions dériver : « Tous les théorèmes de A sont vrais. »

De même que la convention-T était pour Tarski une condition d'adéquation permettant de garantir qu'une théorie de la vérité qui la satisfait saisit bien la notion classique de vérité comme correspondance, l'usage crucial rendant compte de la signification de « vrai » étant identifié alors à l'assertabilité des équivalences-T, de même ici, un usage jugé central du concept de vérité vient fonder la nouvelle condition d'adéquation [11].

2.2 La thèse de la conservativité

De même que la thèse de la réflexion est une réponse au Problème de l'adéquation, la seconde hypothèse de l'argument principal est une réponse au Problème de la frontière. La proposition de Ketland et Shapiro peut être formulée de façon suivante :

Thèse de la conservativité :

> La vérité est une notion explicative si et seulement s'il existe une théorie A, telle que l'extension aléthique de A étend *non-conservativement* A [12].

La notion de conservativité étant définie comme suit :

Définition 1 (Conservativité)**.** Soit T une théorie formulée dans un langage L, et T' une extension de T formulée dans un langage L' tel que $L \subset L'$. T' est conservative sur T si et seulement si, pour tout énoncé ϕ du langage L, s'il existe une preuve de ϕ dans T', il existe une preuve de ϕ dans T.

Pourquoi la thèse de la conservativité est-elle de prime abord plausible ? Supposons qu'il existe une théorie A ne contenant pas de termes sémantiques et telle que l'extension aléthique de A étend non-conservativement A : alors il y a un fait non-sémantique qui peut être décrit dans le vocabulaire de L_A par

11. Remarque : à soi seule la thèse de la réflexion implique que la théorie minimale de la vérité n'est pas adéquate (alors que la théorie récursive, ou tarskienne, l'est). Mais le déflationnisme est en droit indépendant de la thèse selon laquelle l'ensemble des équivalences-T constitue une théorie de la vérité adéquate. Voir par exemple [Field 2001] sur ce point.

12. Quelques précisions. Par « extension aléthique de A », il faut entendre la théorie A augmentée d'une théorie de la vérité adéquate pour rendre compte de tous les usages du terme « vrai » jugés essentiels. Par ailleurs, le passage d'un critère du caractère « substantiel » d'une *théorie relativement à une autre théorie* à un critère portant sur la *notion* de vérité elle-même, s'il n'est jamais explicité, peut être simplement reconstruit comme suit : la vérité est une notion explicative si et seulement si une théorie adéquate de la vérité est « substantielle », autrement dit si et seulement s'il y a une théorie A qu'elle étend non-conservativement.

un énoncé ϕ, qui possède une explication dans la théorie étendue, mais qui ne peut être expliqué à partir des seuls principes couchés dans la théorie A elle-même. La possibilité d'expliquer un fait non-sémantique à partir de principes sémantiques (aléthiques, en l'espèce), alors que ce fait est autrement inexplicable, fournirait un contre-exemple à la thèse déflationniste d'après laquelle la vérité n'a pas de pouvoir d'explication [13]. Réciproquement — mais cette partie ne nous retiendra pas ici — la conservativité de l'extension aléthique indiquerait que la vérité n'est pas véritablement « substantielle ».

2.3 La situation logique

À ce point, nous sommes presque à nos fins. Pour conclure l'argument, Shapiro et Ketland ajoutent que c'est une question de logique qu'une théorie réflexive de la vérité étend de façon non conservative certaines théories. Considérons une théorie A « contenant » sa propre syntaxe. Pour fixer les idées, nous suivrons Shapiro et Ketland et supposerons que A est l'arithmétique de Peano en premier ordre (PA). Supposons également que la syntaxe de A soit faite dans A *via* un codage ; la cohérence d'une telle théorie peut s'exprimer par un énoncé de son langage [14], et Gödel a prouvé qu'un tel énoncé n'était pas prouvable dans la théorie elle-même (second théorème d'incomplétude). Mais il est facile de voir que dans une extension aléthique réflexive de PA, on peut prouver la cohérence de la théorie de base, comme le montrait l'argument informel présenté en introduction. De façon un peu plus détaillée, l'argument est le suivant :

1. L'extension aléthique de PA prouve « Tous les théorèmes de PA sont vrais »
   (Par hypothèse de réflexivité),
2. Or « $\neg(0 = 1)$ » est un théorème de PA,
3. et l'on peut prouver dans PA l'énoncé « $\ulcorner\neg(0 = 1)\urcorner$ est un théorème de PA ».
   (Par 2 et une propriété de la syntaxe, qui fait partie de PA par hypothèse. On peut en effet montrer que si σ est un théorème de PA, alors « $\ulcorner\sigma\urcorner$ est un théorème de PA » est un théorème de la syntaxe.)

