UNIVERSITÉ PARIS-NANTERRE

MÉMOIRE DE MASTER 1

Une approche probabiliste de la théorie de David Lewis pour l'analyse des propositions contrefactuelles

---

Auteur :
Alaë AHMAD

Sous la direction de :
M. Denis BONNAY

Soutenu le 22 juin 2024

Table des matières

Introduction

La syntaxe du possible : la logique modale

Pour étudier la valeur de vérité d'une proposition, il suffit d'analyser la concordance entre son contenu logique et les *faits*. Par exemple, la proposition P : « César a franchi le Rubicon avec son armée » est une proposition vraie, tandis que la proposition Q : « la République romaine a survécu cent ans après César » est une proposition fausse. Cette simple description factuelle du monde est loin d'être triviale en pratique, mais les conditions de vérité ont le mérite d'être claires. En revanche, pour étudier la valeur de vérité de la proposition $\neg P \implies Q$: **« Si César n'avait pas franchi le Rubicon avec son armée, la République aurait survécu cent ans après César »**, qui ne se rapporte à aucun événement physique puisque $\neg P$ ne s'est pas réalisé, l'omniscience « physicaliste » (au sens de la connaissance de l'état de toutes les particules depuis l'origine de l'univers jusqu'à son éventuelle disparition) ne suffit pas. En logique classique, toute implication $FAUX \implies Q$ étant vraie, on est forcés de laisser de côté l'étude de telles propositions. [1]

Pourtant, il semble naturel d'accorder une valeur de vérité plus subtile à certaines de ces propositions dites *contrefactuelles*. Et pour ce faire, on ne voit pas comment procéder autrement que par l'analyse de la concordance entre leur contenu et un certain type de *faits*. Le formalisme majoritairement adopté par les philosophes analytiques pour donner un sens à ces propositions est celui de la logique modale, une extension de la logique propositionnelle avec des opérateurs modaux. En logique modale « classique » (ou *aléthique*) on retrouve les deux couples d'opérateurs : « nécessaire / contingent », « possible / impossible ».

La théorie du contrefactuel s'inscrit donc dans un champ métaphysique plus large : la théorie des **mondes possibles**. Ce concept leibnizien a été réintroduit par Carnap [2] en 1947. La terminologie de « monde possible » n'est pas encore employée par Carnap, mais ses « descriptions d'état » correspondent précisément à la conception du monde possible qui sera adoptée par les philosophes analytiques de la seconde moitié du $20^{\text{ème}}$ siècle, à savoir un ensemble maximal-consistant de propositions atomiques. Formellement, un monde possible $\mathcal{M}$ est un ensemble tel que pour toute proposition atomique P, $P \in \mathcal{M}$ ou (exclusif) $\neg P \in \mathcal{M}$. Ces propositions sont formées à partir de

1. À moins de se contenter du résultat trivial selon lequel toute proposition contrefactuelle est vraie...

2. Rudolf CARNAP, *Signification et nécessité*, Paris Gallimard, 1997. Traduit de l'anglais par F. Rivenc et Ph. de Rouilhan.

constantes (par exemple x = « Socrate ») et de prédicats (par exemple M = « être mortel », auquel cas $Mx \in \mathcal{M}$ et $\neg Mx \notin \mathcal{M}$). Des prédicats d'arité $n \geq 2$ permettent de former des propositions liant plusieurs constantes entre elles : xPy pour « x est le père de y », ...

Introduisons, pour rappel, les opérateurs modaux de nécessité et de possibilité :

- L'opérateur nécessaire (ce qui ne peut pas être faux) noté $\Box$;
- L'opérateur contingent (ce qui peut être faux) noté $\neg\Box$;
- L'opérateur possible (ce qui peut être vrai) noté $\Diamond$;
- L'opérateur impossible (ce qui peut ne peut pas être vrai) noté $\neg\Diamond$.

Un seul de ces opérateurs permet de définir les trois autres, grâce à l'équivalent des « lois de Morgan » modales : $\Diamond P \equiv \neg\Box\neg P$ et $\Box P \equiv \neg\Diamond\neg P$. Ces opérateurs permettent de définir, à partir de propositions se rapportant à un monde, des propositions dont la valeur de vérité s'interprète dans un *système de mondes*. Par exemple, $\Box\neg Mx$ est une proposition vraie s'il existe des mondes dans lesquels Socrate est immortel (quand bien même $\neg Mx$ est fausse dans notre monde), tandis que $\Diamond 2 + 2 = 5$ est une proposition fausse si l'on considère que les vérités mathématiques transcendent notre monde.

Le formalisme de David Lewis pour l'analyse des propositions contrefactuelles

Une fois ces définitions posées, étudier la valeur de vérité d'une proposition contrefactuelle $P \implies Q$ pourrait se ramener à l'analyse (contre)factuelle de l'implication dans le(s) monde(s) où l'impliquant est vrai. Une théorie aussi naïve se montre rapidement insuffisante : a-t-on envie de dire que la proposition **« Si César n'avait pas franchi le Rubicon avec son armée, la République aurait survécu cent ans après César »** est fausse car il existe un monde dans lequel un astéroïde a frappé la Terre 10 ans après le non-franchissement du Rubicon par César ? Si l'on se contente d'une telle définition, on se restreint à décrire des tautologies et des contradictions.

Pour cette raison, les théories qui se proposent de formaliser le contrefactuel sont techniquement plus subtiles. Citons par exemple Kripke[3], Cresswell[4], Stalnaker[5]. Je m'intéresserai exclusivement au formalisme proposé par David Lewis dans son ouvrage *Counterfactuals*[6], qui offre une représentation des conditions de vérité des propositions modales en fonction de la « proximité » avec un monde de référence, en utilisant les ressources fournies par la topologie mathématique.

3. Saul KRIPKE, *La Logique des noms propres*, Editions de Minuit, Paris, 1980. Traduit de l'anglais par François Recanati.
4. M. J. CRESSWELL, *The world is everything that is the case*, Cornell University Press, Ithaca, 1979.
5. Robert C. STALNAKER, *Possible Worlds*, Cornell University Press, Ithaca, 1979.
6. David LEWIS, *Counterfactuals*, Blackwell Publishers, Malden, 1986.

Des contrefactuelles probabilistes au sein du formalisme de David Lewis ?

David Lewis introduit le concept de « conditionnelles variablement strictes » pour relativiser de façon pratique la rigueur de l'implication contrefactuelle que l'on étudie : pour considérer que $P \implies Q$ est vraie, dans quels mondes doit-on s'assurer de la vérité (stricte) de son contenu ? La réponse qu'il fournit rend remarquablement compte de la notion intuitive que nous avons lorsque nous parlons de contrefactuel.

Il est encore possible (et très utile), au sein même de ce formalisme, de relativiser le degré de vérité que l'on attend de « l'implication logique ». Comme la majorité des propositions factuelles (non logico-mathématiques), les propositions contrefactuelles gagnent à être abordées dans les termes probabilistes. En passant de l'implication logique $P \implies Q$ à la probabilité $\mathbb{P}(Q|P)$, je propose une modeste généralisation du formalisme de David Lewis. En plus de l'intérêt de cet ajout pour l'étude des contrefactuelles « binaires », les outils fournis par la théorie des probabilités permettent de donner une valeur numérique à de nouveaux types de propositions contrefactuelles essentiellement quantitatives. Si X désigne la variable aléatoire indiquant la durée de vie de la République romaine après César et P l'événement « César franchit le Rubicon », la comparaison de $\mathbb{E}(X|P)$ et $\mathbb{E}(X|\neg P)$ fournit une information plus pertinente sur l'influence du franchissement du Rubicon par César sur la durée de vie de la République romaine.

Ce mémoire s'appuie sur les critiques contemporaines de la théorie de Lewis [7], qui ont donné lieu à des modèles probabilistes bien différents pour l'étude des contrefactuelles [8]. Je propose d'incorporer ces critiques pour enrichir le formalisme initial de Lewis avec la théorie des probabilités, tout en conservant sa « topologie des mondes possibles ».

7. Voir https ://plato.stanford.edu/entries/counterfactuals

8. Judea PEARL, "Causation, Action, and Counterfactuals", dans *Computational Learning and Probabilistic Reasoning*, 1995.

Le formalisme de David Lewis pour l'analyse des contrefactuelles à partir de la théorie des mondes possibles

Ce mémoire ne prétend pas développer une théorie « concurrente » à celle de David Lewis pour l'analyse des propositions contrefactuelles, mais s'appuie sur le formalisme qu'il a développé dans son ouvrage *Counterfactuals* pour proposer une modeste contribution à sa théorie. Il est donc indispensable de la reprendre dans les grandes lignes.

