David Lewis : La causalité et les contrefactuels

Alexandre Marcellesi*
(University of California, San Diego)

I. Introduction

Vous vous souviendrez sans doute de la faillite de la banque américaine Lehman Brothers en septembre 2008. Une de ses causes est la fameuse crise dite « des subprimes » qui a touché les États-Unis entre 2007 et 2008. Mais que voulons-nous dire exactement quand nous disons que l'un de ces évènements est une *cause* de l'autre ? Et quelle est la nature de la relation dont nous assertons ainsi l'existence ? Un des projets poursuivi par David Lewis, dès 1973 et jusqu'à son décès en 2001, était de développer une analyse de notre concept de causalité. Nous formulons des jugements causaux de façon routinière. C'est par exemple ce que j'ai fait plus haut à propos de la faillite de Lehman Brothers. Formuler de tels jugements implique que nous utilisions le concept de causalité que nous possédons. Et une des manières de cerner les contours de ce concept est d'en formuler une définition, c'est-à-dire une liste de conditions à la fois nécessaires et suffisantes pour son application, et de la confronter à nos jugements causaux. Telle est la méthode adoptée par Lewis pour développer son projet d'analyse conceptuelle de la causalité[1]. L'analyse conceptuelle de la causalité n'est cependant pas le but ultime pour Lewis. Elle est plutôt une étape nécessaire sur le chemin vers une réduction de la causalité. Si vous parvenez à formuler une définition de la causalité qui est à la fois adéquate et réductive en ce sens que son *definiens* ne contient pas de notions

* Alexandre Marcellesi est étudiant doctorant à l'Université de Californie, San Diego (États-Unis) et ancien étudiant de l'Université Paris-Sorbonne et de l'École Normale Supérieure. Ses recherches portent sur la causalité dans les sciences sociales, en particulier la statistique et l'économie empirique. Son article le plus récent ('External Validity : Is There Still a Problem ?', *Philosophy of Science*, à paraitre) traite du problème de la validité externe dans l'évaluation des politiques publiques.

[1] Cf. Ludwig (2012) pour une description bien plus détaillée et précise de la méthodologie de l'analyse conceptuelle chez Lewis.

causales, alors vous pourrez de manière légitime soutenir la thèse que la relation à laquelle notre concept de causalité réfère n'est pas primitive[2].

Le but de cet article est avant tout introductif. Il vise à décrire, de manière brève et incomplète, les tentatives par Lewis de définir la causalité. Je décrirai, et critiquerai, également une tentative plus récente par Christopher Hitchcock (2007) s'inscrivant dans la tradition inaugurée par Lewis. Comme je l'expliquerai plus bas, les projets de Lewis et Hitchcock sont différent, et ce bien qu'ils adoptent la même méthode. Lewis est-il parvenu à ses fins en ce qui concerne le concept de causalité ? Non, et j'expliquerai plus bas les raisons de cet échec. Peut-on espérer qu'un de ses successeurs réussisse là où Lewis a échoué ? Je défendrai, en m'appuyant sur les arguments avancé dans Glymour et al. (2010), la thèse selon laquelle cet espoir est vain, notamment en raison de la méthode adoptée par Lewis. Et comme je l'expliquerai, le pessimisme est également de rigueur pour les philosophes, y compris Hitchcock, qui adoptent la méthode de Lewis tout en poursuivant d'autres buts.

Il m'est à présent nécessaire de faire les quatre remarques suivantes. Premièrement, la relation de causalité qui intéresse Lewis est une relation entre évènements actuels, les évènements étant des particuliers occupant une position dans l'espace et dans le temps. Cette relation est aujourd'hui souvent appelée « causalité actuelle » ou « causalité singulière » afin de la distinguer d'autres relations causales. J'utiliserai donc le mot « causalité » dans cet article pour désigner cette relation en particulier, à l'exclusion des autres relations causales. Deuxièmement, je suivrai Lewis en me limitant aux cas dans lesquels les causes garantissent l'occurrence de leurs effets et en ignorant donc les cas dans lesquels elles ne font qu'en modifier les probabilités. Troisièmement, j'ignorerai ici les deux questions suivantes, et ce bien que Lewis les ai traitées : La causalité est-elle transitive ? Les absences et omissions peuvent-elle être des causes ? Quatrièmement, j'ignorerai ici les critiques, émises par exemple par Phil Dowe (2000, Chapitre 1) ou James Woodward (2014), ayant pour objet non pas les définitions offertes par Lewis ou Hitchcock mais le projet même d'une analyse conceptuelle de la causalité.

Cet article n'a pas pour but d'être une présentation exhaustive des thèses défendues par Lewis à propos de la causalité, des objections soulevées contre ces thèses ou des réponses avancées par Lewis lui-même ou par d'autres. Le nombre de livres et articles sur le sujet est tel qu'une pareille ambition serait déraisonnable. Pour les lecteurs souhaitant s'immerger dans cette littérature, je

[2] Cf. Ludwig (2012) également sur la méthode de la « réduction fonctionnelle ».

recommande de commencer par lire (Nolan, 2005, Chapitre 4) et surtout (Menzies, 2014). Je recommande également (Kistler, 2011) pour une introduction, en français, rigoureuse et détaillée aux différents débats entourant la notion de causalité – y compris celui impliquant Lewis et dont je traiterai plus bas.

II. La définition de 1973

Que voulons-nous dire quand nous disons qu'un évènement *c* est une *cause* d'un autre évènement *e ?* Comment doit-on comprendre le concept de causalité employé dans un tel jugement ? L'approche défendue par Lewis est basée sur l'intuition exprimée dans le passage suivant : « Nous pensons à une cause comme faisant une différence, et la différence que fait cette cause doit être une différence par rapport à ce qui se serait passé en son absence » (Lewis, 1973a, 557)[3]. Ce que nous voulons dire quand nous disons que la crise des *subprimes* est une cause de la faillite de Lehman Brothers, par exemple, c'est que l'occurrence du premier de ces évènements a « fait une différence » en ce qui concerne l'occurrence du second : si la crise des *subprimes* n'avait pas eu lieu, Lehman Brothers n'aurait pas fait faillite. Mais plusieurs étapes séparent cette intuition d'une définition en bonne et due forme de la causalité. Le reste de cette section est une présentation détaillée des étapes menant à la définition proposée par Lewis en 1973 (Lewis, 1973a).

Étape 1. Dans la mesure où la définition de la causalité offerte par Lewis implique la notion de conditionnel contrefactuel, la première étape vers cette définition est de déterminer les conditions de vérité des propositions contrefactuelles[4]. La proposition contrefactuelle correspondant à une paire de propositions *A* et *B*, notée *A>B*,est la proposition que si *A* était le cas, alors *B* serait le cas. Dans quelles conditions la proposition *A>B* est-elle vraie ? Lewis prend au pied de la lettre l'intuition selon laquelle les propositions contrefactuelles ont pour sujet des possibilités non-actuelles et propose donc

[3] Toutes les traductions de l'anglais vers le français sont de moi-même.

[4] Il est à noter que, pour Lewis, les propositions ne sont pas des entités linguistiques. La proposition que *p* est pour Lewis l'ensemble des mondes possibles dans lesquels il est vrai que *p*. La proposition que la neige est blanche, par exemple, est identique à l'ensemble des mondes possibles dans lesquels la neige est blanche. Les phrases « la neige est blanche » et « *snow is white* » expriment toutes deux la proposition que la neige est blanche parce qu'elles sont vraies dans tous les mondes possibles dans lesquels la neige est blanche (et seulement dans ceux-ci).

d'utiliser la notion de monde possible (Lewis, 1986a, Chapitre 1) afin de spécifier les conditions de vérité de ces propositions[5].

Une proposition contrefactuelle $A>B$ est vraie dans un monde w dans deux cas. Appelons un monde dans lequel une proposition A est vraie un monde-A. Le premier cas est celui dans lequel l'antécédent de $A>B$ est impossible du point de vue de w[6]. S'il est impossible que Nicolas Sarkozy soit un mouton, alors, selon Lewis, il est (trivialement) vrai que si Sarkozy était un mouton, il produirait 7 kilos de laine par an. Il est à noter que le monde d'évaluation des contrefactuels sera ici, sauf mention contraire, le monde actuel, dénoté par le symbole @.

Qu'en est-il de la valeur de vérité d'un contrefactuel $A>B$ lorsqu'il existe des mondes-A possibles ? Appelons un monde dans lequel A et B sont toutes deux vraies un monde-$A\&B$. Et appelons un monde dans lequel B est fausse un monde-$\neg B$. Un contrefactuel $A>B$ dont l'antécédent est possible est vrai dans un monde w si et seulement si (ssi) tous les mondes-A les plus proches de w sont aussi des mondes-B[7]. Autrement dit, $A>B$ est vrai dans w ssi il existe un monde-$A\&B$ qui est plus proche de w que ne le sont les mondes-$A\&\neg B$. Quelle est ici la signification du mot « proche » ? Selon Lewis, un monde w_1 est plus proche de w que ne l'est un autre monde w_2 ssi w_1 est plus semblable à w que ne l'est w_2.

Bien que la position officielle de Lewis soit que la relation de similitude entre mondes possibles est primitive, il privilégie néanmoins certains types de similitudes – notamment la similitude quant aux lois de la nature ou quant aux états de faits particuliers. Je reviendrai sur ce point plus bas, car il deviendra important lorsque j'examinerai la manière dont Lewis répond aux contrexemples putatifs avancés contre sa définition. J'ajouterai simplement, pour le moment, que, selon Lewis, pour tout monde w, le monde le plus proche de w est w lui-même. Bien que cette thèse ait été critiquée (List et Menzies 2009, 484-486), elle semble intuitive : Comment un monde w ($w \neq @$) pourrait-

[5] J'ignorerai ici les débats relatifs au statut ontologique de ces mondes possibles. Cf. Chauvier (2012) pour une introduction et Divers (2002) pour un traitement très détaillé.

[6] Pour Lewis, la notion de possibilité, tout comme celle de vérité, est relative à un monde possible. Il est possible pour l'antécédent d'un contrefactuel d'être impossible du point de vue d'un monde w_1 mais possible du point de vue d'un autre monde w_2. Je laisserai ici la relativisation de la notion de possibilité implicite. Il est aussi important de noter que, pour Lewis, la notion de possibilité qui est ici appropriée est celle de possibilité *métaphysique.*

[7] Pour Lewis, il est possible que plusieurs mondes-A soient aussi proches l'un que l'autre de w. Lewis (1973b, 20-21) rejette donc ce qu'il appelle le « principe de limite ».

il être aussi semblable (a fortiori, plus semblable) à @ que @ ne l'est lui-même ?