13. Comme l'avait déjà montré Tarski, l'extension aléthique minimale d'une théorie est justement une extension conservative de cette théorie (voir le mémoire de Tarski sur la vérité, [Tarski 1983, § 5, théorème III]. Plus exactement, le théorème III énonce la cohérence de l'extension d'une théorie par les équivalences-T relativement à cette théorie elle-même. Néanmoins, la preuve que donne Tarski établit en fait la conservativité de cette extension).

14. Le sens exact de l'expression « la cohérence d'une telle théorie peut s'exprimer par un énoncé de son langage » sera reconsidéré plus tard.

4. Donc « $Vr(\ulcorner\neg(0 = 1)\urcorner)$ » est un théorème de l'extension aléthique de PA.
   (Par 1 et 3)
5. Donc « $\neg Vr(\ulcorner 0 = 1\urcorner)$ » est un théorème de l'extension aléthique de PA.
   (Par 4 et adéquation du prédicat de vérité au sens de Tarski [15])
6. Donc « $\ulcorner 0 = 1\urcorner$ n'est pas un théorème de PA » est un théorème de l'extension aléthique de PA.
   (Par 1 et 5)
7. Autrement dit, on peut prouver dans l'extension aléthique de PA que PA est cohérente.

Il s'ensuit que les théories adéquates de la vérité ne sont pas conservatives, et par conséquent que le déflationnisme est faux. Tel est l'argument opposé au déflationniste [16].

3 La thèse de la conservativité : quelqu'un a-t-il changé de sujet ?

Je pense que l'argument présenté dans la section précédente est incorrect à plusieurs titres [17]. Néanmoins, mon intention ici est de me concentrer uniquement sur l'évaluation de la thèse de la conservativité comme réponse au Problème de la frontière. Je voudrais essayer de convaincre le lecteur que le phénomène logique mis en avant par Shapiro est Ketland illustre un cas de non-conservativité *épistémologiquement neutre*, un phénomène sur lequel on ne peut fonder aucun argument en faveur de l'idée que le prédicat de vérité joue un rôle *explicatif*. Avant de présenter mon argument proprement dit, je propose de commencer par une petite expérience de pensée.

3.1 Ethnologie

Imaginons qu'une civilisation très semblable à la nôtre se soit développée à notre insu sur une autre planète. Les terri*a*ns (c'est ainsi qu'ils se nomment

15.
$$\dfrac{\dfrac{Vr(\ulcorner\neg(0 = 1)\urcorner)}{\neg(0 = 1)} \qquad \dfrac{[Vr(\ulcorner 0 = 1\urcorner)]^1}{0 = 1}}{\dfrac{\bot}{\neg Vr(\ulcorner 0 = 1\urcorner)}\,1}$$

16. Dans la suite de l'article, j'adopterai le cadre de discussion de Shapiro, Ketland en supposant que toutes les théories dont nous parlons (théorie-objet, syntaxe, extension aléthique) sont des théories du *premier ordre*.

17. Pour un développement plus complet, je me permets de renvoyer le lecteur à Galinon [2010].

et que nous les nommerons) sont identiques à nous à de nombreux égards, mais en mathématiques ils en sont venus à reconnaître, au-delà de l'arithmétique, l'intérêt d'un domaine connexe, l'arithm*a*tique. Pour une raison qui nous échappe, ces nouveaux pythagoriciens pensent que les n*a*mbres naturels sont les blocs élémentaires de l'univers et, pour cette raison même, leur étude leur importe au plus haut point. Une axiomatisation partielle de l'arithm*a*tique est donnée par $PA \bigcup \{\neg coh_{PA}\}$, où $\neg coh_{PA}$ désigne la négation de l'énoncé *du langage de* PA exprimant la cohérence de PA sous le codage standard de Gödel (cet énoncé est vrai dans $\mathbb{N}$ si et seulement si PA n'est pas cohérent [18]). Des axiomes supplémentaires ont été proposés pour renforcer l'axiomatisation de l'arithm*a*tique, mais ils sont actuellement très controversés et nous les laisserons donc de côté. Pour en rester à l'ethnologie, les terri*a*ns n'utilisent la plupart du temps qu'une axiomatisation très partielle de l'arithm*a*tique lorsqu'ils étudient les n*a*mbres naturels, et cette axiomatisation partielle correspond formellement à PA ; de plus, ils font usage des mêmes conventions et formalismes que nous en matière de logique et, plus admirable encore, ils utilisent eux-mêmes le nom « PA ». En fait, dans leur langage, « PA », les numéraux (« 0 », « 1 », ...) et les symboles d'opérations « successeur », « + », « . » sont ambigus ; ils désignent tantôt une axiomatisation de l'arithmétique, des nombres entiers et les opérations bien connues sur leur domaine, tantôt, respectivement, une axiomatisation partielle de l'arithm*a*tique, les éléments du segment initial des n*a*mbres naturels, et les opérations moins familières qui leurs sont associées ; dans la pratique des terri*a*ns, néanmoins, cette ambiguïté ne pose pas plus de problème que l'homonymie ordinaire, le contexte rendant clair ce dont il s'agit. *Nous* écrirons parfois « PA^* » pour désigner le second système et le distinguer de (notre) PA. PA^* et PA sont donc formellement identiques mais intentionnellement différents, le premier parlant des n*a*mbres, le second devant être interprété comme parlant des nombres. Il nous faut enfin insister sur le fait que cette civilisation est familière des techniques de codage et possède également deux théorèmes d'incomplétude. Ces théorèmes sont, il faut bien l'admettre, tout aussi bons que nos théorèmes de Gödel. Ils les formulent en général, de façon quelque peu relâchée, de la façon suivante :