1.1 L'espace topologique du possible

1.1.1 L'insuffisance de l'implication « stricte »

Par cohérence avec les notations de Lewis, commençons par introduire les principaux opérateurs modaux qu'il utilise dans son ouvrage :

- L'opérateur « cela serait le cas »[1] noté $\Box\!\!\rightarrow$ ($P \Box\!\!\rightarrow Q$ se lit « S'il était le cas que P, alors il serait le cas que Q »);
- L'opérateur « cela pourrait être le cas »[2] noté $\Diamond\!\!\rightarrow$ ($P \Diamond\!\!\rightarrow Q$ se lit « S'il était le cas que P, alors il pourrait être le cas que Q »);
- La négation est notée $\sim$: $\sim(P \Box\!\!\rightarrow Q)$, $\sim(P \Diamond\!\!\rightarrow Q)$

Il ne s'agit pas d'une simple concaténation des opérateurs $\Box$ (ou $\Diamond$) et $\implies$: c'est justement car ce couple naïf n'est pas adapté à l'étude des propositions contrefactuelles que Lewis introduit ces nouveaux opérateurs.

Comme rappelé en introduction, une proposition contrefactuelle $P \Box\!\!\rightarrow Q$ ne peut se réduire à l'implication logique $P \implies Q$ dans tous les mondes possibles, puisqu'il est concevable que l'antécédent P se réalise sans que le conséquent Q ne se réalise, pour tout un tas de raisons loufoques qui nous éloignent de considérations implicites (une sorte de « toutes choses égales par ailleurs »). L'exemple donné par Lewis[3] montre les limites d'une analyse du contrefactuel à partir de l'implication stricte :

1. Opérateur "Would" en anglais.
2. Opérateur "Might" en anglais.
3. [1], p.10.

$$P_1 \boxrightarrow Q$$
$$P_1 \& P_2 \boxrightarrow \sim Q$$

Si P_1 et P_2 désignent respectivement « Alice est venue à la soirée » et « Bob est venu à la soirée », et Q « La soirée était agréable », on aimerait pouvoir simultanément dire « Si Alice était venue à la soirée, la soirée aurait été agréable » et « Si Alice et Bob étaient venus à la soirée, la soirée aurait été désagréable » (si Alice et Bob ne s'entendent pas, par exemple). Pourtant, étant donné que $P_1 \& P_2 \implies P_1$ (si Bob et Alice sont venus à la soirée, en particulier Bob est venu à la soirée), le couple de contrefactuelles ci-dessus est contradictoire. On pourrait bien sûr dire que cela prouve seulement que

$$P_1 \boxrightarrow Q$$

est faux (Alice aurait pu venir sans que la soirée soit agréable, par exemple si Bob était venu), mais on s'interdit alors l'usage le plus courant des propositions contrefactuelles. Dans la littérature philosophique, ce paradoxe est connu sous le nom de « problème du renforcement des antécédents » [4]. Contrairement à l'implication en logique classique, une propriété importante des propositions contrefactuelles est leur *non-monotonie*, au sens où un antécédent P plus contraignant ne rend pas nécessairement plus vraie l'implication $P \implies Q$. [5]

Il faut donc trouver une définition plus adéquate pour les opérateurs $\boxrightarrow$ et $\diamondrightarrow$. Notons qu'une fois l'un d'entre eux défini, le second l'est automatiquement par l'une des relations de dualité modale :

$$\boxrightarrow \equiv \sim \diamondrightarrow \sim \quad \text{et} \quad \diamondrightarrow \equiv \sim \boxrightarrow \sim$$

Nous choisirons donc, comme Lewis, de définir l'opérateur $\boxrightarrow$.

1.1.2 Le système concentrique des mondes possibles

La solution adoptée par David Lewis est de baser la valeur de vérité des propositions contrefactuelles sur une certaine notion de *similarité intermondaine*. Pour ce faire, il introduit une notion de distance topologique entre les mondes [6]. Une proposition contrefactuelle $P \boxrightarrow Q$ est alors vraie si « dans les mondes les plus proches du nôtre où P est vrai, Q est vrai ».

Formellement, si l'on se donne un monde i, et un ensemble $\$_i$ d'ensemble de mondes possibles vérifiant les conditions suivantes :

4. "Antecedent Strenghtening" en anglais. Voir https://plato.stanford.edu/entries/counterfactuals
5. Bien sûr, une implication logique est soit vraie soit fausse, mais si $P_1 \implies Q$ alors $P_1 \& P_2 \implies Q$, ce qui n'est pas le cas pour les propositions contrefactuelles.
6. Il s'agit d'une « distance » ensembliste, non nécessairement numérique.

(C) $\$_i$ est centré en i, c'est-à-dire que $\{i\} \in \$_i$.

(1) Les sphères $S \in \$_i$ sont imbriquées (si $S, T \in \$_i$ alors $S \subset T$ ou $T \subset S$).

(2) $\$_i$ est clos sous union.

(3) $\$_i$ est clos sous intersection.

On vérifie qu'un monde j est plus proche de i qu'un monde k si toutes les sphères $S \in \$_i$ contenant k contiennent également j (et strictement plus proche si la réciproque est fausse). La condition (C) assure que i est le monde le plus proche de lui-même dans ce système (strictement plus que tous les autres).

Visuellement, ce formalisme est très intuitif[7] :

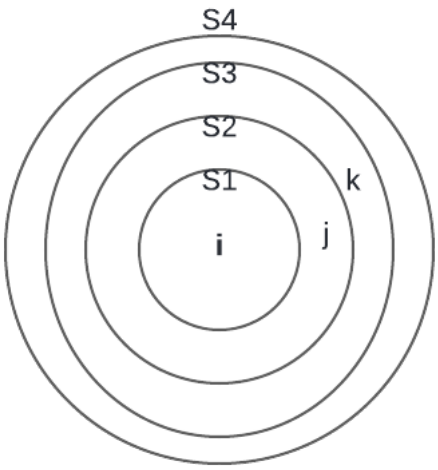

FIGURE 1.1 – Exemple : $j \in S_2$ plus proche de i que $k \in S_3 \setminus S_2$

Un tel système de sphères étant donné (David Lewis ne précise évidemment pas comment l'obtenir : il s'agit de définir les conditions abstraites de vérité des propositions contrefactuelles, pas une méthode pratique pour les déterminer), il ne reste plus qu'à introduire la notion de P-monde comme un monde j dans lequel la proposition P est vraie pour définir la valeur de vérité des propositions contrefactuelles.

Ainsi, la proposition $P \boxrightarrow Q$ est vraie si et seulement si l'une des deux conditions suivantes est vérifiée :

(1) Pour tout $S \in \$_i$, S ne contient pas de P-monde (ou : $\bigcup_{S \in \$_i} S$ ne contient pas de P-monde)

(2) Il existe $S \in \$_i$ contenant un P-monde tel que $P \implies Q$ soit vraie dans tous les mondes de S

Le cas (1) permet de rendre vraies les contrefactuelles avec antécédent toujours faux, tandis que le cas (2) plus intéressant s'interprète comme :

7. En théorie, il peut bien évidemment y avoir un nombre infini voire non dénombrable de telles sphères.

« Dans des mondes suffisamment proches de i, à chaque fois que P est vrai, Q est vrai ». En effet, la condition (2) est de moins en moins restrictive à mesure que l'on se rapproche de i (dans la limite où l'on reste assez loin pour qu'il existe des P-mondes), puisqu'il existe de moins en moins de $P\&\neg Q$-mondes susceptibles de rendre fausse la proposition contrefactuelle.

On serait alors tenté de remplacer (2) par une condition plus simple, à savoir :

$$(2) \quad P \implies Q \text{ dans } S_i^{(P)}$$

où $S_i^{(P)}$ désigne la plus petite sphère autour de i contenant des P-mondes. Comme le précise David Lewis [8], ce n'est pas possible en général car il n'existe pas toujours de plus petit élément pour l'inclusion dans un ensemble infini de telles sphères (ou plus précisément : l'intersection de toutes les sphères contenant des P-mondes, bien qu'étant une sphère par la condition (3), ne contient pas nécessairement de P-monde). Cependant, cette hypothèse simplificatrice se révèle pratique pour limiter la technicité du propos. Nous l'utiliserons donc plus fréquemment que David Lewis.

Avec cette définition des opérateurs modaux, David Lewis définit des propositions contrefactuelles « variablement strictes » : la condition de vérité de la proposition contrefactuelle $P \boxrightarrow Q$ dépend de P (non seulement pour vérifier que l'implication est valide, mais aussi pour déterminer dans quels mondes $S_i^{(P)}$ il est nécessaire de vérifier la validité de l'implication). Illustrons ce phénomène à l'aide de l'exemple de David Lewis sur les séries de contrefactuels avec « antécédents croissants » [9] (de plus en plus restrictifs) :

$$\phi_1 \boxrightarrow \psi$$
$$\phi_1 \& \phi_2 \boxrightarrow \sim\psi$$
$$\phi_1 \& \phi_2 \& \phi_3 \text{ [10]} \boxrightarrow \psi$$

(Pour reprendre l'exemple de Lewis : « Si Alice était venue à la soirée, la soirée aurait été agréable », « Si Alice et Bob étaient venus à la soirée, la soirée aurait été désagréable », « Si Alice, Bob et Charles étaient venus à la soirée, la soirée aurait été agréable »)

Un tel système de sphères permet de rendre simultanément vraies toutes ces propositions contrefactuelles :

8. [1], 1.4 "The Limit Assumption".
9. [1] p. 18.
10. L'associativité permet de se passer des parenthèses.

