Jacques Dubucs
CNRS/Université de Paris I — Panthéon-Sorbonne

LES ARGUMENTS DÉFAISABLES

Nous ne considérons un énoncé mathématique H comme justifié sur la base des axiomes $\mathcal{E}$ que lorsque nous possédons une preuve de H à partir de $\mathcal{E}$. Une preuve est une justification catégorique et définitive : un théorème n'est pas plus ou moins prouvé, et lorsqu'il l'est aucune découverte ultérieure ne peut lui ôter ce statut. En dehors des mathématiques de telles justifications catégoriques et définitives n'existent pas, et il nous arrive d'accepter des énoncés H dont le contenu excède celui des énoncés $\mathcal{E}$ qui les justifient. Il en va ainsi, typiquement, lorsque nous nous fions aux données de la perception, et que nous concluons de « l'objet *a* semble rouge » à « l'objet *a* est rouge ». Si l'on pense que la seule notion envisageable de justification est celle de justification probante qui a cours en mathématiques, force est bien de conclure que toutes nos croyances relatives au monde empirique sont injustifiées : a) les données de l'expérience sensorielle ne nous fournissent aucune garantie stricte de l'existence du monde extérieur ; b) l'observation du comportement verbal et non verbal d'autrui ne nous donne que des indications incomplètes sur ses états mentaux ; c) les données de l'expérience sont incapables de certifier strictement nos prédictions et nos énoncés universels relatifs à l'univers physique.

La seule échappatoire au scepticisme consiste donc à faire droit, dans le domaine empirique, à une notion de justification qui reste en-deçà de la preuve mathématique. Mais aucune contrepartie de la preuve ne s'impose d'elle-même pour les énoncés empiriques, ni aucune façon canonique de procéder pour relâcher, en ce domaine, l'exigence de justification catégorique et définitive des énoncés. Fondamentalement, la situation dans laquelle le contenu d'un énoncé H

excède celui des énoncés $\mathcal{E}$ qui le justifient peut être caractérisée de deux façons distinctes. D'une part en termes métriques de justification partielle : parmi les mondes possibles où les énoncés $\mathcal{E}$ sont réalisés, la proportion de ceux qui vérifient H est, quoique non nulle, inférieure à 1. D'autre part en termes non métriques de justification défaisable : $\mathcal{E}$, quoique non nul, peut être étendu à un ensemble $\mathcal{E}' \supset \mathcal{E}$ tel que $\mathcal{E}'$ implique la négation de H. L'articulation entre ces deux notions de justification est un problème central de la théorie de l'argumentation. C'est par là que je commencerai.

Probabilité et défaisabilité

Dans les années cinquante, Carnap a développé une théorie métrique (probabiliste) des arguments sub-démonstratifs, expressément destinée à définir le degré auquel l'assertion d'un énoncé est justifiée sur la base d'un ensemble de prémisses ne l'impliquant pas. Cette « logique inductive » soulève un certain nombre d'objections, qui paraissent militer en faveur d'une autre analyse de l'argumentation sub-démonstrative. Je n'en mentionnerai que deux, baptisées par commodité l'objection inductiviste et l'objection wittgensteinienne.

L'objection inductiviste

Un argument sub-démonstratif est un argument dans lequel la conclusion est un énoncé dont la fausseté n'est pas incompatible avec la vérité des prémisses. Or Carnap[1] ne concevait pas de cette manière les « inférences inductives », qui pour lui devaient simplement conduire à des « degrés de confirmation ». Si E_t consigne, parmi les données accessibles à un certain agent à l'instant t, la classe de toutes celles qui sont pertinentes pour H, alors une inférence inductive peut seulement permettre à l'agent de conclure que H est confirmé au degré $Pr(H/E_t)$. La conclusion de cette inférence ne consiste pas à détacher l'énoncé H, mais simplement à déconditionnaliser la probabilité qui lui est affectée. En d'autres termes, la logique inductive dans le format carnapien ne constitue pas une véritable théorie des arguments inductifs, c'est-à-dire une théorie expliquant les conditions dans lesquelles il est possible d'accepter de manière partiellement justifiée un énoncé sur la base de certaines prémisses qui ne l'impliquent pas déductivement, mais plutôt une théorie des probabilités expliquant la possibilité d'asserter, de manière catégoriquement justifiée, à quel degré l'énoncé considéré est « confirmé » par les prémisses en question.