Il est temps d'illustrer l'analyse que propose Lewis des conditions de vérité des contrefactuels par un exemple. En supposant que son antécédent soit possible, dans quelles conditions la proposition que si Sarkozy faisait 2m15, il ne porterait pas de talonnettes est-elle vraie ? Selon Lewis, ce contrefactuel est vrai ssi tous les mondes les plus proches dans lesquels Sarkozy fait 2m15 sont aussi des mondes dans lesquels il ne porte pas de talonnettes. Autrement dit, ce contrefactuel est vrai ssi il existe un monde dans lequel Sarkozy fait 2m15 et ne porte pas de talonnettes qui est plus semblable à @ que ne le sont les mondes dans lesquels Sarkozy fait 2m15 mais porte néanmoins des talonnettes. Maintenant que l'analyse des contrefactuels que propose Lewis a été brièvement présentée, nous pouvons passer à la seconde étape sur le chemin vers sa définition de la causalité.

Étape 2. Appelons deux proposition *A* et *B compossibles* ssi il est possible qu'elles soient vraies en même temps. Une famille de propositions B_1, B_2, …, B_n, telle que toute paire dans cette famille est incompossible, *dépend contrefactuellement* d'une famille semblable A_1, A_2, …, A_n ssi il est vrai que $A_i > B_i$ pour $i = 1, \ldots, n$.

Comme je l'ai indiqué plus haut, le but de Lewis est une définition de la causalité entendue comme étant une relation entre évènements. Je n'ai pour l'instant parlé que de propositions et pas d'évènements. Il est cependant aisé de connecter propositions et évènements. D'après Lewis (1973a, 562), « À chaque évènement possible *e* correspond une proposition *O*(*e*) qui est vraie exactement dans les mondes dans lesquels *e* advient. Cette proposition *O*(*e*) est la proposition que *e* advient ». Étant donnée cette connexion entre propositions et évènements, il est aisé d'étendre la définition de la notion de dépendance contrefactuelle aux évènements. On dira qu'une famille d'évènements e^*_1, e^*_2, …, e^*_n (satisfaisant la condition d'incompossibilité) dépend contrefactuellement d'une famille semblable e_1, e_2, …, e_n ssi la famille de proposition $O(e^*_1)$, $O(e^*_2)$, …, $O(e^*_n)$ dépend contrefactuellement de la famille $O(e_1)$, $O(e_2)$, …, $O(e_n)$. Pourquoi s'intéresser à la relation de dépendance contrefactuelle quand le but est d'arriver à une définition de la causalité ? La décision de Lewis d'identifier dépendance contrefactuelle et dépendance causale est le sujet de la prochaine étape.

Étape 3. Comme le note Lewis (1973a, 561), la dépendance contrefactuelle entre « de larges familles d'alternatives est typique des

procédures de mesure, de perception ou de contrôle ». La température indiquée par un thermomètre fonctionnant correctement, par exemple, dépend contrefactuellement de la température du milieu entourant le thermomètre. Et nous inférons naturellement, sur la base de ce constat, qu'il existe une relation causale entre l'un et l'autre. C'est en s'appuyant sur cette idée que des « variations concomitantes », pour utiliser une expression due à John Stuart Mill, à travers un ensemble de circonstances contrefactuelles sont le signe de la présence d'une relation causale que Lewis propose d'identifier dépendance contrefactuelle et dépendance causale. La prochaine étape sur le chemin vers la définition de Lewis est d'expliquer en quoi consiste la dépendance causale entendue comme une relation entre évènements individuels plutôt qu'entre familles d'évènements.

Étape 4. Prenons deux évènements *c* et *e* et appelons $\neg O(e)$ la proposition que l'évènement *e* n'advient pas. Selon Lewis, *c* dépend causalement de *e* ssi la famille $O(e)$, $\neg O(e)$ dépend contrefactuellement de la famille $O(c)$, $\neg O(c)$, c'est-à-dire ssi les deux contrefactuels suivants sont vrais :

(i) $O(c) > O(e)$

(ii) $\neg O(c) > \neg O(e)$

Considérons le cas dans lequel *c* et *e* sont des évènements actuels et dans lequel $O(c)$ et $O(e)$ sont donc actuellement vraies. Dans ce cas-ci, (i) est vrai : dans la mesure où aucun monde n'est aussi proche de @ que ne l'est @ lui-même, le monde-$O(c)$ le plus proche de @ est @ lui-même, et ce monde est aussi un monde-$O(e)$. Autrement dit, lorsque deux propositions *A* et *B* sont vraies dans un monde *w*, les contrefactuels $A>B$ et $B>A$ sont également vrais dans w[8].

Prenons l'exemple de la crise des *subprimes* et de la faillite de Lehman Brothers. Ce sont là, hélas, deux évènements actuels. Selon l'analyse que donne Lewis des contrefactuels, cela implique la vérité de la proposition que si la crise des *subprimes* avait lieu, Lehman Brothers ferait faillite. S'il est aussi vrai que si la crise des *subprimes* n'avait pas eu lieu, Lehman Brothers n'aurait pas faillite, alors la faillite de Lehman Brothers dépend causalement de la crise des *subprimes*. Mais la dépendance causale entre évènements individuels n'est pas encore la causalité. La prochaine, et dernière, étape nous amènera de la dépendance causale à la causalité.

[8] Ceci implique que les propositions contrefactuelles, malgré leur nom, peuvent avoir des antécédents qui sont vrais et qui ne contredisent donc pas les faits.

Étape 5. Selon Lewis, la dépendance causale entre évènements actuels implique la causalité. Si *c* et *e* adviennent dans le monde actuel et si *e* dépend causalement de *c*, alors *c* est une cause de *e*. Si la proposition que si la crise des *subprimes* n'avait pas eu lieu, Lehman Brothers n'aurait pas fait faillite est vraie, alors la crise des *subprimes* est, selon la définition qu'offre Lewis en 1973, une cause de la faillite de Lehman Brothers. Il est important de noter que la crise est *subprimes* est *une* cause, et non *la* cause, de la faillite de Lehman Brothers. Cette faillite a d'autres causes, notamment les mauvaises décisions à répétition des dirigeants de Lehman Brothers. Si l'on s'en tient à la définition qu'offre Lewis, la question de savoir laquelle de ces causes est *la* cause de la faillite est vide de sens. La distinction entre causes et conditions préalables est, pour Lewis, de nature pragmatique.

En avons-nous fini avec la définition de la causalité qu'offre Lewis en 1973 ? Pas tout à fait. Lewis soutient la thèse selon laquelle la causalité est transitive : si *c* est une cause de *d* et *d* est une cause de *e*, alors *c* est une cause de *e*[9]. Mais, comme le montre Lewis (1973b, 32-35), la dépendance contrefactuelle, et donc causale, n'est pas transitive. Cette difficulté est cependant aisément contournée. Appelons une séquence finie d'évènements e_1, e_2, e_3, …, e_n une *chaîne causale* ssi e_2 dépend causalement de e_1, e_3 dépend causalement de e_2 et ainsi de suite jusqu'à e_n. Lewis propose de définir la *causalité* de la manière suivante : un évènement actuel *c* est une cause d'un autre évènement actuel *e* ssi il existe une chaîne causale menant du premier de ces évènements au second. Autrement dit, la causalité est la relation « ancestrale » de la relation de dépendance causale : si *c* est une cause de *e*, alors *c* est un « ancêtre » de *e* dans une chaîne causale.

Comment évaluer la définition de la causalité offerte par Lewis en 1973 ? Comme je l'ai indiqué plus haut, Lewis souhaite arriver à une définition qui soit à la fois adéquate et réductive. Commençons donc par examiner la question de savoir si la définition qu'il avance est adéquate. Dans la mesure où Lewis cherche à définir notre concept de causalité, il convient de confronter sa définition aux jugements, ou « opinions naïves » (Lewis, 1973a, 567, n. 12), que nous formons à l'aide de ce concept. Les cas dans lesquels nos intuitions sont en conflit avec la définition de Lewis constituent des contrexemples à cette définition, alors que les cas dans lesquels elles sont en accord appuient la thèse selon laquelle cette définition est adéquate.

[9] Cf. Hitchcock (2001) pour une critique, parmi d'autres, de cette thèse.

Supposons que *c* est une cause de *e* et, qui plus est, que *e* n'a pas d'autre cause possible. Dans ce cas-ci, il semble vrai que si *e* n'était pas advenu, *c* ne serait pas advenu non plus. Ceci implique, selon la définition de Lewis, que *e* est une cause de *c*. Mais la causalité est une relation dont nous pensons le plus souvent qu'elle est asymétrique : si *c* est une cause de *e*, alors *e* n'est pas une cause de *c*. Pourquoi penser que la causalité est une relation asymétrique ? Notamment parce que nous pensons aussi que les causes précèdent leurs effets dans le temps. Si *c* est une cause de *e*, alors *c* précède *e* dans le temps. Et si *e* suit *c* dans le temps, *e* ne peut, par définition, pas être une cause de *c*. Le cas décrit plus haut semble donc être un contrexemple à la définition de Lewis. Ceci est, selon l'expression de Lewis, le *problème des effets*.

Voici un autre cas problématique, que Lewis appelle le *problème des épiphénomènes*. Supposons que *c* est une cause commune de *e* et *f* mais que *e* n'est pas une cause de *f* et *f* n'est pas une cause de *e*. Supposons également que *f* est postérieur à *e* dans le temps. Si l'on suppose, de plus, que *c* est la seule cause possible de *e* et de *f*, alors il semble vrai que si *e* n'était pas advenu, *f* ne serait pas advenu non plus. La raison en est que si *e* n'était pas advenu, alors *c* ne serait pas advenu, ce qui implique que *f* ne serait pas advenu non plus. Le problème est que si *f* dépend contrefactuellement de *e* alors, selon la définition de Lewis, *e* est une cause de *f*, ce qui contredit notre postulat de départ.