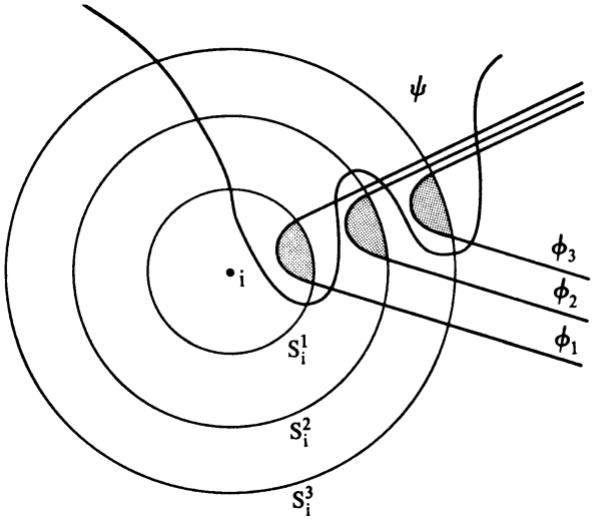

FIGURE 1.2 – Antécédents en série

Dans S_i^1, il existe des ϕ_1-mondes et ils sont tous également des ψ-mondes, ce qui rend vraie la première proposition.

Dans S_i^2, il existe des $\phi_1\&\phi_2$-mondes et aucun n'est un ψ-monde, ce qui rend vraie la deuxième proposition.

Dans S_i^3, il existe des $\phi_1\&\phi_2\&\phi_3$-mondes et ils sont tous également des ψ-mondes, ce qui rend vraie la troisième proposition.

Par la relation de dualité $\diamond\!\!\rightarrow \equiv \sim \Box\!\!\rightarrow\sim$, la proposition $P \diamond\!\!\rightarrow Q$ est vraie si et seulement si les deux conditions suivantes sont vérifiées :

(1) Il existe un P-monde dans une sphère $S \in \$_i$, S (ou : $\bigcup_{S\in\$_i} S$ contient au moins un P-monde)

(2) Pour tout $S \in \$_i$ contenant un P-monde, S contient aussi un $P\&Q$-monde

Avec l'hypothèse simplificatrice que nous avons faite, la condition (2) se réécrit plus simplement :

$$(2) \quad S_i^{(P)} \quad \text{contient un } P\&Q\text{-monde}$$

Autrement dit, la proposition Q n'est pas plus « invraisemblable » que la proposition P (puisqu'en se rapprochant autant que possible de notre monde tout en conservant P, on trouve encore des Q-mondes). Il s'agit bien de l'idée que se l'on fait intuitivement de la possibilité contrefactuelle « Si P était vrai, il serait possible que Q soit vrai ».

1.1.3 Cas limite : les implications « contrefactuelles » avec antécédent factuel

David Lewis examine ensuite des cas limites de sa théorie pour vérifier si la valeur de vérité des contrefactuelles au sein de ce formalisme coïncide toujours avec l'intuition.

On se heurte à une première difficulté dans le cas où la proposition « contrefactuelle » P faisant office d'antécédent est vraie dans notre monde. Dans ce cas, $S_i^{(P)} = \{i\}$ (toujours bien défini, sans hypothèse simplificatrice) d'où :

(1) $P \,\square\!\!\rightarrow Q$ est vraie si et seulement si Q est vraie (dans i)

(2) $P \diamond\!\!\rightarrow Q$ est vraie si et seulement si Q est vraie (dans i)

étant donné l'équivalence entre les quantificateurs « il existe » et « pour tout » dans le singleton $\{i\}$.

Ces deux équivalences sont problématiques. Dans le cas où Q est également vrai dans notre monde, mais pour une raison (a priori) indépendante de P, l'implication contrefactuelle $P \,\square\!\!\rightarrow Q$ est douteuse (mais on veut tout de même pouvoir être capable d'affirmer $P \diamond\!\!\rightarrow Q$). Dire : « Si le ciel était bleu, Alice se serait nécessairement levée à 7h le 3 juin 2024 » semble contre-intuitif, quand bien même ces deux propositions seraient vraies dans notre monde.

Symétriquement, si Q est faux dans notre monde pour des raisons (a priori) indépendantes de P, on veut être en mesure de pouvoir affirmer que $P \diamond\!\!\rightarrow Q$: « Si le ciel était bleu, Alice aurait pu se lever à 8h le 3 juin 2024 ».

Plutôt que de se restreindre aux propositions « strictement » contrefactuelles, Lewis propose [11] de remplacer la condition (C) par des sphères *faiblement* centrées. Formellement, si l'on se donne un monde i, un ensemble $\$_i$ d'ensemble de mondes possibles vérifiant les conditions suivantes :

(CF) $\$_i$ est faiblement centré en i, c'est-à-dire que $\forall S \in \$_i, i \in S$.

(1) Les sphères $S \in \$_i$ sont imbriquées (si $S, T \in \$_i$ alors $S \subset T$ ou $T \subset S$).

(2) $\$_i$ est clos sous union.

(3) $\$_i$ est clos sous intersection.

Le monde i reste le monde le plus proche de lui-même, mais plus nécessairement strictement. Il existe (potentiellement) une sphère $S_i^{(0)}$ de mondes, non réduite au singleton $\{i\}$. Ces mondes permettent de de généraliser la bonne définition des opérateurs contrefactuels même dans le cas où la proposition-antécédent P est vraie. Dans la partie suivante, nous verrons que le formalisme probabiliste est, à certains égards, plus naturel qu'un tel ajout ad hoc (qui conserve la rigueur logique du modèle, mais qui nécessite d'abandonner un axiome assez intuitif...)

1.2 Des reformulations équivalentes

Pour des raisons pratiques, David Lewis fournit plusieurs formalismes équivalents à celui que nous venons de voir [12] (un formalisme étant parfois techniquement plus adapté qu'un autre pour l'étude de certaines propriétés),

11. [1], 1.7 "True Antecedents".
12. [1], 2. "Reformulations".

avec la démonstration de l'équivalence logique des formalismes. Nous nous contenterons des exemples qui s'avéreront utiles dans la partie suivante.

1.2.1 La relation d'ordre de similarité intermondaine

Dans le formalisme topologique que nous venons de voir, les sphères concentriques semblent induire directement une relation d'ordre entre les mondes : plus un monde est dans une sphère rapprochée de i, plus il est proche de i. Il s'avère qu'en définissant correctement la relation d'ordre $\leq_i$ de *similarité intermondaine*, il n'y a pas de perte d'information (on peut également remonter au système de sphères à partir de la relation d'ordre). Si l'on parvient à exprimer la valeur de vérité des contrefactuelles $P \boxrightarrow Q$ et $P \diamondrightarrow Q$ avec cette nouvelle relation d'ordre, nous disposerons donc de deux formalismes équivalents pour la théorie de Lewis.

Formellement, une relation d'ordre $\leq_i$ définie sur un ensemble S_i de mondes vérifiant les conditions :

(1) $\leq_i$ est transitive : si $j \leq_i k$ et $k \leq_i h$, $j \leq_i h$.

(2) $\leq_i$ est une relation d'ordre totale : $\forall j, k \in S_i, j \leq_i k$ ou $k \leq_i j$.

(3) $i \in S_i$ et $\forall j \in S_i, i \leq_i j$.

(C) $\leq_i$ est strictement minimal pour $\leq_i$: si $j \neq i, i <_i j$

est une relation de *similarité intermondaine* (centrée si la condition (C) est vérifiée).

On obtient alors ainsi la valeur de vérité de la proposition $P \boxrightarrow Q$:

(1) S_i ne contient pas de P-monde, ou

(2) Il existe un P-monde $k \in S_i$ tel que pour tout $j \in S_i$, si $j \leq_i k$ alors $P \implies Q$ dans j.

Sous cette forme, il est encore plus clair que $P \boxrightarrow Q$ se réduit à l'implication classique $P \implies Q$ dans des mondes « suffisamment proches » de i.

On remonte de ce formalisme au système de sphères précédent en définissant $\$_i$ comme l'ensemble des parties S de S_i telles que pour tout $j \in S$ et $k \notin S, j <_i k$. On montre alors que $\$_i$ vérifie les conditions (1)-(3) définies en 1.1.2 (ainsi que la condition (C) si $\leq_i$ vérifie (C)), et que les valeurs de vérité de $P \boxrightarrow Q$ sont équivalentes dans les deux formalismes.

Réciproquement, en définissant $\leq_i$ sur $S_i = \bigcup_{S \in \$_i} S$ par : $j \leq_i k$ si et seulement si pour toute sphère $S \in \$_i, k \in S \implies j \in S$, on montre que la relation $\leq_i$ vérifie les conditions (1)-(3) définies ci-dessus (ainsi que la condition (C) si $\leq_i$ vérifie (C)), et que les valeurs de vérité de $P \boxrightarrow Q$ sont équivalentes dans les deux formalismes.

1.2.2 La similarité intermondaine quantitative

La relation d'ordre définie précédemment ne nécessite pas de définir une notion de « distance » numérique entre les mondes. Cependant, si l'on disposait d'une fonction degré de similarité $d_i(j)$ définie sur S_i, on dériverait directement une relation de similarité intermondaine $\leq_i$ par :

$$j \leq_i k \quad \text{ssi} \quad d_i(j) \geq d_i(k)$$

Dans ce cas, on reformule plus simplement la valeur de vérité de la proposition $P \,\Box\!\!\rightarrow Q$:

(1) S_i ne contient pas de P-monde, ou

(2) Il existe $d \in \mathbb{R}^+$ tel qu'il existe des P-mondes similaires à i à degré au moins d, et pour tout $j \in S_i$, si $d_i(j) \geq d$ alors $P \implies Q$ dans j.