Théorème 1 (Premier théorème)**.** Si T est une théorie cohérente, récursivement énumérable, et suffisamment riche, alors T est incomplète.

Théorème 2 (Second théorème)**.** Si T (comme ci-dessus) est cohérente, alors

$$T \nvdash \neg coh_T \text{ et } T \nvdash coh_T$$

En particulier, l'énoncé coh_{PA} et sa négation sont indépendants de PA [19].

18. L'intérêt que nous portons à cet énoncé du langage de PA est fondamentalement lié à l'interprétation syntaxique que nous en faisons. On supposera que chez les terri*a*ns c'est son interprétation arithm*a*tique qui soulève, en premier lieu, l'intérêt.

19. Où coh_{PA} est l'énoncé du langage de PA mentionné plus haut.

Tout ceci est parfaitement standard chez les terri*a*ns. Ceci étant dit, bien entendu, quand ils nous entendent affirmer que les théorèmes d'incomplétude montrent qu'il y a « des énoncés vrais indécidables dans PA », ces gens sont d'accord avec nous, mais ils ne sont pas d'accord sur le fait que coh_{PA} en fait partie [20] !

Le contexte de notre expérience de pensée étant posé, considérons maintenant le scénario suivant. Un jour, Pierre, un terrien qui ne connaît pas grand chose aux terri*a*ns, engage une conversation sur le pouvoir explicatif de la vérité avec l'un d'eux. Voici une relation de la conversation [21] :

> Pierre — Croyez-vous que les axiomes de PA soient vrais ?
> L'Étranger — Oui, je crois que les axiomes de PA^* sont vrais.
> Pierre — Et croyez-vous que les règles d'inférences préservent la vérité ?
> L'Étranger — Je le crois en effet.
> Pierre — Vous croyez donc que tous les théorèmes de PA sont vrais ?
> L'Étranger — Oui [22].
> Pierre, tout à coup excité — Puisque PA prouve que $0 \neq 1$, vous croyez que c'est vrai, donc que $0 = 1$ n'est pas vrai, et par conséquent vous croyez donc que PA ne prouve pas que $0 = 1$, autrement dit vous croyez que PA est cohérent.
> L'Étranger — Tout à fait, je crois que PA^* est cohérent.
> Pierre — Êtes-vous d'accord par conséquent qu'une bonne théorie de la vérité pour le langage de PA doit avoir pour conséquence la cohérence de PA ?
> L'Étranger — En effet, c'est une bonne chose qu'une théorie de la vérité permette de rendre compte de ce raisonnement.
> Pierre — Vous savez, la théorie de la vérité de Tarski pour le langage de PA fait précisément cela.
> L'Étranger *(sincèrement)* — Oui, c'est assurément un beau travail que T*a*rski nous a laissé là.
> Pierre — Mais voyez : PA ne prouve pas la cohérence de PA, tandis que PA augmentée des principes généraux qui gouvernent le prédicat de vérité, cette théorie, dis-je, *prouve* la cohérence de PA. Mobiliser nos principes aléthiques permet d'expliquer des faits exprimables dans le langage de PA que PA ne peut pas expliquer. Mon acceptation de PA ne m'engage pas logiquement à accepter

20. Il n'y a rien de profond ici : c'est simplement que $\mathbb{N}$ n'est pas l'interprétation saillante de PA dans les contextes conversationnels terri*a*ns. Et, par hypothèse, l'énoncé désigné par coh_{PA} est *faux* dans l'interprétation arithm*a*tique du langage formel PA qui est saillante dans les contextes conversationnels terri*a*ns.

21. J'ai essayé de désambiguïser les occurrences de « PA » en utilisant « PA^* » quand c'était nécessaire. Quand il n'est pas possible de trancher, j'ai noté $PA^{(*)}$.