Lewis défend une solution unique aux deux problèmes des effets et des épiphénomènes. Selon Lewis, « la solution correcte à apporter à ces deux problèmes est simplement, je pense, de nier les contrefactuels qui en sont la source » (1973a, 566). Prenons le problème des effets. J'ai dit plus haut que, étant données les suppositions que fait Lewis au sujet des évènements *c* et *e*, il semble vrai que si *e* n'était pas advenu, alors *c* ne serait pas advenu. Lewis nie tout bonnement que ce soit le cas. Comment son affirmation se justifie-t-elle ? Souvenez-vous que la valeur de vérité d'un contrefactuel dépend de la relation de similitude entre mondes possibles que l'on considère. Dans la mesure où nous avons supposé que *c* et *e* sont deux évènements actuels, *c* dépend causalement de *e* ssi le monde-¬*e* le plus proche de @ est un monde-¬*c*. Et c'est précisément cette affirmation que Lewis rejette dans le passage suivant :

> Il est faux de dire que si *e* n'était pas advenu alors *c* ne serait pas advenu non plus… Il est plutôt le cas que *c* serait advenu tout comme dans le monde actuel mais n'aurait pas causé *e*. On s'éloigne moins du monde actuel en se débarrassant de *e* tout en gardant *c* fixe – et en abandonnant certaines des lois de la nature et des circonstances qui font que *c* ne peut, dans le monde actuel,

> échouer à causer *e* – qu'en conservant ces lois et circonstances et en se débarrassant de *e* en remontant le temps de manière à éliminer *c*. (Lewis, 1973a, 566).

Ce que Lewis nie dans cette citation c'est qu'il faille adopter une interprétation qu'il appelle « rétrograde » – car remontant dans le temps – du contrefactuel $\neg e > \neg c$ (Lewis, 1979, 457). Qu'est-ce qui justifie la position adoptée par Lewis dans le passage que je viens de citer ?

Selon l'analyse des lois de la nature que défend Lewis (1983, 365-368), les lois déterminent (logiquement) le cours de l'histoire – elles déterminent quels évènements adviennent où et quand. Un monde possible dans lequel *e* n'advient pas est donc un monde possible dont les lois ne sont pas celles du monde actuel. Autrement dit, on doit violer les lois de la nature – par ce que Lewis (1973a, 560) appelle un « miracle » – pour passer du monde actuel à un monde possible dans lequel *e* n'advient pas. Supposons que, dans le monde actuel, *c* advient à l'instant t_0 et *e* à l'instant t_1. Et comparons les deux mondes possibles suivants. Dans le monde w_1, le miracle qui est requis advient à l'instant t_0-ε, très peu de temps avant t_0. Dans le monde w_2, en revanche, le miracle en question advient à l'instant t_1-ε, très peu de temps avant t_1 mais bien après t_0.

La thèse défendue par Lewis est que w_2 est plus semblable à @ que ne l'est w_1. Ces deux mondes diffèrent de @ en ce qu'ils ne partagent pas ses lois de la nature. Mais l'histoire de w_2 est identique à celle de @ de son point de départ jusqu'à l'instant t_1-ε alors que celle de w_1 n'est identique à celle de @ que jusqu'à l'instant t_0-ε. Si les mondes tels que w_2 sont plus semblables à @ que ne sont les mondes tels que w_1, alors tous les mondes-$\neg e$ les plus proches de @ sont des mondes-*c*, qui qui implique la fausseté du contrefactuel $\neg e > \neg c$. Et si ce contrefactuel est faux, la définition de Lewis n'implique pas que *c* dépende causalement (a fortiori, soit un effet) de *e*. Adopter une interprétation non-rétrograde des contrefactuels – en comprenant la relation de similitude entre mondes possibles de la manière suggérée par Lewis – permet donc de résoudre le problème des effets.

Qu'en est-il du problème des épiphénomènes ? Supposons que *c* advient à l'instant t_0, *e* à l'instant t_1 et *f* à l'instant t_2. Lewis soutient qu'un monde dans lequel le miracle requis advient à l'instant t_1-ε, très peu de temps avant t_1 mais après t_0, est plus semblable à @ que ne l'est un monde dans lequel ce miracle advient à l'instant t_0-ε, très peu de temps avant t_0. Si cette affirmation est vraie, alors tous les mondes-$\neg e$ les plus proches de @ sont des mondes-$c\&f$, ce qui

implique la fausseté du contrefactuel $\neg e > \neg f$. Et la fausseté de ce contrefactuel implique à son tour que *f* ne dépend pas causalement (a fortiori, n'est pas un effet) de *e* selon la définition de Lewis.

Le fait que la définition de la causalité offerte par Lewis permette de résoudre le problème des effets ainsi que le problème des épiphénomènes supporte la thèse selon laquelle cette définition est adéquate. Lewis a montré que cette définition est en accord avec nos « opinions naïves » dans les cas analogues à ceux que je viens de discuter. Il faut cependant noter que je n'ai jusqu'ici considéré que des cas dans lesquels la définition de Lewis semble rendre le verdict que deux évènements sont liés par une relation de causalité alors que nos intuitions nous disent que ce n'est pas le cas. Ceci n'est évidemment pas la seule façon pour une définition de la causalité de faillir à sa mission. La définition qu'offre Lewis sera également inadéquate s'il existe des cas dans lesquels elle implique qu'il n'existe pas de relation de causalité entre deux évènements alors que nos intuitions nous disent le contraire.

Existe-t-il de tels cas problématiques pour la définition de Lewis ? Oui, les cas que Lewis appelle des cas de « causalité redondante », c'est-à-dire des cas dans lesquels « deux causes potentielles distinctes d'un effet sont présentes ; chacune des deux causes aurait suffi par elle-même pour que l'effet advienne ; et l'effet ne dépend donc d'aucune de ces deux causes » (2000, 182). Il existe plusieurs types de cas de causalité redondante. J'en présenterai ici trois.

Considérons le cas suivant : Suzy lance une pierre vers une bouteille et la brise. Billy, qui se tient près de Suzy, aurait lancé une autre pierre si Suzy n'avait pas lancé la sienne, et il aurait aussi, par son jet, brisé la bouteille. Bien que nos intuitions nous disent que le jet de Suzy est une cause du bris de la bouteille, le bris ne semble pas dépendre causalement du jet. Il est faux de dire que si Suzy n'avait pas jeté sa pierre, la bouteille ne se serait pas brisée. La raison en est simple : si Suzy n'avait pas lancé sa pierre, Billy aurait lancé la sienne, et il aurait ainsi brisé la bouteille. Ce cas est, dans le vocabulaire de Lewis (2000, 184), un cas de « préemption précoce ».

Il est relativement aisé de montrer que la définition de Lewis s'accorde avec nos intuitions dans les cas de préemption précoce. Il suffit pour cela de supposer qu'il existe un évènement *d* intermédiaire entre le jet de Suzy et le bris de la bouteille – par exemple l'évènement qui consiste en l'occupation par la pierre de Suzy d'une région de l'espace située entre l'endroit où Billy et Suzy se tiennent et la bouteille – qui soit tel que le bris de la bouteille dépende causalement de *d* et que *d* dépende causalement du jet de Suzy. S'il est vrai que

(i) si Suzy n'avait pas lancé sa pierre, *d* ne serait pas advenu et (ii) si *d* n'était pas advenu, la bouteille ne se serait pas brisée, alors il existe une chaine causale menant du jet de Suzy au bris de la bouteille. Et dans la mesure où la causalité est la relation ancestrale de la dépendance causale, le jet de Suzy est bien une cause du bris de la bouteille selon la définition de Lewis.

Les cas de préemption précoce sont néanmoins, d'après Lewis, des « cas faciles » de causalité redondante. Considérons une version légèrement modifiée du cas décrit dans le paragraphe précèdent, due à Ned Hall (2004, 235) : Suzy et Billy jettent tous deux une pierre vers une bouteille, mais la pierre de Suzy arrive en premier et brise la bouteille. La pierre de Billy arrive à peine un instant plus tard et traverse l'espace, à présent vide, que la bouteille occupait un instant plus tôt, avant de se briser. Si Suzy n'avait pas jeté sa pierre, celle jetée par Billy aurait brisé la bouteille. Ce cas est un cas de « préemption tardive » (Lewis 2000, 184). La solution que Lewis apporte au problème soulevé par les cas de préemption précoce n'est pas applicable aux cas de préemption tardive. La raison en est que, puisque Suzy et Billy jettent tous deux une pierre vers la bouteille, il n'existe aucun évènement *d* intermédiaire entre le jet de Suzy et le bris de la bouteille tel que le bris de la bouteille dépend causalement de cet évènement. Il n'existe pas, en conséquence, de chaîne causale menant du jet de Suzy au bris de la bouteille. Et la définition de Lewis implique donc que ce jet n'est pas une cause du bris de la bouteille, contrairement à ce que nos intuitions nous incitent à croire.

Comment Lewis propose-t-il donc de traiter des cas de préemption tardive ? Sa proposition originale (Lewis, 1986b, 206--207) est la suivante. considérons un monde possible dans lequel se déroule le même scenario que celui décrit plus haut, à un détail près : dans ce monde possible, appelons-le *w*, Billy et sa pierre sont absents. Il existe, dans ce monde *w*, une chaîne causale menant du jet de Suzy au bris de la bouteille. Mais la causalité, semble-t-il, est une relation intrinsèque entre évènements. Que Billy jette sa pierre ou qu'il ne la jette pas ne devrait pas, intuitivement, avoir d'influence sur la relation entre le jet de Suzy et le bris de la bouteille. Dans la mesure où la séquence d'évènements menant du jet de Suzy au bris de la bouteille dans le monde actuel est une « copie intrinsèque » (Lewis, 2000, 184) de la séquence d'évènements dans le monde *w*, elle doit donc, elle aussi, être une chaîne causale. Lewis défend donc la thèse selon laquelle les copies intrinsèques de chaînes causales sont elles aussi des chaînes causales. Ces copies sont des séquences d'évènements entretenant des relations de « quasi-dépendance »

(Lewis, 2000, 184). Le concept de quasi-dépendance permet à Lewis d'éviter que les cas de préemption tardive ne soient des contrexemples à sa définition de la causalité. Puisqu'il existe une chaine de quasi-dépendance menant du jet de Suzy au bris de la bouteille, le premier de ces évènements est bien une cause du second.

Existe-t-il des cas de causalité redondante autres que les cas de préemption précoce et de préemption tardive ? Oui, par exemple les cas de « préemption coupante » dus à Jonathan Schaffer (2000). Considérons le scenario suivant (attribué par Lewis à Bas van Fraassen) :

> Le sergent et le major crient leurs ordres aux soldats. Les soldats savent qu'en cas de conflit entre les deux, ils doivent obéir à l'officier le plus haut gradé, c'est-à-dire au major. Mais il n'y a pas ici de conflit. Le sergent et le major crient simultanément "Avancez !" ; les soldats entendent les deux ordres ; les soldats avancent. Leur mouvement est causé de manière redondante : les soldats auraient avancé si le sergent avait crié "Avancez !" pendant que le major gardait le silence. Mais ils auraient également avancé si le major avait crié "Avancez !" pendant que le sergent gardait le silence. La redondance est ici asymétrique : dans la mesure où les soldats obéissent à l'officier le plus haut gradé, ils avancent parce que le major leur a donné l'ordre d'avancer, pas parce que le sergent leur a donné le même ordre. Le major préempte causalement le sergent. L'ordre du major *coupe* celui du sergent. (Lewis, 2000, 183, italiques originales).