Avec cette hypothèse simplificatrice, on peut définir [13] pour toute proposition P son degré de vraisemblance (relativement à i) $deg_i(P)$ comme la borne inférieure des réels $d \in \mathbb{R}^+$ tels qu'il existe des P-mondes similaires à i à degré $d_i(j) \geq d$. Par exemple $deg_i(FAUX) = 0$ et $deg_i(P) = d_i(i)$ si i est un P-monde.

Cette nouvelle notation nous permet de définir une nouvelle relation de comparaison [14] sur les propositions contrefactuelles :

$$P \preccurlyeq_i Q \quad \text{ssi} \quad deg_i(P) \geq deg_i(Q)$$

Par exemple : ni les chiens ni les cailloux ne parlent, mais l'on a envie de pouvoir dire qu'un monde dans lequel les chiens parlent est moins farfelu qu'un monde dans lequel les cailloux parlent. Toute proposition contrefactuelle est automatiquement plus farfelue que toute proposition vraie selon cette définition, puisque $deg_i(Q) \leq d_i(i)$.

Notons tout de même que cette « reformulation » n'est pas équivalente à la précédente (et donc au formalisme initial), puisque, comme Lewis le remarque lui-même [15], elle restreint l'étude à des mondes possibles dont le cardinal des différents degrés de similarité avec i est borné par $Card(\mathbb{R}) = \aleph_1$.

1.2.3 Les fonctions de sélection

Une autre reformulation du formalisme topologique, inspirée par une théorie alternative des contrefactuelles développée par Stalnaker [16], nécessite de se placer dans l'hypothèse simplificatrice selon laquelle il existe une sphère $S_i^{(P)}$ minimale contenant des P-mondes pour toute proposition contrefactuelle (non contradictoire) P.

Dans ce cas, on définit :

13. Lewis ne le fait pas, mais je l'introduis ici car j'utiliserai cette notation dans la suite.
14. que Lewis définit plus généralement, sans hypothèse simplificatrice, dans [1] 2.5 "Comparative Possibility".
15. [1] p.51.
16. Robert C. STALNAKER, Possible Worlds, Cornell University Press, Ithaca, 1979.

$$f_i(P) = \begin{cases} \{\text{les } P\text{-mondes appartenant à } S_i^{(P)}\} & \text{s'il existe des } P\text{-mondes} \\ \varnothing & \text{sinon} \end{cases}$$

Résumons : $S_i^{(P)}$ désigne la sphère la plus proche centrée en i qui contient des P-mondes (mais pas que), et $f_i(P)$ sélectionne les P-mondes de cette sphère.

On dérive de cette définition directement les valeurs de vérité de $P \mathbin{\Box\!\!\rightarrow} Q$ et $P \mathbin{\Diamond\!\!\rightarrow} Q$:

(1) $P \mathbin{\Box\!\!\rightarrow} Q$ est vraie si et seulement si $\forall j \in f_i(P)$, Q est vraie dans j.

(2) $P \mathbin{\Diamond\!\!\rightarrow} Q$ est vraie si et seulement si $\exists j \in f_i(P)$, Q est vraie dans j.

Attention : dans le cas où $S_i^{(P)}$ n'est pas bien défini (cf. 1.1.2), aucun P-monde n'appartient à l'intersection de toutes les sphères contenant des P-mondes, donc si l'on tente de « généraliser » ainsi la fonction précédente par

$$f_i(P) = \begin{cases} \text{les } P\text{-mondes appartenant à toutes les sphères contenant des } P\text{-mondes} & \text{s'il existe des } P\text{-mondes} \\ \varnothing & \text{sinon} \end{cases}$$

On peut se retrouver dans des cas où $f_i(P) = \varnothing$ quand bien même il existe des P-mondes, auquel cas $P \mathbin{\Box\!\!\rightarrow} Q$ serait trivialement considérée comme « vraie » avec la définition précédente alors même qu'il n'existe peut-être pas de sphère $S \in \$_i$ contenant un P-monde et dans laquelle $P \implies Q$ soit vraie. Il n'est donc possible d'utiliser cette « reformulation » que sous l'hypothèse simplificatrice de l'existence de $S_i^{(P)}$ pour toute proposition P.

Réciproquement, Lewis montre que l'on peut remonter des fonctions de sélection au formalisme topologique initial [17], ce que nous admettrons ici.

17. [1] 2.7 "Selection Functions".

Les contrefactuelles probabilistes à partir du formalisme de David Lewis

2.1 Pourquoi des probabilités ?

Le formalisme de David Lewis résout le problème de la rigueur de l'implication logique classique, trop stricte pour étudier les propositions contrefactuelles en la généralisant simplement à l'ensemble des mondes possibles à l'aide des opérateurs modaux, en restreignant l'ensemble des « mondes possibles » à des sous-ensembles de taille raisonnable, adaptée aux propositions contrefactuelles que l'on considère.

De façon similaire, on peut rendre les implications contrefactuelles plus « flexibles » en modifiant un autre paramètre : le degré de vérité que l'on attend qu'elles vérifient. Cet ajout n'est pas simplement là pour généraliser la théorie de Lewis à d'autres types de contrefactuelles, pour lesquelles les tables de vérité binaires sont peu adaptées. Il s'agit de montrer que, même pour les exemples donnés par Lewis, son formalisme ne permet parfois pas d'accorder une valeur de vérité cohérente avec l'intuition.

2.1.1 L'insuffisance de la logique formelle

Considérons la proposition contrefactuelle « *Si les voisins n'avaient pas fait la fête, j'aurais dormi correctement.* » [1] que l'on note $P \boxrightarrow Q$.

Pour analyser sa valeur de vérité dans notre monde i avec le formalisme de David Lewis, plaçons-nous dans la sphère $S_i^{(P)}$ [2], la plus petite sphère contenant des P-mondes (c'est-à-dire des mondes dans lesquels mes voisins n'ont pas fait la fête). Pour pouvoir affirmer $P \boxrightarrow Q$, il faut que tous les P-mondes de cette sphère soient également des Q-mondes, autrement dit : dans **tous** les mondes assez proches du nôtre dans lesquels mes voisins n'ont pas fait la fête, j'ai dormi correctement.

1. Toute une partie de la théorie de David Lewis est consacrée au problème de l'identité transmondaine : qui est « je » dans les autres mondes ? Je n'ai pas repris son analyse du concept de « contreparties » pour répondre à cette question dans la partie précédente car il s'agit d'une question indépendante de mon propos.
2. Cela nécessite l'hypothèse simplificatrice mentionnée en 1.1.2, que j'utilise uniquement par commodité technique.

Il suffit qu'il existe un P-monde dans $S_i^{(P)}$ dans lequel mes voisins n'ont pas fait la fête, mais j'ai mal dormi à cause d'un moustique, pour rendre faux $P \boxrightarrow Q$. L'implication logique reste trop rigoureuse, même restreinte à des « mondes proches », pour formaliser le contrefactuel intuitif.

Si, par « chance », tous les mondes dans lesquels un moustique se trouvait dans ma chambre sont plus éloignés de i que le plus petit monde dans lequel mes voisins n'ont pas fait la fête, le problème se pose de toute façon pour l'implication contrefactuelle « *S'il n'y avait pas de moustique dans ma chambre, j'aurais dormi correctement.* » ($R \boxrightarrow Q$), rendue fausse par des mondes plus proches dans lesquels les voisins ont fait la fête.

Le problème fondamental réside dans le fait que le système de sphères (ou la relation de comparaison intermondaine) défini dans le formalisme de Lewis ne dépend pas de la proposition P : il s'agit d'une donnée absolue. Comme l'étude de $P \boxrightarrow Q$ restreint seulement ensuite cet ensemble à des P-mondes, il les restreint aussi potentiellement à des R-mondes, dans lesquels R peut influer sur Q (de façon contraire à P).

Le philosophe Kit Fine donne un exemple plus frappant[3] avec la proposition contrefactuelle « Si Nixon avait appuyé sur le bouton (P), il y aurait eu un holocauste nucléaire (Q) ». Cette proposition contrefactuelle devrait intuitivement être considérée comme vraie. Pourtant, au sein du formalisme de Lewis, pour peu que la distance entre notre monde actuel et un monde dans lequel un holocauste nucléaire s'est produit soit supérieure à un monde dans lequel une différence minime, comme un dysfonctionnement du système de détonation (R) empêche l'holocauste nucléaire (Q), alors la contrefactuelle $P \boxrightarrow Q$ est fause. En effet, les mondes les *plus proches* du nôtre dans lesquels Nixon appuie sur le bouton ne sont pas des mondes dans lesquels un holocauste nucléaire se produit, alors même qu'il est évident qu'il existe une forte relation causale entre P et Q.