22. L'Étranger croit donc que tous les théorèmes de PA^* sont vrais...

> l'énoncé coh_{PA} du langage de PA exprimant la cohérence de PA, mais une fois reconnue la vérité de PA, je suis forcé d'accepter coh_{PA}. Il y a donc un fait non-sémantique qui est expliqué par l'attribution de vérité à PA.
> L'Étranger *(étonné)* — Il me semble que vous déraisonnez : vous ne pouvez pas avoir dérivé un énoncé faux, $coh_{PA^{(*)}}$, à partir d'une théorie vraie (PA^*), et de principes aléthiques vrais, par des règles de preuves correctes !

À ce point, pressentant que la spontanéité de l'échange pourrait compromettre l'esprit bon enfant de cette rencontre interculturelle, les deux protagonistes préfèrent mettre fin poliment à la conversation.

3.2 La situation logique, précisée

Pour comprendre l'enjeu de ce dialogue et l'argument que je vais présenter, il faut tout d'abord prêter attention à deux points laissés dans l'ombre par Shapiro et Ketland, et qui sont pourtant des conditions nécessaires pour observer le phénomène de non-conservativité qui est au cœur de leur argument. La première concerne la relation entre le langage de la théorie de départ, ici le langage de PA, et le langage de l'extension aléthique. En effet, pour formuler une théorie de la vérité pour un langage donné, il est nécessaire d'avoir à sa disposition la syntaxe du langage en question [23]. Cette syntaxe peut être développée de plusieurs façons. Soit elle est interprétée (codée) dans le langage de base, celui de l'arithmétique ici, soit elle est développée dans un vocabulaire spécifique dédié (comme dans la monographie de Tarski sur le concept de vérité par exemple, voir [Tarski 1983]). Si la syntaxe est développée dans un vocabulaire spécifique, l'extension aléthique tarskienne de PA qui mobilise cette théorie syntaxique, notons la $T_{syn}(PA)$ dans cette variante « sans codage », permettra de prouver la cohérence de PA, mais l'énoncé de la cohérence de la théorie PA qui est prouvé dans cette extension est un énoncé du vocabulaire spécifique déployé pour faire la syntaxe de la théorie de départ (PA) et non un énoncé du langage de PA lui-même. Dans ce cas, on n'observe donc pas de phénomène de non-conservativité sur PA [24]. En revanche, lorsque la syntaxe est interprétée dans le langage de PA *via* codage, alors l'énoncé de la cohérence de la théorie qui est dérivé dans l'extension aléthique est bien, cette fois, un énoncé du langage de cette théorie (PA). Mais cet énoncé n'est

23. C'est-à-dire une théorie des notions de termes, formules, de concaténation des symboles, etc., qui permette de parler des énoncés et de formuler des lois comme la commutativité du prédicat de vérité avec la négation.

24. Sauf à avoir posé par ailleurs des lois de correspondance entre le langage de la théorie de base et le langage de la syntaxe, mais je laisserai cette possibilité de côté ici pour ne pas alourdir l'exposé. Pour plus détail, je renvoie le lecteur à [Galinon 2010].

pas prouvable dans PA, d'où maintenant la non-conservativité. Il est donc crucial, pour pouvoir opposer la non-conservativité de l'extension aléthique au déflationniste, que la syntaxe soit formulée dans le langage de la théorie de départ. Nous allons voir que cette situation est la première cause d'une équivoque épistémologique.

Pour comprendre la seconde condition logique dont dépend l'observation du phénomène de non-conservativité, il faut introduire une distinction clé entre deux façons de comprendre les théories contenant des schémas d'axiomes — comme l'arithmétique de Peano en premier ordre, ou ZFC. Selon une première lecture, les schémas d'axiomes sont compris comme des *listes,* tandis que, selon une seconde lecture, les schémas d'axiomes sont compris comme des *règles.* Les schémas compris comme des listes dans la formulation d'une théorie ne sont qu'un artifice métalinguistique pour formuler finiment un ensemble infini, *mais bien défini,* d'axiomes, à savoir les instances du schéma obtenues en remplaçant les lettres schématiques qui y figurent par des formules d'un vocabulaire approprié, en général entendu comme le vocabulaire ambiant de la théorie. Les schémas compris comme *règles* sont des *schémas ouverts* (*open-ended schemas*) [25]. Ils doivent être compris comme engendrant sans fin de nouveaux axiomes à mesure que le langage environnant s'enrichit de nouveaux moyens d'expression, y compris des moyens d'expression que nous n'avions pas à notre disposition lorsque nous avons formulé la théorie pour la première fois. Le point important est que lorsqu'une théorie est formulée avec des schémas ouverts, il existe un fossé entre leur portée épistémologique réelle et leur pouvoir logique (formel) relativement à un état donné du langage, entre ce à quoi nous nous engageons en les acceptant et l'ensemble de leurs conséquences logiques strictes dans un environnement théorique donné. Appliquée au cas de PA, l'arithmétique de Peano en premier ordre, cette distinction montre qu'il y a en somme deux théories de l'arithmétique de Peano en premier ordre : celle que, faute de mieux, on pourrait appeler la théorie formalisée standard, dont la liste des axiomes est exactement spécifiée [26], et puis il y a la théorie schématique dans la formulation de laquelle le schéma d'induction est compris comme étant un schéma ouvert. Pour distinguer les deux formulations de l'arithmétique de Peano en premier ordre, je noterai PA l'axiomatisation ordinaire et $PA(S)$ l'axiomatisation schématique, celle où le schéma est compris comme étant ouvert.