L'intuition que nous sommes ici supposés avoir est que l'ordre du major est une cause du mouvement des soldats alors ce n'est pas le cas de l'ordre du sergent. Cependant, dans la mesure où le mouvement des soldats ne dépend causalement ni de l'ordre du major ni de l'ordre du sergent, aucun de ces deux ordres n'en est une cause d'après la définition de Lewis.

Les cas de préemption coupante faisant partie des cas de préemption tardive, il n'existe aucun évènement *d* intermédiaire entre le mouvement des soldats et l'ordre du major tel que le mouvement des soldats dépend causalement de cet évènement. Peut-on ici faire appel à la notion de quasi-dépendance ? Considérons un monde w_1 dans lequel existe une copie intrinsèque de la séquence d'évènements menant de l'ordre du major au mouvement des soldats mais dans lequel il n'y a pas de sergent. Dans ce monde, le mouvement des soldats dépend causalement de l'ordre du major et cet ordre est donc, dans le monde w_1, une cause du mouvement des soldats.

Ceci signifie que le mouvement des soldats dans le monde actuel quasi-dépend de l'ordre du major. Ce qui implique que l'ordre du major est bien, dans le monde actuel, une cause du mouvement des soldats. Jusqu'ici tout va bien.

Mais considérons à présent un monde w_2 dans lequel existe une copie intrinsèque de la séquence d'évènements menant de l'ordre du sergent au mouvement des soldats mais dans lequel il n'y a pas de major. Dans ce monde w_2, le mouvement des soldats dépend causalement de l'ordre du sergent et cet ordre est donc, dans w_2, une cause des mouvements des soldats. Ceci signifie que le mouvement des soldats dans le monde actuel quasi-dépend de l'ordre du sergent. Ce qui implique que l'ordre du sergent est, dans le monde actuel, une cause du mouvement des soldats, contrairement à ce que nos intuitions semblent indiquer. La notion de quasi-dépendance ne permet donc pas de résoudre le problème soulevé par les cas de préemption coupante.

Comment, alors, traiter des cas de préemption coupante ? Plutôt que de modifier encore sa définition originale, Lewis (2000) propose une nouvelle définition. Cette nouvelle définition est proche de sa définition originale mais promet de résoudre tous les problèmes rencontrés par celle-ci. Je me tourne à présent vers cette nouvelle définition.

III. La définition de 2000

Considérons à nouveau le cas de préemption tardive présenté plus haut, un cas dans lequel, intuitivement, le jet de Suzy est une cause du bris de la bouteille alors que ce n'est pas le cas du jet de Billy. Il semble évident que Suzy aurait pu jeter sa pierre d'une façon qui n'est pas la façon dont elle l'a actuellement jetée. Elle aurait pu la jeter un peu avant ou un peu après, avec plus ou moins de force, etc. Suivant Lewis (2000, 188), appelons chacune de ces versions légèrement différentes du jet de Suzy (y compris sa version actuelle) une « altération » du jet de Suzy. Il semble que si l'une des altérations non-actuelles du jet de Suzy était advenue, alors une altération non-actuelle du bris de la bouteille serait également advenue. Si Suzy avait jeté sa pierre un peu plus tôt, par exemple, la bouteille se serait aussi brisée un peu plus tôt.

Lewis propose d'utiliser la notion d'altération définie dans le paragraphe précèdent afin de formuler une nouvelle définition de la causalité. Commençons d'abord par définir la notion d'influence. Soit deux évènements actuels c et e. Selon Lewis (2000, 190), c influence e ssi il existe « un ensemble conséquent » c_1, c_2, …, c_n d'altérations de c (y compris son altération actuelle) et un

ensemble e_1, e_2, …, e_n d'altérations de *e* tels que le contrefactuel $c_i > e_i$ est vrai pour i = 1, …, *n*. Si *c* influence *e* alors, intuitivement, la manière (le lieu, le moment, etc.) dont *e* advient dépend causalement de la manière (le lieu, l'instant, etc.) dont *c* advient. Mais l'influence n'est pas encore la causalité. La raison en est que, comme le montre Lewis (2000, §VIII), la relation d'influence n'est pas transitive. La solution à cette difficulté est simplement de définir la causalité comme la relation ancestrale de la relation d'influence. Autrement dit, selon la nouvelle définition de Lewis, un évènement actuel *c* est une cause d'un autre évènement actuel *e* si il existe une chaîne d'influence menant du premier de ces évènements au second.

Comment cette nouvelle définition permet-elle à Lewis de traiter des cas de causalité redondante ? Commençons par les cas de préemption tardive. Que se passerait-il si le jet de Billy était altéré, c'est-à-dire si une altération non-actuelle de son jet advenait ? Le bris de la bouteille ne changerait pas, ou très peu. Quelque effet infime qu'ait le jet de Billy sur le bris de la bouteille – par exemple, par le truchement de la force gravitationnelle qu'exerce la pierre jetée par Billy sur la bouteille – cet effet ne changerait pas, ou très peu, si une altération non-actuelle du jet de Billy advenait. Mais qu'en serait-il si l'on altérait de façon semblable le jet de Suzy tout en laissant celui de Billy inchangé ? Il semble que, dans ce cas, le bris de la bouteille changerait dans des proportions bien plus importantes, puisque c'est la pierre jetée par Suzy qui, au bout du compte, entre en contact avec la bouteille. C'est en tout cas la thèse que soutient Lewis. Et c'est, selon lui, parce que le jet de Suzy a une influence sur le bris de la bouteille qui est bien plus grande que celle du jet de Billy que nous jugeons intuitivement que le jet de Suzy est une cause du bris de la bouteille alors que ce n'est pas le cas de celui de Billy[10].

Qu'en est-il des cas de préemption coupante qui posent tant de problèmes à la première définition de la causalité offerte par Lewis ? Lewis résout le problème posé par les cas de préemption coupante de la même manière. Pourquoi jugeons-nous intuitivement que l'ordre du major est une cause du mouvement des soldats alors que ce n'est pas le cas de l'ordre du sergent ? Parce qu'altérer l'ordre du major – en le changeant par exemple de « Avancez ! » à « Mettez-vous à l'abri ! » – sans altérer celui du sergent aurait, selon toute vraisemblance, un effet conséquent sur le mouvement des soldats, alors que l'inverse n'est pas le cas. Changer l'ordre du sergent sans changer celui du major n'aurait aucun effet sur le mouvement des soldats puisque, en

[10] Le même raisonnement s'applique aux cas de préemption précoce.

cas de conflit, les soldats suivront l'ordre du major. L'ordre du major a donc une influence importante sur le mouvement des soldats alors que l'ordre du sergent n'en a, semble-t-il, aucune. Et c'est pourquoi, soutient Lewis, nous jugeons intuitivement que l'ordre du major est une cause du mouvement des soldats alors que ce n'est pas le cas de l'ordre du sergent.

Les objections et contrexemples à la seconde définition de la causalité offerte par Lewis sont bien trop nombreux pour que je puisse ici les cataloguer. Je me limiterai ici à présenter un contrexemple putatif, proposé par Tomasz Bigaj[11]. Voici le scenario qui, selon Bigaj (2012, 8), constitue un contrexemple à la définition de la causalité en termes d'influence avancée par Lewis :

> Considérons une des situations les plus simples que l'on puisse imaginer : une ligne de chemin de fer se divisant en deux voies, et un aiguillage déterminant la direction de la circulation des trains. Supposons également qu'une des deux voies mène vers une impasse. Un train de passagers doit passer la fourche où se trouve l'aiguillage à 18h pour atteindre sa destination. Cependant, un malandrin change la position de l'aiguillage en tirant sur un levier à 17h, de sorte que le train devant passer la fourche à 18h se retrouve aiguillé sur la voie menant vers une impasse. Malheureusement, personne ne s'en rend compte, le train continue son chemin à toute vapeur sur la mauvaise voie et fini, au bout du compte, par s'écraser contre un mur.

Bigaj soutient, à raison me semble-t-il, que dans le cas que je viens de décrire, nous jugeons intuitivement que le basculement de l'aiguillage par le malandrin est une cause de l'accident du train. Mais le basculement de l'aiguillage n'influence pas, dans le sens que Lewis donne à ce terme, l'accident. Aucune altération possible du basculement de l'aiguillage n'aurait pour résultat une altération de l'accident du train, comme le soutient Bigaj (2012, 9) dans le passage suivant :

> Vous êtes libre d'imaginer toutes sortes d'altérations de la façon dont le malandrin tire le levier dans le monde actuel – à 17h05, à 16h55, avec la main gauche, avec son pied, de manière lente et délicate, de manière brusque, etc. Aucune de ces altérations n'aura pour résultat une différence dans la manière dont se déroule l'accident de train.

[11] Cf. Kvart (2001), Schaffer (2001), Strevens (2003) ou Stone (2009), entre autres, pour d'autres contrexemples.

Il est important de noter que, dans la mesure où une altération non-actuelle d'un évènement est, par définition, une version légèrement différente de cet évènement, le cas dans lequel le malandrin ne bascule pas l'aiguillage n'est pas une altération de l'évènement advenu dans le monde actuel. Autrement dit, ne pas basculer un aiguillage du tout n'est pas une façon de basculer un aiguillage. Si c'était le cas, alors le basculement de l'aiguillage par le malandrin influencerait l'accident et serait donc, selon la nouvelle définition de Lewis, une cause de cet accident[12].

Quel enseignement tirer du contrexemple proposé par Bigaj ? Que la nouvelle définition de la causalité proposée par Lewis n'est pas adéquate. Devrions-nous donc, puisqu'aucune des deux définitions proposées par Lewis n'est adéquate, abandonner l'espoir d'un jour arriver à une définition qui soit adéquate et réduise la causalité à la dépendance contrefactuelle ? Non, il n'est pas encore temps de paniquer. De fait, nombre de philosophes proposant aujourd'hui des définitions de la causalité ont la même ambition que Lewis, c'est-à-dire réduire la causalité à la dépendance contrefactuelle, et ce même si certains d'entre eux (Hitchcock, par exemple) ne défendent pas explicitement la thèse – défendue par Lewis – selon laquelle la dépendance contrefactuelle peut être analysée sans recours à des notions causales. C'est le cas, par exemple, de plusieurs des philosophes qui développent leurs définitions de la causalité en s'appuyant sur le formalisme des équations structurelles. Je me tourne à présent vers l'une de ces définitions, proposée par (Hitchcock, 2007).