On retrouve donc une problématique similaire à celle du problème initial, qui justifiait de passer de la modalité sur l'ensemble des mondes possibles à des sous-ensembles bien choisis pour rendre compte de l'intuition derrière les propositions contrefactuelles. Dans le formalisme de Lewis, au moins pour certaines propositions contrefactuelles, on se retrouve dans des cas où la valeur de vérité de $P \boxrightarrow Q$ est influencée par des facteurs arbitraires et non pertinents. Dans le premier exemple, cela dépend (entre autres) de l'affectation initiale de distance inter-mondaine entre les mondes dans lesquels les voisins sont silencieux, et ceux dans lesquels un moustique est présent dans la chambre.

La valeur de vérité binaire pour l'implication contrefactuelle semble donc encore trop restrictive pour ne pas reprendre systématiquement fausses des

3. Kit FINE, "Review of Lewis' Counterfactuals", *Mind*, 1975.

propositions de type $P \boxed{\rightarrow} Q$, et systématiquement vraies des propositions de type $P \diamond\!\!\rightarrow Q$.

2.1.2 Les probabilités comme fondement de la connaissance (contre)factuelle

Dans la réflexion épistémologique moderne, on ne peut manquer de remarquer la prééminence des probabilités sur l'implication logique en tant que fondements de la connaissance (hors logico-mathématique). Et pour cause : cette idée peut être illustrée de manière particulièrement évidente par la science médicale. Supposons que les propositions P et Q désignent respectivement « prendre le vaccin X » et « contracter la maladie Y ». Si l'on se fie à l'implication logique stricte, $P \implies \neg Q$ sera considéré comme faux dès lors qu'un seul cas de figure viendra contredire cette affirmation, par exemple une personne ne réagissant pas bien au vaccin.

C'est précisément pour cette raison que l'utilisation de la probabilité conditionnelle $\mathbb{P}(\neg Q|P)$ se révèle beaucoup plus pertinente. Cette approche probabiliste permet de nuancer la relation de causalité entre P et Q quand il ne s'agit pas d'une conséquence logique qui déroule d'un raisonnement mathématique implacable, en tenant compte de la fréquence observée des guérisons suite à la prise du vaccin. Même si $\mathbb{P}(\neg Q|P)$ n'est pas strictement égal à 1, cela ne signifie pas pour autant que l'on doit s'empêcher d'affirmer que « le vaccin X empêche de contracter la maladie Y ». Au contraire, cela apporte une information supplémentaire, également très utile.

L'analogie avec la théorie de Lewis se clarifie avec cet exemple : cette fois, P et Q désignent respectivement « Alice prend le vaccin X » et « Alice contracte la maladie Y ». De la même façon que, dans notre monde, un individu sur 1000 est susceptible de contracter la maladie Y après avoir pris le vaccin X, il en est de même pour 1000 contreparties d'Alice (qui, certes, sont plus proches entre elles que 1000 individus choisis aléatoirement dans notre monde, mais Lewis refuse toute forme d'essentialisme intermondain. En particulier, les contreparties ne sont pas nécessairement physiologiquement identiques [4]). On ne pourrait, dans le formalisme de Lewis, pas conclure que $P \boxed{\rightarrow}\!\!\sim Q$ pour peu qu'il existe une contrepartie d'Alice dans un monde « assez proche » contractant quand même la maladie Y après avoir pris le vaccin X. [5]

Cette idée selon laquelle les probabilités sont plus adaptées que l'implication logique pour fonder la connaissance n'est pas nouvelle en philosophie. Elle a notamment été défendue par des auteurs tels que Carnap [6] ou

4. [1], 1.9 "Potentialities".
5. Rappelons que la proximité des mondes n'est pas définie, a priori, en fonction de la réaction d'Alice au vaccin.
6. Rudolf CARNAP, *Logical Foundations of Probability*, 1950.

encore Frank Ramsey[7], qui ont souligné la nécessité de prendre en compte l'incertitude et la variabilité dans la démarche scientifique, en recourant à des outils probabilistes plutôt que de se contenter d'une approche trop rigide. Les premiers philosophes analytiques du contrefactuel, dont David Lewis, proposent étonnamment des théories assez « conservatrices » au sens de leur attachement à la logique formelle, quand bien même leurs formalismes peuvent parfois être enrichis par la théorie des probabilités sans les dénaturer. C'est le cas, me semble-t-il, de la théorie de Lewis.

Les études plus récentes pour l'étude des propositions contrefactuelles s'appuient essentiellement sur la théorie des probabilités[8] pour pallier ce qui semble être une insuffisance de la logique modale. Nous n'étudierons pas en profondeur ces modèles dans le cadre de ce mémoire. Le philosophe et informaticien Judea PEARL synthétise ainsi la différence entre cette nouvelle famille de modèles probabilistes et la théorie de Lewis :

« Contrairement à la théorie de Lewis, les contrefactuelles ne sont pas basées sur une notion abstraite de similarité entre les mondes hypothétiques ; au lieu de cela, elles reposent directement sur les mécanismes (ou *lois*, pour être élégant) qui produisent ces mondes et sur les propriétés invariantes de ces mécanismes. Les « miracles » insaisissables de Lewis sont remplacés par des interventions principielles qui représentent le changement minimal (à un modèle) nécessaire pour établir l'antécédent... Ainsi, les similarités et les priorités - si elles sont jamais nécessaires - peuvent être lues dans les interventions comme une pensée ultérieure... mais elles ne sont pas fondamentales pour l'analyse. »[9]

La suite de ce mémoire s'appuie à montrer qu'il est possible de conserver la « notion abstraite de similarité entre les mondes » du formalisme de Lewis pour répondre aux objections mentionnées précédemment et rendre compte de la *causalité contrefactuelle* dans son acception la plus intuitive (en particulier, en rendant vraies des contrefactuelles comme « Si les voisins n'avaient pas fait la fête, j'aurais dormi correctement » ou « Si Nixon avait appuyé sur le bouton, il y aurait eu un holocauste nucléaire »).

2.1.3 L'étude de contrefactuelles quantitatives

Revenons à la proposition contrefactuelle de l'introduction : **« Si César n'avait pas franchi le Rubicon avec son armée, la République romaine aurait survécu cent ans après César »**. Il est en théorie possible d'étudier la valeur de vérité de cette proposition contrefactuelle de type $\sim P \boxrightarrow Q$ au sein du formalisme de David Lewis. On se place dans $S_i^{(\sim P)}$, et si l'ensemble des $\sim P$-mondes de $S_i^{(\sim P)}$ sont également des Q-mondes (c'est-à-dire des mondes

7. Frank Ramsey, *Truth and Probability*, 1926.

8. Judea PEARL, "Causation, Action, and Counterfactuals", dans *Computational Learning and Probabilistic Reasoning*, 1995.

9. Voir https ://plato.stanford.edu/entries/counterfactuals (3.3) pour la version originale.

dans lesquels la République romaine a survécu cent ans après César), l'implication contrefactuelle est vraie.

Pour illustrer le problème qui se pose avec ce formalisme, considérons la famille de contrefactuelles R_x (indexée par $x \in \mathbb{R}^+$) de la forme $\sim P \boxrightarrow Q_x$ où Q_x désigne : « La République romaine survit (au moins) x années après César ». Comme l'antécédent est identique pour toutes ces contrefactuelles, les sphères $\sim P$-permissives sont également identiques : pour étudier la valeur de vérité de n'importe laquelle de ces propositions, on se place dans $S_i^{(\sim P)}$ [10]. Cependant, les propositions Q_x sont de plus en plus restrictives à mesure que x croît ($Q_y \implies Q_x$ pour $x \leq y$) et donc la valeur de vérité des implications R_x « décroît » à mesure que x augmente. Étant donné que seules les valeurs de vérité 0 ou 1 sont possibles pour toutes les R_x, en fixant X comme la borne supérieure des réels $x \in \mathbb{R}^+$ pour lesquels R_x est vrai, on doit admettre que pour tout $\varepsilon > 0$:

$$R_{X-\varepsilon} \equiv V \quad \text{et} \quad R_{X+\varepsilon} \equiv F$$

Autrement dit : au sein du formalisme de David Lewis, on doit pouvoir affirmer « Si César n'avait pas franchi le Rubicon avec son armée, la République romaine aurait survécu $X - \varepsilon$ années après César », tout en niant aussi catégoriquement « Si César n'avait pas franchi le Rubicon avec son armée, la République romaine aurait survécu $X + \varepsilon$ années après César », pour tout $\varepsilon > 0$.

2.1.4 La variable aléatoire contrefactuelle

L'exemple de la section précédente montre que, pour certaines propositions contrefactuelles quantitatives (ce qui représente en réalité beaucoup plus de contrefactuelles qu'il n'y paraît : même les exemples introductifs de David Lewis « Si A était venu, la soirée aurait été *agréable* », « Si A et B étaient venus, la soirée aurait été *désagréable* » sont essentiellement quantitatifs, puisque le caractère *agréable* d'une soirée est un spectre), le formalisme de David Lewis nous confine à une certaine forme de « manichéisme contrefactuel ».