Cette distinction étant faite, revenons à présent à la dérivation de l'énoncé de la cohérence de PA dans son extension aléthique. Dans le cas qui nous intéresse, où la syntaxe est codée dans l'arithmétique et où l'on observe la

25. Cette distinction est classique. Voir par exemple [Burgess & Rosen 1997] pour une discussion dans le contexte philosophique du débat entre nominalisme et réalisme en mathématique, [Feferman 1991] pour une discussion en termes logiques.

26. En particulier, les instances de substitution du schéma d'induction sont celles, et seulement celles, que l'on obtient en substituant des formules *du langage de* PA aux lettres schématiques dans le schème d'induction.

« non-conservativité », la dérivation de l'énoncé de la cohérence dans l'extension aléthique utilise une instance du schéma d'induction dans laquelle figure le prédicat de vérité [27]. C'est ce qui est visible dans la dérivation informelle de la cohérence de PA présentée en introduction, dont le passage crucial ici est :

1. ...
2. Tous les axiomes de PA sont vrais ;
3. Toutes les règles d'inférences préservent la vérité ;
4. *Donc* tous les théorèmes de PA sont vrais
   (À partir de 2 et 3, par une instance du schéma d'induction sur les théorèmes de A mettant en jeu le prédicat Vr, qui n'était pas définissable dans L_{PA}).
5. ...

Pour dériver l'énoncé arithmétique de la cohérence de PA dans notre extension aléthique, on a donc mobilisé des instances du schéma d'induction de l'arithmétique de Peano *qui ne figuraient pas dans la liste initiale des axiomes de* PA — de nouveaux axiomes, donc. Or le passage par cette extension n'est pas un accident. En effet si, avec les notations proposées, nous avons d'un côté le fait sur lequel s'appuie l'argument de la conservativité contre le déflationnisme :

Fait 1. L'extension aléthique tarskienne de $PA(S)$ est une extension non-conservative de $PA(S)$,

nous avons d'un autre côté le fait logique non moins important suivant :

Fait 2. L'extension aléthique tarskienne de PA étend *conservativement* PA [28].

Autrement dit, pour dériver l'énoncé codé de la cohérence dans l'extension aléthique, il faut non seulement ajouter à PA les axiomes aléthiques tarskiens proprement dits, mais encore étendre strictement le schéma d'induction de PA ; et le Fait 2 vient préciser que, sans l'élargissement du schéma, la seule extension de PA par les axiomes récursifs de la vérité est en fait *conservative* sur PA. Le point de logique que venons de rappeler est bien connu, et ce qui nous intéresse à présent est son interprétation épistémologique.

3.3 L'argument de la conservativité, réexaminé

Pour réévaluer l'argument de la conservativité à la lumière des précisions qui viennent d'être faites, revenons au dialogue présenté plus haut et à son interprétation, en clarifiant les équivoques possibles entre syntaxe et arithmétique

27. Dans le cas où la syntaxe est développée dans un vocabulaire séparé, cette induction se fait dans dans la théorie syntaxique, en étendant *son* schéma d'induction.
28. Voir par exemple [Halbach 1999, Th. 3.1, 359], pour une preuve « syntaxique ».

et entre théorie ordinaire et schématique. Pour commencer, il faut souligner que si la vérité de PA et la vérité de PA^* sont deux faits bien distincts, en revanche la cohérence de PA et la cohérence de PA^* sont un seul et même fait : les deux théories sont formellement identiques, et la propriété de cohérence est une propriété des systèmes formels. De plus, l'énoncé coh_{PA} du langage $\mathcal{L}_{PA}$ (morphologiquement identique à $\mathcal{L}_{PA^*}$) exprime la cohérence de PA^* (i.e., la cohérence de PA). Mais en quel sens ? D'abord et avant tout au sens où, modulo un codage dans PA de la syntaxe de PA^* (i.e., la syntaxe de PA), il est vrai *dans* $\mathbb{N}$ si et seulement si PA^* est cohérent. Mais ce n'est bien entendu *pas* le cas que coh_{PA} est vrai dans l'arithm*a*tique si et seulement si PA^* est cohérente. Le fait que coh_{PA} soit vrai ou non dans son interprétation arithm*a*tique n'a tout simplement rien à voir avec la cohérence de PA.