IV. Équations structurelles, contrefactuels et causalité

Considérons à nouveau le contrexemple de Bigaj à la définition de la causalité en termes d'influence que propose Lewis. Il semble intuitivement vrai que si le malandrin n'avait pas basculé l'aiguillage avant 18h, alors le train n'aurait pas connu d'accident. Dans la mesure où le malandrin a, dans le monde actuel, basculé l'aiguillage avant 18h et où le train a bien connu un accident, il est aussi vrai que si le malandrin avait basculé l'aiguillage avant 18h, alors le train aurait connu un accident. C'est en tout cas le verdict que rend l'analyse des conditions de vérité des contrefactuels que propose Lewis (cf. Section 2).

[12] Cf. Bigaj (2012, §4) pour une objection à la proposition selon laquelle Lewis pourrait simplement adopter la thèse suivant : *c* est une cause de *e* ssi il existe une chaîne ou bien de dépendance causale ou bien d'influence menant de l'un à l'autre.

Soient deux variables binaires *M* et *C*. La valeur de *M* est 1 lorsque le malandrin bascule l'aiguillage avant 18h et 0 lorsque ce n'est pas le cas. La valeur de *C* est 1 lorsque le train connait un accident au bout de la voie menant vers un impasse et 0 lorsque ce n'est pas le cas[13]. Étant données ces deux variables, on peut représenter la relation entre le basculement de l'aiguillage par le malandrin et l'accident à l'aide de l'équation suivante :

$$C = M \qquad \text{(E)}$$

Le symbole « = » dans cette équation doit être compris comme représentant non seulement une égalité entre les valeurs de *C* et celles de *M* mais aussi une relation de dépendance contrefactuelle. Autrement dit, (E) n'exprime pas simplement le fait que les valeurs de *C* et celles de *M* covarient. Dans la mesure où les relations de dépendance contrefactuelle sont, d'après Lewis, asymétriques, ce doit également être le cas du symbole « = » qui apparait dans (E). Je suivrais ici les conventions établies en stipulant que l'équation (E) doit être comprise comme exprimant la proposition que les valeurs de *C* dépendent contrefactuellement des valeurs de *M*. Étant donnée cette interprétation, l'équation (E) n'est rien d'autre qu'une façon brève et compacte d'exprimer les deux contrefactuels suivant : $M = 1 > C = 1$ et $M = 0 > C = 0$. Et ces deux contrefactuels sont, bien évidemment, ceux dont j'ai affirmé la vérité dans le paragraphe ouvrant la présente section.

Pourquoi les équations telles que (E) sont-elles souvent appelées « équations structurelles » ? De nombreux philosophes s'intéressant à la causalité pensent, tout comme Lewis, que dépendance causale et dépendance contrefactuelle sont une seule et même relation (il faut ici se souvenir que Lewis défend une interprétation non-rétrograde des contrefactuels). Pour ces philosophes, les équations telles que (E) expriment donc aussi des relations de dépendance causale, ou ce que Hitchcock (2007, 504) appelle la « structure causale singulière ». Mais alors, pourquoi qualifier ces équations de « structurelles » plutôt que de « causales » ? Parce que l'utilisation d'équations pour représenter les relations causales trouve son origine dans les disciplines de la statistique et de l'économétrie à une époque où les notions causales avaient mauvaise réputation (cf. Frisch et Waugh, 1933, 390, par exemple)[14]. Bien que nous vivions à une époque différente et que les notions causales aient

[13] Je me limiterai ici au cas dans lequel les variables sont binaires et représentent le fait qu'un évènement advienne ou n'advienne pas.

[14] Cf. Drouet (2012, Chapitre 2) pour une présentation de la manière dont équations et modèles structurels sont utilisés pour tirer des inférences causales à partir de données empiriques.

aujourd'hui une bien meilleure réputation (Hoover, 2004), l'adjectif « structurel » nous est resté, comme un vestige de ce passé pas si lointain. Et c'est donc pourquoi les équations telles que (E) sont aujourd'hui encore appelées « équations structurelles ».

Il est important de noter que l'utilisation d'équations afin de représenter les relations causales n'est pas l'apanage des philosophes identifiant dépendance causale et dépendance contrefactuelle. On peut très bien se servir d'une équation comme (E) pour exprimer la proposition que les valeurs de *C* dépendent causalement des valeurs de *M* sans pour autant identifier dépendance causale et dépendance contrefactuelle. Il faudra cependant pour cela interpréter le symbole « = » apparaissant dans (E) d'une autre manière que celle postulée plus haut. Michael Baumgartner (2013) et Nancy Cartwright (2015), par exemple, utilisent tous deux le cadre des équations structurelles afin de formuler des définitions de la causalité, et ce bien que ni l'un ni l'autre ne souscrivent à la thèse identifiant dépendance causale et dépendance contrefactuelle.

Comment les équations structurelles peuvent-elles servir de fondation sur laquelle construire une définition de la causalité ? Il existe de nombreuses réponses à cette question. Comme je l'ai indiqué plus haut, je me limiterai ici à une seule de ces réponses, proposée par Hitchcock (20007). Et j'utiliserai une version légèrement modifiée du cas décrit par Bigaj afin de l'illustrer. Dans cette variante du cas de Bigaj, un malandrin suppléant se tient prêt à basculer l'aiguillage avant 18h au cas où le premier malandrin ne le ferait pas. Comme vous l'aurez peut-être deviné, cette modification fait du cas de Bigaj un cas de préemption précoce.

Commençons par stipuler qu'un *modèle causal* est un couple <**V**, **E**> dans lequel **V** est un ensemble de variables et **E** un ensemble d'équations structurelles reliant ces variables (Hitchcock, 2007, 499). Dans la variante du cas de Bigaj décrite dans le paragraphe précédent, **V** = {*U*, *M*, *B*, *C*} et **E** = {(1), (2), (3)}. Les variables *M* et *C* ont déjà été définies plus haut. *B* est une variable binaire dont la valeur est 1 lorsque le malandrin suppléant bascule l'aiguillage avant 18h et 0 lorsque ce n'est pas le cas. Et *U* est une variable binaire résumant les causes du basculement de l'aiguillage avant 18h par le premier malandrin. La valeur de *U* est 1 lorsque ces causes sont présentes et 0 lorsqu'elles sont absentes. Les équations (1), (2) et (3) ci-dessous expriment les relations de dépendance causale existant entre ces variables :

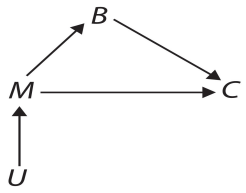
$$M = U \qquad (1)$$

$$B = 1 - M \quad (2)$$

$$C = \text{MAX}\{B, M\} \quad (3)$$

La première équation exprime la proposition que le basculement de l'aiguillage avant 18h par le malandrin original dépend, à la fois contrefactuellement et causalement, d'un certain nombre de causes qui ne sont pas ici explicitement mentionnée. La seconde équation exprime la proposition que le basculement de l'aiguillage avant 18h par le malandrin suppléant dépend du comportement du premier malandrin à l'égard de l'aiguillage. Et la troisième équation exprime la proposition que l'accident du train dépend à la fois du comportement du premier malandrin et de celui du malandrin suppléant à l'égard de l'aiguillage.

Le modèle causal que je viens de définir n'est rien d'autre qu'une représentation formelle de la variante du cas que Bigaj décrit de manière informelle. Il est possible de transformer cette représentation formelle en une représentation graphique.



(Figure 1)

La procédure pour transformer un modèle causal en un « graphe causal » est simple : dessinez une flèche de X vers Y dans un graphe causal ssi il existe une équation structurelle de la forme $Y = f(X\ldots)$ dans le modèle causal correspondant. Les flèches dans la Figure 1 représentent donc des relations de dépendance causale et aussi, pour Lewis et nombre d'autres philosophes, des relations de dépendance contrefactuelle. Il est à noter que les graphes causaux sont moins informatifs que les modèles causaux qu'ils représentent. Le graphe causal dans la Figure 1, par exemple, indique que les valeurs de B dépendent des valeurs de M mais n'indique pas, contrairement à l'équation (2), la forme fonctionnelle de cette dépendance. Deux points de terminologie avant de passer à l'étape à suivante. Premièrement, s'il existe une flèche unique de X vers Y dans un graphe causal, alors X est un « parent » de Y et Y un « enfant » de X. Deuxièmement, une série de variables $X_1, \ldots, X_n$ dans un graphe causal telle que chaque X_i(pour $i = 1, \ldots, X_{n-1}$) est un parent de X_{i+1} est dite être un « chemin dirigé » de X_1 vers X_n.

L'étape suivante sur le chemin vers la définition de la causalité que propose Hitchcock est de déterminer les conditions dans lesquelles un

contrefactuel est *vrai dans un modèle causal*. Soit <**V**, **E**> un modèle causal dans lequel **V** = {X_1, X_2, ..., X_n}. Pour chaque variable X_i dans **V**, **E** contient une équation (E_{Xi}) de la forme $X_i = f_i(X_j\ldots)$ reliant cette variables aux autres éléments de **V**. Selon Hitchcock (2007, 501), le contrefactuel $X_i = x_i > X_j = x_j$ est vrai dans <**V**, **E**> ssi la valeur de $X_j$est x_j dans le modèle causal <**V**, **E'**> que l'on obtient en remplaçant (E_{Xi}) par l'équation $X_i = x_i$[15].

Bien que l'analyse donnée dans le paragraphe précédent – une analyse que Hitchcock emprunte à Judea Pearl (2000, §7.1) – puisse sembler compliquée, elle est similaire dans l'esprit à celle proposée par Lewis et présentée plus haut (Section II). Dans les deux cas, lorsque l'on veut déterminer ce qui se passerait si un évènement non-actuel advenait (ou si un évènement actuel n'advenait pas), il faut considérer une situation dans laquelle l'évènement en question advient (ou n'advient pas) alors que rien d'autre ne change (ou change le moins possible). Le modèle causal modifié <**V**, **E'**> est en ceci similaire à un monde possible dans lequel l'évènement $X_i = x_i$ advient suite à un miracle, c'est-à-dire, dans le cas présent, suite à une violation de l'équation (E_{Xi})[16].