Les outils fournis par la théorie des probabilités permettent de donner des valeurs numériques plus adaptées à ces propositions contrefactuelles essentiellement quantitatives. Si X désigne la variable aléatoire indiquant la durée de vie de la République romaine après César et P l'événement « César franchit le Rubicon », la comparaison de $\mathbb{E}(X|P)$ et $\mathbb{E}(X|\sim P)$ (toujours au sein d'un sous-ensemble de mondes possibles adéquat) fournit une information plus pertinente sur l'influence du franchissement du Rubicon par César sur la durée de vie de la République romaine.

10. Je rappelle que cette hypothèse simplificatrice n'est pas nécessaire, mais permet simplement d'alléger le propos du point de vue technique.

On peut même se contenter d'une notion d'espérance intermondaine sur les variables aléatoires contrefactuelles, puisqu'en choisissant correctement la variable aléatoire on retrouve facilement la probabilité associée à tout type d'événement « non quantitatif » :

$$\mathbb{P}(X \in A|P) = \mathbb{E}[\mathbf{1}_{X\in A}|P]$$

2.2 Le nouveau modèle « Lewis - probabiliste »

Dans cette section, je propose une façon d'incorporer les considérations probabilistes au sein du formalisme de David Lewis.

2.2.1 Formulation topologique

Même avec un formalisme probabiliste, il reste déraisonnable d'étudier les propositions contrefactuelles sur l'ensemble des mondes possibles. On désire toujours pouvoir affirmer avec certitude [11] des propositions contrefactuelles évidemment vraies du type « Si Elsa avait sauté de la tour Eiffel, Elsa serait morte », que le formalisme initial de David Lewis permettent d'affirmer sans se préoccuper des mondes trop éloignés dans lesquels Elsa dispose de super-pouvoirs.

La topologie des mondes possibles de Lewis permet de restreindre les mondes possibles aux sous-ensembles pertinents. Considérons un monde i, et un ensemble $\$_i$ d'ensemble de mondes possibles vérifiant les conditions suivantes :

(C) $\$_i$ est centré en i, c'est-à-dire que $\{i\} \in \$_i$.

(1) Les sphères $S \in \$_i$ sont imbriquées (si $S, T \in \$_i$ alors $S \subset T$ ou $T \subset S$).

(2) $\$_i$ est clos sous union.

(3) $\$_i$ est clos sous intersection.

Nous remplaçons les deux opérateurs $\Box\!\!\rightarrow$ et $\Diamond\!\!\rightarrow$ par une famille d'opérateurs $\Box\!\!\rightarrow_p$ indexée par $p \in [0,1]$.

On lit $P \Box\!\!\rightarrow_p Q$ comme : « S'il était le cas que P, alors il serait le cas que Q avec une probabilité supérieure à p ». On a donc en particulier $\Box\!\!\rightarrow \equiv \Box\!\!\rightarrow_1$ et $\Diamond\!\!\rightarrow \equiv \bigcup_{p>0} \Box\!\!\rightarrow_p$

Pour $p \in [0,1]$, la proposition $P \Box\!\!\rightarrow_p Q$ est vraie si et seulement si l'une des deux conditions suivantes est vérifiée :

(1) Pour tout $S \in \$_i$, S ne contient pas de P-monde (ou : $\bigcup_{S\in\$_i} S$ ne contient pas de P-monde)

(2) Il existe $S \in \$_i$ contenant un P-monde tel que $\mathbb{P}(Q|P) \geq p$ [12] soit vraie dans S, ce que l'on note $\mathbb{P}_S(Q|P) \geq p$

11. ou plus précisément : « avec probabilité égale à 1 », ce qui n'équivaut pas exactement à la certitude logique dans le cas d'une infinité de mondes.

12. Cette « probabilité » correspond à la fréquence de réalisation de Q dans ces P-mondes

Le cas (1) permet toujours de rendre vraies les implications contrefactuelles avec antécédent toujours faux, tandis que le cas (2) s'interprète comme : « Dans des mondes suffisamment proches de i, quand P est vrai, Q est vrai avec probabilité au moins p ».

Revenons sur l'exemple de la proposition contrefactuelle « *Si les voisins n'avaient pas fait la fête, j'aurais dormi correctement.* » ($P \boxrightarrow Q$). Avec le formalisme probabiliste, on ne peut accorder une valeur de vérité qu'à des implications du type $P \boxrightarrow_p Q$, et cette valeur de vérité est encore binaire. De façon immédiate, c'est en prenant la borne supérieure des $p \in [0,1]$ tels que $P \boxrightarrow_p Q$ soit vraie qu'on peut alors définir une valeur de vérité *continue* à la contrefactuelle.

On définit donc $V(P,Q)$ comme étant la valeur de vérité de l'implication contrefactuelle. En particulier :

$$P \boxrightarrow Q \Leftrightarrow V(P,Q) = 1 \quad \text{et} \quad P \diamond\!\!\rightarrow Q \Leftrightarrow V(P,Q) > 0$$

Pour certaines contrefactuelles, il peut s'avérer utile pour mesurer à quel point P influence Q, de comparer $\mathbb{P}(Q|P)$ avec $\mathbb{P}(Q|{\sim}P)$ (A quel point ai-je mieux dormi que si mes voisins avaient fait la fête ? A quel point la soirée aurait été agréable si Bob était venu ?). Cela n'aurait cependant pas de sens de comparer $V(P,Q)$ à $V({\sim}P,Q)$, puisque le conditionnement se fait sur des mondes différents. En l'occurrence, si on étudie une contrefactuelle *stricte* (P et Q faux dans i), alors ${\sim}P \boxrightarrow {\sim}Q$ est vrai (car ${\sim}P \implies {\sim}Q$ dans tous les mondes de $\{i\}$) et $V({\sim}P,Q) = 0$.

Au lieu d'abandonner l'axiome (C) pour élargir la plus petite sphère aux mondes « très semblables » à i afin de donner une valeur non triviale à $V({\sim}P,Q)$, il semble intuitif de vouloir définir $V({\sim}P,Q) = \mathbb{P}_S(Q|{\sim}P)$, où l'ensemble de mondes S est le même que celui qui permettait d'obtenir la valeur $V(P,Q) = \mathbb{P}_S(Q|P)$. Nous sommes amenés à faire la même hypothèse simplificatrice que dans la première partie par commodité technique, car un tel ensemble de mondes S n'est pas nécessairement défini [13]. On considère que tel est toujours le cas, et on note $S(P,Q)$ l'ensemble de mondes vérifiant $V(P,Q) = \mathbb{P}_{S(P,Q)}(Q|P)$.

Finalement, on définit *l'influence causale contrefactuelle* $C(P,Q)$ de P sur Q par la relation :

$$C(P,Q) = \frac{\mathbb{P}_{S(P,Q)}(Q|P) - \mathbb{P}_{S(P,Q)}(Q|{\sim}P)}{\mathbb{P}_{S(P,Q)}(Q|P) + \mathbb{P}_{S(P,Q)}(Q|{\sim}P)}$$ [14]

Vérifions que cette définition fait sens avec notre exemple, et permet de retrouver un sens plus proche de l'implication contrefactuelle que la valeur

13. $V(P,Q)$ est défini comme une borne supérieure : il existe donc une famille $S(p)$ d'ensembles de mondes pour lesquels $V(P,Q) \geq \mathbb{P}_{S(p)}(Q|P)$ pour tout $p \in [0, V(P,Q)]$ sans nécessairement que la « limite » des ensembles de mondes $S(p)$ existe.

14. Par analogie avec le « contraste » en physique.

de vérité $V(P,Q)$.

Si dans tous les mondes (au sein de $S(P,Q)$) où les voisins font la fête ($\sim P$), je dors mal ($\sim Q$), alors $\mathbb{P}_{S(P,Q)}(Q|{\sim}P) = 0$ et $C(P,Q) = \frac{\mathbb{P}_{S(P,Q)}(Q|P)}{\mathbb{P}_{S(P,Q)}(Q|P)} = 1$, quand bien même $V(P,Q) = \mathbb{P}_{S(P,Q)}(Q|P) < 1$ [15] à cause de moustiques qui m'auraient empêché de dormir dans des mondes très proches. Ainsi, il n'est pas nécessaire que $P \boxrightarrow Q$ pour que $C(P,Q) = 1$.

Si, dans 10% des mondes (au sein de $S(P,Q)$) dans lesquels les voisins font la fête, j'arrive à dormir correctement, contre 90% dans les mondes où les voisins ne font pas la fête, alors $C(P,Q) = \frac{0.9-0.1}{0.9+0.1} = 0.8$.

S'il y a autant de mondes (au sein de $S(P,Q)$) dans lesquels j'arrive à dormir quand les voisins font la fête et quand ils ne la font pas, alors $C(P,Q) = 0$: le fait que les voisins fassent la fête n'influence ni positivement ni négativement mon sommeil (quand bien même on pourrait avoir $V(P,Q) > 0$, voire même $V(P,Q) = 1$ et donc $P \boxrightarrow Q$: si je dors toujours correctement par exemple).

S'il y a plus de mondes (au sein de $S(P,Q)$) dans lesquels j'arrive à dormir correctement quand les voisins font la fête, cette définition reste cohérente : on obtient simplement $C(P,Q) < 0$.