Par ailleurs, au cours de son argument, Pierre raisonne par induction de la façon suivante : les axiomes sont vrais, les règles préservent la vérité, donc tous les théorèmes sont vrais. Cette induction sur les démonstrations est correcte (c'est un principe de preuve valide de la syntaxe du langage), mais comme aurait pu le remarquer l'Étranger, une fois la syntaxe interprétée dans le langage de PA, on a affaire à une inférence arithmétiquement correcte, mais non arithm*a*tiquement correcte : une fois reformulée la syntaxe dans le langage de PA, cette induction mobilise un vocabulaire qui n'appartient pas au langage de PA (il contient « vrai ») et des instances du schéma d'induction qui sont *fausses* dans l'interprétation attendue de PA^* [29].

Nous avons donc une situation dans laquelle deux individus croient parler d'une même théorie. Même après en avoir donné tous les axiomes (et par conséquent tous les théorèmes), à la faveur d'une homonymie extraordinaire, la confusion est entretenue. Les deux individus ont par ailleurs la même théorie du concept de vérité. Tous deux pensent de la théorie dont ils croient respectivement qu'elle est l'objet de la conversation, qu'elle est vraie, et tous deux s'accordent sur le fait qu'à partir de cette théorie, d'une théorie de la syntaxe et de leur théorie de la vérité, ils peuvent conclure que cette théorie est cohérente. Pourtant l'Étranger n'est pas en mesure d'en conclure quoi que ce soit de neuf relativement au domaine d'objets qui était celui de sa théorie de base : tous deux ont expliqué la cohérence de la théorie de base, mais seul Pierre a ce faisant expliqué un fait relevant du domaine de sa théorie de base (le domaine de l'arithmétique). Mais comment Pierre peut-il *savoir* qu'il a par là même expliqué un nouveau fait arithmétique ? Il y a là quelque chose à expliquer, puisque l'Étranger n'est pas en droit de faire de même — il n'a de fait rien expliqué de neuf relativement au domaine qui est son sujet d'étude. Nous avons déjà donné la réponse : Pierre a en fait engagé dans le cours de l'argument de nouveaux principes de preuve qu'il sait être arithmétiquement corrects et qui n'avaient pas été précisés pour commencer dans la description

29. Si elles étaient vraies, coh_{PA} devrait l'être aussi, or par hypothèse c'est la *négation* de coh_{PA} qui est un énoncé arithm*a*tique vrai.

de PA, principes que l'Étranger n'admet pas comme principes de preuve pour les faits relevant du domaine qu'il a en vue, celui des n*a*mbres naturels.

Il est par conséquent naturel d'interpréter les faits logiques de la façon suivante : la théorie de la vérité par elle-même ne permet pas d'*apprendre* ou d'*expliquer* quelque chose de nouveau à propos des entiers, en dépit de la non-conservativité, parce que pour observer le phénomène en question, nous devons à un certain point étendre notre théorie arithmétique de départ *d'une façon qui n'est dictée ni par notre compréhension du concept de vérité lui-même, ni par les connaissances explicitement couchées dans les axiomes de la théorie de base*. Que l'on regarde cette extension comme une décision purement pragmatique, comme pourrait le faire un conventionnaliste en matière d'arithmétique, ou qu'on la regarde comme dictée par la volonté de formuler une connaissance arithmétique que nous avions déjà mais restée jusque-là implicite, cela ne change rien au fond : au bout du compte, la non-conservativité ne fait que refléter la décision que nous avons prise à un moment de renforcer nos axiomes, ou avérer la possibilité que nous avons eue à un moment de le faire. C'est cette décision que Pierre a prise subrepticement en utilisant une instance du schéma d'induction formulée à l'aide du prédicat de vérité pour prouver la cohérence de PA.

Au bout du compte, il apparaît donc que l'argument de la conservativité pivote sur une équivoque entre théorie ordinaire et théorie schématique. Mais de deux choses l'une : soit la théorie de base est une théorie ordinaire (avec des schémas compris comme listes), et alors son extension aléthique est en fait conservative. Dans ce cas, la thèse de la conservativité est peut-être vraie, mais l'argument n'atteint plus le déflationniste. Soit la théorie de base est schématique, mais alors la thèse de la conservativité est fausse car l'interprétation de la non-conservativité des extensions aléthiques sur les théories *schématiques* en termes de pouvoir explicatif des principes aléthiques est douteuse.