Il convient à présent d'illustrer l'analyse que donne Hitchcock des contrefactuels au moyen de la variante du cas de Bigaj. Quelle est la valeur de vérité – dans le modèle causal <**V**, **E**> défini plus haut – de la proposition que si le premier malandrin n'avait pas basculé l'aiguillage avant 18h, alors le train n'aurait pas connu d'accident, formellement $M = 0 > C = 0$? Commençons par remplacer l'équation qui relie *M* à son unique parent dans **V**, c'est-à-dire (1), par l'équation $M = 0$. Quelle est la valeur de *B* dans le modèle causal ainsi obtenu ? Selon l'équation (2), la valeur de *B* est 1 lorsque celle de *M* est 0. On peut maintenant déterminer la valeur de *C*. Et, selon l'équation (3), la valeur de *C* est 1 lorsque $B = 1$ et $M = 0$. Le contrefactuel $M = 0 > C = 0$ est donc faux dans le modèle causal <**V**, **E**>. Autrement dit, il est, dans ce modèle, faux que si le premier malandrin n'avait pas basculé l'aiguillage avant 18h, alors le train n'aurait pas connu d'accident. La raison en est que si le premier malandrin n'avait pas basculé l'aiguillage avant 18h, le malandrin suppléant l'aurait fait.

---

[15] Il est à noter que $X_i = x_i$ n'est pas, à proprement parler, une équation structurelle et que <**V**, **E'**> n'est donc pas non plus un modèle causal selon la définition adoptée plus haut. $X_i = x_i$ n'exprime pas une relation de dépendance contrefactuelle mais exprime plutôt le postulat selon lequel la valeur de $X_i$est x_i.

[16] Cf. (Briggs, 2012) pour une comparaison concise et précise entre l'analyse des contrefactuels que propose Lewis et celle que propose Pearl.

L'analyse que donne Hitchcock des contrefactuels permet de déterminer les valeurs de vérité de contrefactuels ayant des antécédents complexes – dans les limites énoncées par Briggs (2012) – de manière simple et mécanique[17]. Cette analyse rend-elle celle de Lewis obsolète et redondante ? Non, car nous avons encore besoin de l'analyse de Lewis, ou d'un substitut adéquat, pour ce qui est des contrefactuels exprimés par les équations structurelles du modèle causal que l'on considère. Revenons un instant au cas original de Bigaj, celui dans lequel il n'y a pas de malandrin suppléant. Soit le modèle causal pour ce cas le couple <**V**, **E**> dans lequel **V** = {*C*, *M*} et **E** = {(E)}. Peut-on utiliser la procédure correspondant à l'analyse des contrefactuels que propose Hitchcock pour répondre à une question concernant la valeur de vérité de la proposition que si le malandrin n'avait pas basculé l'aiguillage avant 18h, alors le train n'aurait pas connu d'accident, formellement $M = 0 > C = 0$? Non, en tout cas pas si l'on espère une réponse à cette question qui soit informative et non pas triviale. Supposer que l'équation (E) reliant *M* et *C* est une équation *structurelle* est aussi, par définition, supposer la vérité du contrefactuel $M = 0 > C = 0$[18].

L'analyse que donne Hitchcock des conditions de vérité des contrefactuels lui permet de définir la notion de *dépendance contrefactuelle dans un modèle causal* de manière simple. Soit *X* et *Y* deux variables binaires, *x* et *y* étant leurs valeurs actuelles respectives. On dira que *X* dépend contrefactuellement de *Y* dans un modèle causal <**V**, **E**> ssi il existe des valeurs non-actuelles *x*'et *y*' de *X* et *Y*, respectivement, qui sont telles que le contrefactuel $X = x' > Y = y'$ est vrai dans <**V**, **E**>.

L'étape suivante sur le chemin vers la définition que propose Hitchcock est de caractériser les notions de « valeur par défaut » et « valeur déviante » d'une variable. Si j'emploie ici le verbe « caractériser » plutôt que le verbe « définir » c'est parce qu'Hitchcock n'offre pas de définitions précises de ces notions. Selon Hitchcock (2007, 506), « la valeur par défaut d'une variable est la valeur que l'on s'attend à observer en l'absence d'information à propos d'interférences causales». Mais il est difficile de définir ce en quoi consiste une

[17] Pour que cette procédure réponde correctement aux questions concernant les valeurs de vérité des contrefactuels, il faut cependant que le modèle causal adopté soit correct, ou « approprié » (Hitchcock, 2007, 503), c'est-à-dire qu'il décrive correctement les relations de dépendance contrefactuelle existant « hors du modèle ».

[18] Ceci n'est évidemment le cas que si, comme je l'ai fait plus haut, on suppose que le symbole « = » dans les équations structurelles représente à la fois une relation de dépendance causale et une relation de dépendance contrefactuelle.

« interférence causale » et, a fortiori, de déterminer les valeurs que l'on s'attend à observer en l'absence de telles interférences.

Considérons la variante du cas de Bigaj. Il semble que la valeur par défaut de C, dénotée $Def(C)$, doive être 0 et donc sa valeur déviante, dénotée $Dev(C)$, 1 : En l'absence d'information au sujet des deux malandrins, on s'attendrait sans aucun doute à ce que le train poursuive sa route sur la voie principale et arrive à bon port sans connaitre d'accident. C'est en tout cas ce que soutient Hitchcock (2007, 506) dans un cas analogue à la variante du cas de Bigaj. Il est important de noter que la valeur par défaut d'une variable n'est pas toujours sa valeur actuelle. La valeur par défaut de C, par exemple, est 0 alors que sa valeur actuelle est 1. Qu'en est-il des variables M et B ? Quelles sont leurs valeurs par défaut et leurs valeurs déviantes ? Cette question ne semble pas admettre de réponse qui soit indiscutable. Le comportement « par défaut » du premier malandrin, par exemple, est-il de basculer l'aiguillage avant 18h ? Comme je l'expliquerai plus bas, cette question n'est pas anodine dans la mesure où les notions de valeur par défaut et de valeur déviante jouent un rôle crucial dans la définition de la causalité que propose Hitchcock.

Quelle est l'étape suivante sur le chemin vers la définition que propose Hitchcock ? Elle est de définir la notion de « réseau causal » (Hitchcock, 2007, 509). Soit $<\mathbf{V}, \mathbf{E}>$ un modèle causal tel que $X, Y \in \mathbf{V}$. Le réseau causal reliant X à Y dans ce modèle causal est l'ensemble $\mathbf{N} \subseteq \mathbf{V}$ dont les éléments sont X, Y ainsi que toutes les variables situées sur un chemin dirigé menant de X à Y dans le graphe causal correspondant à $<\mathbf{V}, \mathbf{E}>$. Dans la variante du cas de Bigaj, par exemple, le réseau causal reliant M à C est l'ensemble $\{M, B, C\}$. Un réseau causal $\mathbf{N}$ reliant X à Y dans un modèle causal $<\mathbf{V}, \mathbf{E}>$ est dit « autonome » (Hitchcock, 2007, 509) ssi, pour toute variable $Z \in \mathbf{N}$, si Z a des parents dans $\mathbf{N}$ alors la valeur de Z est sa valeur par défaut lorsque (i) chacun de ses parents dans $\mathbf{N}$ prend sa valeur par défaut et (ii) chacun de ses parents hors de $\mathbf{N}$ prend sa valeur actuelle.

Il est enfin temps de présenter la définition de la causalité qu'offre Hitchcock, maintenant que ces importants préliminaires conceptuels et terminologiques sont derrière nous. Voici donc la définition que propose Hitchcock (2007, 511) :

> *TC* : Soit $<\mathbf{V}, \mathbf{E}>$ un modèle causal tel que $X, Y \in \mathbf{V}$, $X = x$ et $Y = y$. Si le réseau causal reliant X à Y dans $<\mathbf{V}, \mathbf{E}>$ est autonome, alors X

> = x est une cause de $Y = y$ dans <**V**, **E**> ssi Y dépend contrefactuellement de X dans <**V**, **E**>.

Comme le lecteur l'aura sans doute remarqué, il y a deux différences majeures entre *TC* et les définitions offertes par Lewis. Premièrement, *TC* définit la causalité comme étant une relation entre les valeurs de certaines variables et non pas comme étant une relation entre évènements actuels. Deuxièmement, *TC* rend la causalité relative aux modèles causaux : Il est possible que $X = x$ soit une cause de $Y = y$ dans un modèle causal mais pas dans un autre. Peut-on rapprocher *TC* des définitions offertes par Lewis ? Oui, de la manière décrite par Hitchcock dans le passage suivant :

> Un évènement *c* est une cause d'un autre évènement *e* ssi : (i) $X = x$ est, selon *TC*, une cause de $Y = y$ dans un modèle causal <**V**, **E**> ; (ii) $X = x$ représente l'occurrence de *c* et $Y = y$ celle de *e* ; et (iii) <**V**, **E**> est un modèle causal *approprié* de la situation dans laquelle *c* et *e* adviennent. (Hitchcock, 2007, 503, italiques originales).

Dans quelles conditions un modèle causal est-il approprié ? Au moins lorsqu'il n'implique pas de contrefactuels qui soient faux. Il faut cependant noter qu'Hitchcock ne dit pas grand-chose, du moins dans Hitchcock (2007), sur ce sujet pourtant important et concède que le caractère approprié ou non d'un modèle causal dépendra souvent de facteurs pragmatiques.

Comment la définition de Hitchcock s'applique-t-elle à la variante du cas de Bigaj ? Il faut se souvenir, tout d'abord, que le modèle causal <**V**, **E**> correspondant à ce cas est tel que **V** = {U, M, B, C} et **E** = {(1), (2), (3)}, et aussi que les valeurs actuelles des variables dans **V** sont les suivantes : $U = 1$, $M = 1$, $B = 0$ et $C = 1$. La définition de la causalité que propose Hitchcock implique-t-elle que le basculement de l'aiguillage avant 18h par le premier malandrin est une cause de l'accident de train si l'on suppose que le modèle causal <**V**, **E**> est approprié ?

Afin de répondre à cette question, il faut tout d'abord déterminer les valeurs par défaut et les valeurs déviantes des variables dans **V**. J'ai dit plus haut qu'il semble que la valeur par défaut de C doive être 0 et sa valeur déviante 1. Qu'en est-il de U, M et B ? Je postulerai ici que, pour toutes les

variables Z dans **V**, Def(Z) = 0 et Dev(Z) = 1[19]. Quel est le réseau causal reliant M à C dans <**V**, **E**> ? Il s'agit simplement de l'ensemble **N** = {M, B, C}.