Dans tous les cas, $C(P,Q) \in [-1,1]$ semble mesurer *l'impact contrefactuel* de P sur Q de façon à la fois plus intuitive et nuancée que $P \boxrightarrow Q$ et $P \diamondrightarrow Q$.

2.2.2 Formulation avec la relation de similarité intermondaine

La topologie des mondes possibles que nous utilisons étant strictement identique à celle du formalisme initial de David Lewis, les sphères concentriques induisent toujours une relation d'ordre de similarité intermondaine.

Rappelons les conditions que doit vérifier $\leq_i$ définie sur un ensemble S_i de mondes :

(1) $\leq_i$ est transitive : si $j \leq_i k$ et $k \leq_i h$, $j \leq_i h$.

(2) $\leq_i$ est une relation d'ordre totale : $\forall j,k \in S_i, j \leq_i k$ ou $k \leq_i j$.

(3) $i \in S_i$ et $\forall j \in S_i, i \leq_i j$.

(C) $\leq_i$ est strictement minimal pour $\leq_i$: si $j \neq i, i <_i j$

On obtient alors ainsi la valeur de vérité de la proposition $P \boxrightarrow_p Q$:

(1) S_i ne contient pas de P-monde, ou

(2) Il existe un P-monde $k \in S_i$ tel que $\mathbb{P}_{j \leq_i k}(Q|P) \geq p$.

15. pour peu que $V(P,Q) > 0$, c'est-à-dire que $P \diamondrightarrow Q$.

La probabilité $\mathbb{P}_{j\leq_i k}(Q|P)$ est prise sur l'ensemble des mondes plus proches de i qu'un certain monde k.

Une fois $P \boxed{\rightarrow}_p Q$ défini, on en déduit $V(P,Q)$ et $C(P,Q)$ comme en 2.2.1.

On remonte de ce formalisme au système de sphères précédent en définissant $\$_i$ comme l'ensemble des parties S de S_i telles que pour tout $j \in S$ et $k \notin S, j <_i k$. On montre alors que $\$_i$ vérifie toutes les conditions définies en 2.2.1, et que les valeurs de vérité de $P \boxed{\rightarrow}_p Q$ sont équivalentes dans les deux formalismes.

Réciproquement, en définissant $\leq_i$ sur $S_i = \bigcup_{S\in\$_i} S$ par : $j \leq_i k$ si et seulement si pour toute sphère $S \in \$_i$, $k \in S \implies j \in S$, on montre que la relation $\leq_i$ vérifie les conditions définies ci-dessus, et que les valeurs de vérité de $P \boxed{\rightarrow}_p Q$ sont équivalentes dans les deux formalismes.

2.2.3 Formulation avec les fonctions de sélection

Ce formalisme nécessite de se placer dans l'hypothèse simplificatrice selon laquelle il existe une sphère $S_i^{(P)}$ minimale contenant des P-mondes pour toute proposition contrefactuelle (non contradictoire) P. Dans le cadre de cette hypothèse, nous allons utiliser les fonctions de sélection définis dans la partie précédente pour introduire de façon plus simple et plus intuitive les outils développés dans 2.2.1.

Pour rappel, on définit :

$$f_i(P) = \begin{cases} \{\text{les } P\text{-mondes appartenant à } S_i^{(P)}\} & \text{s'il existe des } P\text{-mondes} \\ \emptyset & \text{sinon} \end{cases}$$

On peut, à partir de cette fonction, donner la valeur de vérité de $P \boxed{\rightarrow}_p Q$ par :

$$P \boxed{\rightarrow}_p Q \quad \text{est vraie si et seulement si} \quad \mathbb{P}_{f_i(P)}(Q) \geq p$$

Mais ce n'est pas nécessaire, car on peut définir directement $V(P,Q)$ et $C(P,Q)$ par :

$$V(P,Q) = \mathbb{P}_{f_i(P)}(Q)$$

$$C(P,Q) = \frac{\mathbb{P}_{f_i(P)}(Q|P) - \mathbb{P}_{S_i^{(P)}}(Q|{\sim}P)}{\mathbb{P}_{f_i(P)}(Q|P) + \mathbb{P}_{S_i^{(P)}}(Q|{\sim}P)}$$

Réciproquement, Lewis montre que l'on peut remonter des fonctions de sélection au formalisme topologique initial (avec l'hypothèse simplificatrice), ce que nous admettrons ici.

2.3 Comparaison avec le formalisme non probabiliste

Nous avons déjà discuté dans la section 2.1 de l'intérêt d'ajouter une notion de probabilités pour pallier certaines conséquences contre-intuitives de la théorie de David Lewis. Dans cette section, nous vérifions que le formalisme « Lewis-probabiliste » résout effectivement ces problèmes.

2.3.1 Irréductibilité de « Lewis-probabiliste » au modèle de Lewis

Revenons sur l'exemple de proposition contrefactuelle que j'ai introduit pour justifier la nécessité d'enrichir la théorie de Lewis avec les probabilités : « *Si les voisins n'avaient pas fait la fête, j'aurais dormi correctement.* » que l'on note $P \boxrightarrow Q$.

J'ai prétendu que, dans $S_i^{(P)}$, on pouvait trouver des P-mondes dans lesquels les voisins ne font pas la fête, mais un moustique m'empêche de dormir, ce qui permet de trouver un $P\&{\sim}Q$-monde dans $S_i^{(P)}$, invalidant donc $P \boxrightarrow Q$.

On pourrait être tenté de répondre que les mondes dans lesquels les voisins ne font pas la fête ET un moustique est présent dans ma chambre sont tous strictement plus éloignés de notre monde que ceux où la « seule différence » est l'absence de bruit de mes voisins. Il n'y aurait donc, par construction, pas de $P\&R$-mondes dans $S_i^{(P)}$ (où R désigne la présence d'un moustique dans ma chambre). Pourtant, dans le cas où $P \boxrightarrow Q$, il n'y a pas une mais au moins deux propositions qui changent par rapport au monde i : ${\sim}P \leftarrow P$ ainsi que ${\sim}Q \leftarrow Q$. Dans les mondes où les voisins n'ont pas fait la fête mais un moustique m'a empêché de dormir, il n'y a aussi (au minimum) que deux propositions qui sont modifiées : ${\sim}P \leftarrow P$ et $R \leftarrow {\sim}R$, puisqu'on conserve ${\sim}Q$ dans ces mondes.

Comme la distance intermondaine, donnée préalablement à toute étude contrefactuelle, n'a pas de raison d'être calibrée pour rendre les Q-mondes plus proches que les R-mondes, on ne peut pas utiliser cet argument pour disqualifier d'office la nécessité de prendre en considération les R-mondes au sein de $f_i(P)$. Il n'y a pas de raison que les mondes où je dors correctement (et il n'y a, comme dans i, pas de moustique) soient plus éloignés que ceux dans lesquels il y a un moustique dans ma chambre (et où, comme dans i, je dors mal).

De même, en reprenant l'exemple de Kit (« Si Nixon avait appuyé sur le bouton (P), il y aurait eu un holocauste nucléaire (Q) »), les P-mondes les plus proches incluent probablement des mondes dans lesquels un dysfonctionnement empêche l'holocauste nucléaire (et c'est d'autant plus vrai avec cet exemple car ces mondes sont *justement* plus proches du nôtre du fait que l'holocauste ne s'est pas produit).

2.3.2 Les propositions contrefactuelles binaires

Pour l'instant, considérons que le fait de dormir « bien » ou « mal » est une information binaire, qui se traduit par D ou $\sim D$. On note (de façon plus intuitive...) respectivement V et M la présence de voisins bruyants et de moustiques dans ma chambre (pour une nuit donnée).

On suppose que i est un $(V\&\sim M\&\sim D)$-monde, c'est-à-dire un monde dans lequel mes voisins ont fait la fête, il n'y a pas de moustique dans ma chambre et j'ai mal dormi la nuit.

Nous avons vu (cf. 2.1.1) que dans le formalisme de Lewis, on ne peut pas affirmer que $\sim V \boxrightarrow D$ dès qu'il y a des $(\sim V\&M)$-mondes dans $S_i^{(\sim V)}$, c'est-à-dire, avec la reformulation en terme de similarité intermondaine, s'il existe un monde j avec moustique et voisins silencieux vérifiant $j \leq_i k$ pour tout monde k avec voisins silencieux dans lesquels j'ai bien dormi la nuit. La valeur de vérité de $\sim V \boxrightarrow D$ dépend donc de l'affectation initiale de distances intermondaines, ce qui est normal (et admis par Lewis), mais pas d'une façon satisfaisante.

En effet, quand on veut étudier la valeur de vérité de l'implication contrefactuelle $P \boxrightarrow Q$, les « interférences » dues à des propositions R indépendantes de P mais qui influent également sur Q devraient pouvoir être évitées, dans la mesure du possible (sauf à ne considérer que des contrefactuelles pour lesquelles Q est une conséquence logique de P, mais Lewis est très loin de se restreindre à celles-ci, même avec l'utilisation de l'opérateur $\boxrightarrow$).