4 Conclusion

Savoir qu'une théorie interprétée A est vraie est suffisant pour inférer que A est cohérente pour peu que la syntaxe de A soit suffisamment bien comprise. Pourtant savoir que A est vraie ne donnera jamais aucune nouvelle connaissance des faits relevant du domaine d'investigation qui est celui de la théorie A au-delà de celles qui étaient déjà couchées dans A elle-même, à moins que nous n'ayons su depuis le début que A était d'une manière ou d'une autre une formulation défectueuse de notre connaissance réelle de son domaine, et n'ayons eu de plus le moyen de reconnaître quelques extensions strictes de A comme étant correctes relativement à cette connaissance. Autrement, pour faire écho à des réflexions que Shapiro a faites ailleurs [Shapiro 1998a, 618 sq.], comment pourrions-nous être certains que nous n'avons pas changé de sujet ?

Que nous ne sommes pas subrepticement passés d'un discours sur les n*a*mbres à un discours sur les nombres ? En montrant que l'interprétration des phénomènes de non-conservativité des extensions aléthiques proposée par Shapiro et Ketland n'est pas contraignante, nous sauvons l'idée que la vérité est une notion neutre et que nos principes aléthiques peuvent s'appliquer universellement à tout domaine de discours.

L'application classique des théorèmes de non-conservativité en philosophie des mathématiques est fournie par les arguments d'indispensabilité, et la parenté entre l'argument opposé au déflationnisme aléthique par Shapiro et Ketland et l'argument classique opposé aux nominalismes contemporains en mathématique est claire. Mais l'argument, cette fois-ci, n'atteint pas sa cible. Pour le comprendre, il faut d'abord noter le déplacement conceptuel opéré par le déflationniste en matière de vérité par rapport aux thèses nominalistes contemporaines. Le déflationniste n'est pas un réductionniste ou un éliminativiste en matière de vérité, puisqu'il soutient précisément la thèse du caractère *indispensable* de la notion de vérité. Ce que le déflationniste cherche à faire, nous l'avons dit, c'est à tracer une frontière *dans* l'ensemble des notions qu'il juge indispensables, entre celles qui relèvent de quelque chose comme l'appareil descriptif-explicatif d'une part, et celles qui relèvent de l'appareil logico-expressif d'autre part. Nous avons un certain nombre d'intuitions sur la différence de nature entre des notions comme la notion d'électron ou de cardinal fortement inaccessible d'un côté, et la notion d'identité ou de conjonction d'un autre. Il est vrai que ces intuitions sont difficiles à clarifier, mais elles n'en supportent pas moins depuis longtemps un effort philosophique soutenu pour en montrer le bien-fondé [30]. Dans ce cadre, il apparaît donc que la formulation de la thèse de la conservativité comme réponse au Problème de la frontière n'est pas une simple reformulation de positions prises antérieurement dans le débat autour des nominalismes en mathématiques. Ce que la thèse de la conservativité est supposée établir désormais, ce n'est pas l'indispensabilité de la vérité, mais le caractère « explicatif » de nos usages de la notion de vérité. Or, ce que nous avons vu est que, dans le cas des extensions aléthiques, l'observation d'un phénomène de non-conservativité sur la théorie de base dépend essentiellement du caractère schématique de cette théorie de base. Et nous avons montré que lorsque la thèse de la conservativité est comprise comme s'appliquant également à des théories de ce type, alors elle est hautement problématique. Dans ces conditions, le gain en pouvoir explicatif relativement à la théorie de départ résulte conjointement de l'ajout des axiomes aléthiques *et* des nouveaux axiomes de la théorie de base, et nous n'avons aucune rai-

30. C'est par exemple tout le sens des recherches sur la caractérisation sémantique des constantes logiques. Que le critère de conservativité ne trouve pas à s'appliquer sans difficulté dans cette problématique-ci est du reste attesté par les phénomènes bien connus de non-conservativité de lois logiques sur d'autres lois logiques. La portée de ce genre de phénomène est discutée par exemple dans [Dummett 1991, 290 sq.].

son de conclure que les vérités aléthiques, les lois axiomatiques de la vérité jouent le moindre rôle explicatif. Au contraire, une interprétation du phénomène qui ferait droit à l'idée que les axiomes tarskiens font de la vérité une notion purement expressive se présente naturellement. Plus spécifiquement, on peut factoriser de la façon suivante le processus conduisant à l'observation de phénomènes de non-conservativité :