Ce réseau causal est-il autonome ? Commençons par M. Dans la mesure où M n'a pas de parent dans **N**, on peut ici l'ignorer. Qu'en est-il de B ? B a un parent dans **N**, en l'occurrence M. Et selon l'équation structurelle (2), B prend une valeur déviante, en l'occurrence 1, lorsque M prend sa valeur par défaut, c'est-à-dire 0. Ceci implique que le réseau causal **N** reliant M à C n'est pas autonome[20]. Et puisque ce réseau n'est pas autonome, *TC* ne s'y applique pas et ne peut donc pas répondre à la question de savoir si le basculement de l'aiguillage avant 18h par le premier malandrin est une cause de l'accident de train. Cette conclusion est pour le moins déconcertante. Bien que la définition de la causalité qu'offre Hitchcock soit relativement complexe, elle échoue, comme le reconnait lui-même Hitchcock (2007, 521), dans un cas apparemment simple de préemption précoce, un cas qui ne pose aucun problème à la première définition de Lewis, celle datant de 1973.

Ce résultat aurait-il été différent si d'autres valeurs par défaut avaient été choisies pour les variables qui composent **V ?** Considérons le cas dans lequel les valeurs par défaut et les valeurs déviantes de B et C sont inversées, c'est-à-dire le cas dans lequel Def(B) = Def(C) = 1 et Dev(B) = Dev(C) = 0, la valeur par défaut de M restant inchangée. Le réseau causal **N** reliant M à C est-il, dans ce cas, autonome ? Commençons par ignorer M puisque, comme je l'ai indiqué plus haut, cette variable n'a pas de parent dans **N**. Passons donc à B. Selon l'équation (2), la valeur de B est 1 lorsque la valeur de M est 0. B prend donc bien sa valeur par défaut lorsque M, son unique parent dans **N**, prend également sa valeur par défaut. Qu'en est-il de C ? L'unique parent de C dans **N** est B[21]. Selon l'équation (3), la valeur de C est 1 lorsque celle de B est 1. Dans la mesure où *Def*(C) = *Def*(B) = 1, C prend sa valeur par défaut lorsque B prend aussi sa valeur par défaut. Ce qui signifie que le réseau causal reliant M à C est bien autonome et donc que *TC* s'y applique.

Quel est alors, étant donnée cette nouvelle distribution de valeurs par défaut, le verdict rendu par *TC* ? Pour répondre à cette question, il faut déterminer si C dépend contrefactuellement de M dans le modèle causal <**V**, **E**>. La valeur actuelle de M étant la même que celle de C, c'est-à-dire 1, C

[19] Ce postulat est en accord avec celui que fait Hitchcock (2007, 506).
[20] Étant donnée la relation inverse entre valeurs de M et valeurs de B, la même conclusion suivra du postulat selon lequel *Def*(M) = *Def*(B) = 1.
[21] M n'est pas un parent de C puisqu'il y a deux flèches, et non une seule, entre ces deux variables.

dépend contrefactuellement de *M* dans <**V**, **E**> ssi le contrefactuel *M* = 0 >*C* = 0 est vrai dans <**V**, **E**>. Et pour que ce soit le cas, la valeur de *C* doit être 0 dans le modèle causal <**V**, **E'**> obtenu en substituant à l'équation (1) l'équation *M* = 0. Selon l'équation structurelle (2), la valeur de *B* est 1 lorsque celle de *M* est 0. Et selon l'équation structurelle (3), la valeur de *C* est 1 lorsque celle de *B* est 1. *C* ne dépend donc pas contrefactuellement de *M* dans le modèle causal <**V**, **E**>. Ceci implique, selon *TC*, que le basculement de l'aiguillage avant 18h par le premier malandrin n'est pas une cause de l'accident de train.

Ce résultat dépend bien sûr de la distribution de valeurs par défaut adoptée plus haut. Peut-être était-il illégitime de supposer que Def(*M*) = 0 et Def(*B*) = Def(*C*) = 1. Il semble en effet difficile de réconcilier le postulat selon lequel la valeur par défaut de *C* est 1 avec ce que soutient Hitchcock, c'est-à-dire que « la valeur par défaut d'une variable est la valeur que l'on s'attend à observer en l'absence d'information à propos d'interférences causales» (2007, 506). Mais qu'en serait-il si nous vivions dans un monde différent, un monde dans lequel les accidents de train sont extrêmement fréquents, et ce malgré les efforts de la SNCF locale. Il semble qu'il soit légitime, dans un tel monde, de supposer que la valeur par défaut de *C* est 1 : Dans ce monde, nous nous attendrions à ce que le train connaisse un accident même en l'absence d'information au sujet des deux malandrins. Mais quelles seraient nos intuitions quant à la cause de l'accident de train dans la variante du cas de Bigaj si nous vivions dans un tel monde ? Il semble que, même si l'on suppose que les accidents de train sont extrêmement fréquents, le basculement de l'aiguillage par le premier des deux malandrins reste, intuitivement, une cause de *cet* accident. C'est lui, après tout, qui manipule le levier aiguillant le train sur la voie menant à une impasse.

Qu'en est-il du postulat selon lequel la valeur par défaut de *B* est 1 ? Comment le défendre ici ? Dans la mesure où j'ai supposé que le malandrin suppléant bascule l'aiguillage avant 18h ssi le premier malandrin ne le fait pas, il semble que la distribution de valeurs par défaut pour les variables *M* et *B* doive satisfaire une condition de cohérence. Si le comportement par défaut du premier malandrin est de ne pas basculer l'aiguillage avant 18h, alors il semble que celui du malandrin suppléant doive être de le faire. Si le comportement par défaut d'un enseignant est d'être absent, alors celui de son suppléant – c'est-à-dire de l'enseignant qui est présent ssi le premier enseignant ne l'est pas – doit être, il semble, d'être présent.

Quelles conclusions tirer de la discussion menée plus haut ? Premièrement, Hitchcock soutient que les cas de préemption sont des cas dans lesquels il n'y a « pas de dépendance contrefactuelle dans un réseau non-autonome » (2007, 529). Cette affirmation est erronée. Comme le montre la variante du cas de Bigaj, il est possible pour un cas de préemption – en l'occurrence, un cas de préemption précoce – de ne pas appartenir à la catégorie décrite par Hitchcock. Tout dépend de la distribution de valeurs par défaut que l'on adopte. Étant donnée la seconde des deux distributions de valeurs par défaut considérées plus haut, la variante du cas de Bigaj est un cas de préemption précoce dans lequel il n'y a pas de dépendance contrefactuelle dans un réseau causal qui est bien autonome. Deuxièmement, Hitchcock soutient que, dans les cas dans lesquels il n'y pas de dépendance contrefactuelle dans un réseau causal autonome, « nous tendons fortement à penser que *c* n'est pas une cause de *e* » (2007, 529). Si la seconde des deux distributions de valeurs par défaut considérées plus haut est légitime, alors cette affirmation est également erronée. En d'autres termes, s'il est légitime de supposer que $Def(M) = 0$ et $Def(B) = Def(C) = 1$, alors le verdict rendu par *TC* contredit l'intuition selon laquelle le premier malandrin est bien une cause de l'accident de train, une intuition qui reste inchangée lorsque l'on se place dans un monde dans lequel les accidents de train sont extrêmement fréquents. Autrement dit, si l'argument que j'ai développé plus haut est sain, la définition de la causalité offerte par Hitchcock n'est pas adéquate. Troisièmement, et de manière plus générale, bien que les notions de valeur par défaut et de valeur déviante jouent un rôle crucial dans la définition de Hitchcock – ainsi que dans les définitions défendues, par exemple, par Peter Menzies (2004), Ned Hall (2007) ou par Hitchcock lui-même dans un article récemment co-écrit avec Joseph Halpern (2014) – il n'existe pas (i) de définitions strictes de ces notions, ni (ii) de procédure pour décider, dans chaque cas, quelle distribution de ces valeurs adopter[22].

La définition de la causalité offerte par Hitchcock (2007) est bien plus complexe que celles offertes par Lewis[23]. Cette définition rencontre pourtant des difficultés lorsque l'on s'intéresse à des cas apparemment simples de préemption précoce, puisque les verdicts qu'elle rend ne s'accordent avec nos jugements causaux que si la distribution de valeurs par défaut est soigneusement choisie. Cet état de fait doit-il inviter au pessimisme quant à la

[22] Même l'article récent de Halpern et Hitchcock (2014) donne au lecteur peu d'informations sur ce sujet épineux.

[23] C'est a fortiori le cas de la définition récemment proposée par Halpern et Hitchcock (2014).

possibilité d'une réduction de la causalité à la dépendance contrefactuelle s'appuyant sur le cadre des équations structurelles ? On peut penser que, dans la mesure où il existe d'autres tentatives que celle de Hitchcock, un tel pessimisme ne se justifie pas. Il existe cependant de bonnes raisons d'être pessimiste quant à la possibilité d'une définition de la causalité – quelle qu'elle soit – qui soit adéquate. Et c'est vers ces raisons que je me tourne à présent.

V. Analyses conceptuelles de la causalité : un regard pessimiste

Le projet de Lewis tel que je l'ai décrit plus haut est celui d'une analyse conceptuelle de la causalité qui serve d'appui à une réduction de la causalité à la dépendance contrefactuelle. Le projet de Hitchcock est distinct de celui de Lewis. On peut le décrire comme un projet de modélisation formelle du mécanisme générant nos jugements causaux. Cela ne signifie pas que la définition offerte par Hitchcock ou, plus généralement, le formalisme des équations structurelles, ne puisse, en principe, être utilisé afin d'arriver à la définition réductive recherchée par Lewis. C'est par exemple ce que propose de faire Ned Hall (2007a ; 2007b)[24]. Si la relation représentée par le symbole « = » dans les équations structurelles est la dépendance contrefactuelle telle que Lewis la comprend, alors toute définition adéquate de la causalité formulée au moyen du formalisme des équations structurelles accomplira, de fait, la réduction souhaitée par Lewis (si tant est qu'aucune notion causale ne soit introduite dans la construction de cette définition).

Malgré le fait que les projets de Lewis et Hitchcock soient distincts, ces deux philosophes adoptent néanmoins des méthodes très similaires. De fait, tous les articles considérés plus hauts, de Lewis (1973a) à Halpern et Hitchcock (2014), ont la même structure : dans un premier temps, l'auteur de l'article construit une définition de la causalité ; et dans un deuxième temps, cet auteur examine un ensemble de « cas d'école », par exemple des cas de préemption, pour montrer que, dans ces cas, les verdicts rendus par sa définition s'accordent avec nos jugements causaux.