Le formalisme Lewis-probabiliste, quant à lui, permet de quantifier l'influence de $\sim V$ sur D, selon un sens qui se rapproche le plus de notre idée intuitive de « causalité ». L'influence causale contrefactuelle $C(\sim V, D)$ détermine, dans toutes les configurations [16], une valeur numérique qui ne correspond pas simplement à la probabilité conditionnelle « Aurais-je dormi correctement si mes voisins n'avaient pas fait la fête ? », mais qui fait la différence avec la façon dont j'ai dormi dans les mondes similairement proches où mes voisins ont été bruyants. Le ratio permet ensuite d'éliminer les « interférences » dues aux causes indépendantes comme M, qui influent les probabilités conditionnelles à la fois au numérateur et au dénominateur. Dans notre exemple :

$$C(\sim V, D) = \frac{\mathbb{P}_{S(\sim V,D)}(D|\sim V) - \mathbb{P}_{S(\sim V,D)}(D|V)}{\mathbb{P}_{S(\sim V,D)}(D|\sim V) + \mathbb{P}_{S(\sim V,D)}(D|V)} = \frac{\mathbb{P}_{S(\sim V,D)}(D|\sim V)}{\mathbb{P}_{S(\sim V,D)}(D|\sim V)} = 1$$

en considérant que dans tous les mondes assez proches où les voisins ont fait la fête, je n'ai pas dormi correctement. [17]

16. cf. 2.2.1.
17. Si ce n'est pas non plus le cas pour d'autres raisons, l'impact contrefactuel est un peu inférieur à 1 mais peut rester largement supérieur à $\mathbb{P}_{S(\sim V,D)}(D|\sim V)$.

2.3.3 Les propositions contrefactuelles quantitatives

Pour répondre à la question « Combien de temps aurait duré la République romaine si César n'avait pas franchi le Rubicon ? », on a vu que le formalisme de Lewis était non seulement peu adapté (car cette question ne correspond à aucune contrefactuelle du type $P \boxdot\!\!\rightarrow Q$), et donne même des résultats intuitivement très douteux si l'on s'efforce d'étudier les propositions contrefactuelles R_x qui font sens au sein de cette théorie (cf. 2.1.3.).

Plaçons-nous dans $f_i(\sim P)$ (l'ensemble des $\sim P$-mondes de $S_i^{(\sim P)}$, c'est-à-dire les mondes les plus proches dans lesquels César n'a pas franchi le Rubicon). Il s'agit d'un ensemble de mondes dans lesquels la durée de vie de la République romaine prend des valeurs distinctes. Il est naturel de définir X comme la variable aléatoire indiquant la durée de vie de la République romaine après César afin de calculer $\mathbb{E}_{f_i(\sim P)}(X) = \mathbb{E}_{S_i^{(\sim P)}}(X|\sim P)$, la durée de vie moyenne de la République romaine dans les mondes les plus proches où César n'a pas franchi le Rubicon.

Cette information ne suffit pas à mesurer l'impact contrefactuel du franchissement du Rubicon par César sur la durée de vie de la République romaine. Il faut comparer cette valeur à la durée de vie moyenne de la République romaine dans les mondes les plus proches où César a franchi le Rubicon (et pas au sens de Lewis, puisque ces mondes se réduisent au nôtre, étant donné que l'on conditionne cette fois par une proposition P s'étant réalisée). Comme pour les probabilités, on conditionne donc au sein de $S_i^{(\sim P)}$ afin de calculer $\mathbb{E}_{S_i^{(\sim P)}}(X|P)$, l'espérance de vie de la République romaine dans les P-mondes de $S_i^{(\sim P)}$. On calcule enfin :

$$C(\sim P, X) = \frac{\mathbb{E}_{S_i^{(\sim P)}}(X|\sim P) - \mathbb{E}_{S_i^{(\sim P)}}(X|P)}{\mathbb{E}_{S_i^{(\sim P)}}(X|\sim P) + \mathbb{E}_{S_i^{(\sim P)}}(X|P)}$$ [18]

L'influence causale contrefactuelle s'interprète de la même façon que dans le cas précédent. Vérifions que cette définition semble rendre compte de l'intuition.

Si dans tous les mondes (au sein de $S_i^{(\sim P)}$) où César franchit le Rubicon (P), la République romaine ne lui survit pas alors $C(\sim P, X) = 1$ et ce peu importe la durée de vie de moyenne $\mathbb{E}_{S_i^{(\sim P)}}(X|\sim P)$ de la République romaine dans les mondes où César ne franchit pas le Rubicon (pour peu que $\mathbb{E}_{S_i^{(\sim P)}}(X|\sim P) > 0$).

18. J'utilise la même notation qu'en 2.2.1 pour l'influence causale contrefactuelle $C(P, Q)$, mais X est désormais une variable aléatoire. Il s'agit en réalité d'une généralisation, puisqu'on retrouve la définition précédente en posant $X = \mathbf{1}_Q$.

Si la République romaine dure en moyenne 10 ans après la mort de César dans les mondes (au sein de $S_i^{(\sim P)}$) dans lesquels César franchit le Rubicon contre 90 ans dans les mondes où il ne le franchit pas, alors $C(\sim P, X) = 0.8$.

Si la République romaine dure en moyenne aussi longtemps après la mort de César dans les mondes (au sein de $S_i^{(\sim P)}$) dans lesquels César franchit le Rubicon et ceux dans lesquels il ne le franchit pas, alors $C(\sim P, X) = 0$.

Si la République romaine dure en moyenne plus longtemps après la mort de César dans les mondes (au sein de $S_i^{(\sim P)}$) dans lesquels César franchit le Rubicon que ceux dans lesquels il ne le franchit pas, alors $C(\sim P, X) < 0$.

Conclusion

L'analyse du contrefactuel est une problématique importante en philosophie contemporaine. Outre son intérêt évident pour l'étude métaphysique du *possible*, les théories morales conséquentialistes (telles que l'utilitarisme) nécessitent de mesurer *l'impact contrefactuel* d'une action. Les débats sur le libre-arbitre sont aussi intimement liés au concept du possible, au point que les positions compatibilistes et anti-compatibilistes peuvent (en partie) se réduire à des questions sur l'interprétation du concept de contrefactuel [19].

Dans la lignée de Wittgenstein, pour qui seuls les *faits* (qui peuvent se traduire [20] en des propositions) caractérisent un monde, les philosophes analytiques à partir de Carnap généralisent cette intuition pour définir des mondes possibles, qui diffèrent du nôtre uniquement par la valeur de vérité de ces propositions. La logique modale semble être le cadre naturel pour étudier les propositions contrefactuelles.

L'analyse en terme de conditionnelles strictes se montre cependant insatisfaisante du fait de divergences entre les propriétés attendues attendues de l'implication contrefactuelle par rapport à l'implication formelle (comme le paradoxe du renforcement des antécédents). David Lewis introduit (comme Robert Stalnaker) une notion de « similarité » topologique entre mondes et offre ainsi un cadre adéquat pour appliquer les outils de la logique modale, en se restreignant à des sous-ensembles de mondes dans lesquels il n'est pas automatiquement tautologique ou contradictoire de parler de nécessité et de possibilité.

Les théories les plus récentes (réseaux bayésiens, modélisation par équations structurelles) développées vers le début des années 2000 préfèrent l'utilisation des probabilités conditionnelles comme outil de base de l'analyse du contrefactuel, en se concentrant sur certaines relations entre les faits, plutôt que sur les similarités entre les mondes. Ces modèles, qui s'inspirent des sciences cognitives et informatiques, offrent des prédictions claires sur des propositions contrefactuelles particulières, sans se préoccuper de considérations métaphysiques (autant que possible).

En incorporant les probabilités au sein même du formalisme de Lewis, on résout certains paradoxes de sa théorie. Même si cette nouvelle version ne permet évidemment toujours pas de fournir d'algorithme pratique pour

19. Voir https ://plato.stanford.edu/entries/counterfactuals/AgenMindRati
20. pas forcément dans le langage...

déterminer la valeur de vérité des propositions contrefactuelles (contrairement aux modèles d'équations structurelles), *l'influence causale contrefactuelle* ainsi définie fournit des conditions abstraites de vérité pour les propositions contrefactuelles plus satisfaisantes (ou du moins, plus conformes à l'intuition) que les opérateurs de nécessité et de possibilité de Lewis, pas assez flexibles pour capturer leur complexité.

Bibliographie

[1] David LEWIS, *Counterfactuals*, Blackwell Publishers, Malden, 1986. : https ://perso.uclouvain.be/peter.verdee/counterfactuals/lewis.pdf

[2] Rudolf CARNAP, *Signification et nécessité*, Paris Gallimard, 1997. Traduit de l'anglais par F. Rivenc et Ph. de Rouilhan.

[3] Saul KRIPKE, *La Logique des noms propres*, Editions de Minuit, Paris, 1980. Traduit de l'anglais par François Recanati.

[4] M. J. CRESSWELL, *The world is everything that is the case*, Cornell University Press, Ithaca, 1979.

[5] Robert C. STALNAKER, *Possible Worlds*, Cornell University Press, Ithaca, 1979.

[6] Stanford Encyclopedia of Philosophy, Article *Counterfactuals*, 2019.

[7] Kit FINE, "Review of Lewis' Counterfactuals", *Mind*, 1975.

[8] Judea PEARL, "Causation, Action, and Counterfactuals", dans *Computational Learning and Probabilistic Reasoning*, 1995.