1. Il est partie intégrante de notre compréhension de l'arithmétique que le schéma d'induction soit compris comme une règle. On ne peut pas comprendre complètement ce que sont les entiers naturels, sans être en position de reconnaître certaines théories de l'arithmétique plus forte que PA comme correctes.
2. Il y a de l'implicite et de l'explicite dans la représentation logique de nos connaissances lorsque nous formulons des théories où les schémas sont compris comme règles. Les connaissances laissées implicites ont des conséquences logiques latentes, mais ces conséquences ne sont pas déployées dans le langage ambiant. Autrement dit, ce qui est déductible formellement d'une théorie dont certains schémas sont compris comme règles n'épuise pas ce qui suit logiquement de ce qui était accepté et compris comme le contenu complet de la théorie.
3. Les principes de preuves ouverts deviennent naturellement plus forts quand le langage ambiant devient expressivement plus riche. Dans ce processus, certains principes de preuves dictés par notre compréhension du sujet de la théorie de départ sont explicités.
4. C'est à ce point que la vérité entre en jeu : la vérité est un outil expressif, et elle permet d'expliciter certains engagements théoriques que nous avons relativement à la théorie de départ.
5. Donc la raison véritable de la non-conservativité des théories de la vérité sur certaines théories de bases, lorsqu'elle est avérée, est à chercher dans le caractère *ouvert* de notre compréhension de la théorie de base (le caractère ouvert des principes de preuves arithmétiques) et le pouvoir *expressif* de la vérité.

Ce tableau de la situation semble raisonnable et constitue une réponse attractive à l'argument de la conservativité opposé au déflationisme aléthique. Sous cette interprétation le phénomène de non-conservativité corrobore l'idée que le prédicat de vérité sert à *expliciter ce qui était implicite* dans notre acceptation de la théorie de base, que la vérité nous a permis d'*exprimer*, ou de *formuler* une théorie plus forte des faits arithmétiques que celle que nous étions en mesure de formuler auparavant. Si nous nous souvenons du fait que, du point de vue déflationniste, le pouvoir expressif de la vérité est sa véritable raison d'être, la non-conservativité bien comprise des extensions aléthiques est même un phénomène de nature à le conforter [31].

31. Comme nous l'avons mentionné plus haut note 5, [Field 1999] a lui aussi contesté l'argument de la conservativité. Field remarque qu'on observe un phénomène de non-

Bibliographie

Burgess, John & Rosen, Gideon

1997 *A Subject with No Object : Strategies for Nominalistic Interpretation of Mathematics*, Oxford : Clarendon Press.

Dummett, Michael

1991 *The Logical Basis of Metaphysics*, The William James lectures ; 1976, Cambridge, Mass. : Harvard University Press.

Feferman, Solomon

1991 Reflecting on incompleteness, *Journal of Symbolic Logic*, 51, 1–48.

Field, Hartry

1999 Deflating the conservativness argument, *Journal of Philosophy*, 96, 533–540.

2001 *Truth and the Absence of Fact*, Oxford : Clarendon Press.

Frege, Gottlob

1971 *Écrits logiques et philosophiques*, Paris : Éditions du Seuil.

Galinon, Henri

2010 *Recherches sur la vérité*, Thèse de doctorat, Université Paris 1.

Halbach, Volker

1999 Conservative theories of classical truth, *Studia Logica*, 62, 353–370.

Halbach, Volker & Horsten, Leon (éd.)

2002 *Principles of Truth*, Frankfurt-am-Main : Ontos Verlag.

Ketland, Jeffrey

1999 Deflationism and Tarski's paradise, *Mind*, 108, 69–94.

---

conservativité des extensions aléthiques tarskienne de PA sur PA à la seule condition d'étendre les axiomes d'induction de PA. L'argument de Field est alors le suivant : le principe d'induction est un principe mathématique, et non aléthique. Donc le phénomène de non-conservativité est une indication de la « substance » du principe d'induction, non de la vérité. Nous sommes à la fois plus prudents dans notre conclusion et plus précis dans notre explication de la façon dont l'interaction entre le schéma d'induction et lois de la vérité permet une interprétation de la non-conservativité compatible avec les thèses déflationnistes. Le rôle logique et la portée épistémologique de l'hypothèse de l'interprétation de la syntaxe dans la théorie de base *via* codage dans l'argument de la conservativité tel qu'il a été formulé originalement n'a, à ma connaissance, jamais été souligné clairement.

Quine, Willard Van Orman

1970 *Philosophy of Logic*, Cambridge, Mass : Harvard University Press.

Shapiro, Stewart

1998a Induction and indefinite extensibility : The Gödel sentence is true, but did someone change the subject ?, *Mind*, 107(427), 597–624.

1998b Proof and truth : Through thick and thin, *Journal of Philosophy*, 95(10), 493–521.

Tarski, Alfred

1983 *Logic, Semantics, Metamathematics*, Indianapolis : Hackett pub.

Tennant, Neil

2002 Deflationism and the Gödel-phenomena, *Mind*, 111, 551–582.