Existe-t-il de bonnes raisons de penser que suivre la méthode des cas d'école nous permettra un jour d'arriver à une définition de la causalité qui soit adéquate ? Considérer l'histoire des définitions de la causalité n'incite guère à l'optimisme. Les philosophes ont tendance à être très ingénieux. Lorsqu'une définition d'un concept – la causalité, par exemple – est proposée, on peut

[24] Cf. Hitchcock (2009) pour une série d'objections à Hall (2007a).

parier sans risquer de se ruiner que des contrexemples suivront sans délai. Mais le pessimisme quant à la possibilité d'une définition de la causalité qui soit adéquate ne se justifie pas que par une induction s'appuyant sur les échecs des définitions offertes jusqu'à présent. Clark Glymour et un certain nombre de ses collègues (Glymour et al., 2010) ont récemment proposé d'autres arguments supportant un tel pessimisme et décriant ce qu'ils appellent la « méthode socratique », une méthode qui prévaut dans les travaux philosophiques sur la causalité. Bien qu'un examen détaillé de ces arguments ne soit pas ici possible, je présenterai néanmoins brièvement deux d'entre eux.

Telle que Glymour et al. la décrive, la méthode adoptée par Lewis et Hitchcock, parmi tant d'autres, est une méthode « par induction sur la base de nos intuitions dans un nombre infinitésimal d'exemples et de contrexemples possibles » (2010, 169). Les articles proposant une définition de la causalité considèrent rarement plus d'une douzaine de cas d'école avant de conclure, au moins de manière implicite, que leur définition est adéquate. La première thèse défendue par Glymour et al. (2010, §2) est que, dans la mesure où le nombre de cas possibles est très important, et ce même lorsque l'on considère un nombre restreint de causes, le fait qu'une définition de la causalité s'accorde avec nos jugements causaux dans une douzaine de cas est loin d'être suffisant pour justifier la conclusion selon laquelle la définition en question est adéquate.

Le deuxième argument avancé par Glymour et al. que j'examinerai ici met en question à la fois la fiabilité et le caractère représentatif des intuitions sur lesquelles s'appuient Lewis, Hitchcock et la plupart des autres philosophes ayant proposé des définitions de la causalité. Comme le disent Glymour et ses collègues,

> [t]outes les applications de la méthode socratique à la causalité dont nous ayons connaissance s'appuient sur les jugements causaux d'un groupe restreint de philosophes, même pour les cas atypiques. La supposition selon laquelle les jugements des philosophes à propos de ces cas atypiques font, ou devrait faire, autorité est à la fois rassurante et injustifiée. (2010, 186).

Si les définitions offertes par Lewis ou Hitchcock sont censées définir le concept de causalité possédé par les membres d'une population donnée (en l'occurrence, la population humaine dans son ensemble), alors pourquoi penser qu'il soit légitime de recourir aux intuitions de quelques philosophes, plutôt qu'aux intuitions typiques de cette population en général, pour évaluer ces

définitions[25] ? L'idée exprimée par cette question est à l'origine du mouvement de philosophie expérimentale qui s'est développé ces dernières années (Cf. Alexander, 2012). Comme le soulignent Glymour et ses collègues, il existe peu de travaux – que ce soit en psychologie ou en philosophie expérimentale – portant sur les cas d'école (par exemple, les cas de préemption) couramment utilisés pour évaluer les définitions de la causalité. C'est a fortiori le cas pour les myriades de cas possibles qui n'ont pas encore été « découverts » par les philosophes s'intéressant à la causalité. Même si Hitchcock, par exemple, arrivait à formuler une définition de la causalité qui soit en accord avec *ses* jugements causaux, il ne suivrait donc pas immédiatement qu'il a par là réussi à définir le concept de causalité que possèdent les humains en général.

VI. Conclusion

Faut-il donc abandonner l'espoir d'arriver un jour à une définition du concept de causalité qui soit adéquate et donc, a fortiori, d'arriver à réduire la causalité à la dépendance contrefactuelle telle que l'entend Lewis ? Le premier des deux arguments avancés par Glymour et ses collègues appuie la thèse selon laquelle cet espoir est vain. Et, quoique l'on pense de ce premier argument, le deuxième de leurs arguments implique en tous les cas que la méthode socratique jusqu'ici privilégiée devrait être abandonnée[26]. Il est important de noter que les arguments développés dans (Glymour et al., 2010) n'impliquent ni que la dépendance causale ne soit pas identique à la dépendance contrefactuelle telle que l'entend Lewis ni que ce soit une erreur de penser que causalité et dépendance contrefactuelle entretiennent des liens privilégiés. Mais ces arguments impliquent bien, si tant est qu'on les prenne au sérieux, que bon nombre de philosophes s'intéressant à la causalité sont, à la manière du train de Bigaj, engagés sur une voie menant vers une impasse[27].

---

[25] Cette critique s'applique aussi bien aux intuitions invoquées pour défendre les définitions de la causalité qu'aux intuitions invoquées pour les critiquer. Elle s'applique donc à l'intuition que j'ai moi-même invoquée plus haut (Section 4) dans ma critique de la définition offerte par Hitchcock (2007).

[26] Cf. Hall et Paul (2013) pour une méthode supposée permettre une réduction de la causalité sans devoir en passer par une analyse conceptuelle.

[27] Je remercie Nancy Cartwright ainsi qu'un relecteur anonyme pour *Klēsis* pour leurs commentaires sur une version antérieure de cet article.

Bibliographie

J. Alexander, *Experimental Philosophy : An Introduction*, Cambridge, Polity Press, 2012.

M. Baumgartner, « A Regularity Theoretic Approach to Actual Causation », in *Erkenntnis*, 78/1, 2013, pp. 85-109.

T. Bigaj, « Causation Without Influence », in *Erkenntnis*, 76/1, 2012, pp. 1-22.

R. Briggs, « Interventionist Counterfactuals », in *Philosophical Studies*, 2012, 160/1, pp. 139--166.

N. Cartwright, « Single Case Causes : What is Evidence and Why », manuscript, 2015.

S. Chauvier, « Le déversoir modal », », in *Klēsis*, 24, 2012, pp. 56--77.

J. Divers, *Possible Worlds*, London, Routledge, 2002.

P. Dowe, *Physical Causation*, Cambridge, Cambridge University Press, 2000.

R. Frisch et F. Waugh, « Partial Time Regressions as Compared with Individual Trends », in *Econometrica*, 1933, 1/4, pp. 387-401.

C. Glymour, D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, P. Spirtes, C. M. Teng, et J. Zhang, « Actual Causation : A Stone Soup Essay », in *Synthese*, 175/2, 2010, pp. 169-192.

N. Hall, « Two Concepts of Causation », in J. Collins, N. Hall et L.A. Paul (éds.), *Causation and Counterfactuals*, Cambridge, Mass., MIT Press, 2004.

N. Hall, « Structural Equations and Causation », in *Philosophical Studies*, 132/1, 2007a, pp. 109--136.

N. Hall, "Structural Equations and Causation (version longue) », manuscript, 2007b.

N. Hall et L.A. Paul, « Metaphysically Reductive Causation », in *Erkenntnis*, 78/1, 2013, pp. 9-41.

J. Halpern et C. Hitchcock, « Graded Causation and Defaults », in *British Journal for the Philosophy of Science*, 2014.

C. Hitchcock, « The Intransitivity of Causation Revealed in Equations and Graphs », in *Journal of Philosophy*, 98, 2001, pp. 273-299.

C. Hitchcock, « Prevention, Preemption, and the Principle of Sufficient Reason », in *Philosophical Review*, 2007, 116/4, pp. 495-532.

C. Hitchcock, « Structural Equations and Causation : Six Counterexamples », in *Philosophical Studies*, 144/3, 2009, pp. 391-401.

K. Hoover, « Lost Causes », in *Journal of the History of Economic Thought*, 26/2, 2004, pp. 149-164.

M. Kistler, « La causalité », in Anouk Barberousse, Denis Bonnay et Mikaël Cozic (dir.), *Précis de philosophie des sciences*, Paris, Vuibert 2011, p. 100-140

I. Kvart, « Lewis's "Causation as Influence" », in *Australasian Journal of Philosophy*, 79/3, 2001, pp. 409-421

D. Lewis, « Causation », in *Journal of Philosophy*, 70/17, 1973a, pp. 556--567.

D. Lewis, *Counterfactuals*, Oxford, Blackwell, 1973b

D. Lewis, « Counterfactual Dependence and Time's Arrow », in *Noûs*, 13/4, 1979, pp. 455--76.

D. Lewis. « New Work for a Theory of Universals », in *Australasian Journal of Philosophy*, 61/4, 1983, pp. 343-77.

D. Lewis, *On the Plurality of Worlds*, Oxford, Basil Blackwell, 1986a.

D. Lewis, « Postscripts to "Causation" », in *Philosophical Papers*, volume II, Oxford, Oxford University Press, 1986b, pp. 159-213.

D. Lewis, « Causation as Influence », in *Journal of Philosophy*, 97/4, 2000, pp. 182-197.

C. List et P. Menzies, « Non-reductive physicalism and the limits of the exclusion principle », in *Journal of Philosophy*, 106/9, pp. 475-502.

P. Ludwig, « Analyse conceptuelle et connaissance phénoménale », in *Klēsis*, 24, 2012, pp. 194-238.

P. Menzies, « Difference-making in context », in J. Collins, N. Hall et L.A. Paul(éds.), *Causation and Counterfactuals*, Cambridge, Mass., MIT Press, 2004.

P. Menzies, « Counterfactual Theories of Causation », in E. Zalta (éd.), *The Stanford Encyclopedia of Philosophy*, edition Printemps 2014, 2014.

D. Nolan, *David Lewis*, Chesham, Acumen Publishing, 2005.

J. Pearl, *Causality : Models, Reasoning, and Inference*, Cambridge, Cambridge University Press, 2000.

J. Schaffer, « Trumping Preemption », in *Journal of Philosophy*, 97/4, 2000, pp. 165--181.

J. Schaffer, « Causation, Influence, and Effluence », in *Analysis*, 61/1, 2001, pp. 11--19.

J. Stone, « Trumping the Causal Influence Account of Causation », in *Philosophical Studies*, 142/2, 2009, pp. 153--160.

M. Strevens, « Against Lewis's New Theory of Causation : A Story With Three Morals », in *Pacific Philosophical Quarterly*, 84/4, 2003, pp. 398-412.

J. Woodward, « A Functional Account of Causation ; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters – Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment) », in *Philosophy of Science*, 81/5, 2014, pp. 691-713.