L'expression « accepter une hypothèse » est elle-même ambiguë, signifiant tantôt « juger, ayant considéré l'hypothèse, qu'elle est correcte », tantôt « s'apprêter à agir comme si l'hypothèse était correcte ». Carnap soutenait qu'il ne peut jamais être rationnel d'accepter (au premier

sens) une hypothèse qui n'est pas impliquée (et dont nous savons qu'elle n'est pas impliquée) par l'évidence disponible, éventuellement jointe aux énoncés que nous acceptons déjà. Accepter l'hypothèse H, pour un agent rationnel, consiste simplement à entreprendre l'action A_H qui serait la plus appropriée pour atteindre ses objectifs si H était le cas : le comportement de l'agent n'est pas le signe extérieur — déchiffrable à qui connaîtrait ses objectifs — d'une adhésion intellectuelle à l'hypothèse, il n'est pas l'extériorisation de l'acceptation, mais l'acceptation elle-même. Au fond, les véritables « conclusions » d'un « argument » inductif sont les actions, en sorte que, comme le dit J. Neyman, on devrait, dans cette perspective, parler de « comportements inductifs » plutôt que d'« inférences inductives » : si $Créd_t(H_i)$ représente le degré auquel H_i est justifié par les données disponibles à l'instant t, et si u_{ij} est le degré de satisfaction éprouvé devant le résultat de l'action A_j dans le cas où l'hypothèse H_i serait la bonne, la « conclusion » tirée des données est simplement, via les désirs de l'agent, l'action qui maximise l'utilité espérée $\sum_i$ $créd_t$ (H_i). u_{ij} .

Dans cette perspective, on maintient donc, pour l'essentiel, que les seuls énoncés réellement justifiés sont ceux qui sont infailliblement garantis par l'évidence disponible. Et comme dans le domaine empirique une telle garantie est hors d'atteinte, on concède donc au scepticisme humien la destruction de la rationalité cognitive (entendue comme adéquation entre les informations disponibles et les énoncés acceptés), en se proposant seulement de maintenir intacte l'idée d'une rationalité comportementale (entendue comme adéquation entre les fins poursuivies et les moyens mis en œuvre). Désormais les seuls énoncés acceptables (au sens théorique de « assertables ») sont les tautologies, mais n'importe quel énoncé non contradictoire peut être accepté (au sens pragmatique de « fournir la base d'une action ») : dans une loterie de 1 million de billets dont un seul gagnant, l'hypothèse que votre billet va être le bon, si invraisemblable soit-elle, doit être acceptée si le lot est plus d'un million de fois supérieur au prix du billet, car dans ces conditions c'est en achetant un billet que vous maximiserez votre utilité espérée. Cette perspective réussit donc à montrer qu'il peut être pragmatiquement rationnel d'accepter un énoncé qui n'est pas déductivement impliqué par les données disponibles. Mais elle échoue, et c'est le sens de l'objection « inductiviste » qui peut lui être faite, à montrer qu'une telle acceptation peut être intellectuellement rationnelle[2].

L'objection wittgensteinienne.

Wittgenstein, qui n'est que très marginalement préoccupé par le scepticisme sur l'induction, soutient qu'une perspective comme celle qui vient d'être exposée est en tout cas radicalement inapte à triompher du scepticisme relatif aux états mentaux d'autrui, et que le relâchement exigible dans la notion de justification catégorique et définitive ne peut pas se limiter, dans ce domaine, à faire place à une relation de confirmation partielle.

1. Si un individu manifeste les comportements caractéristiques de la douleur (gesticulations, cris, etc), l'énoncé qui relate ce comportement confère un certain soutien à l'énoncé selon lequel l'individu souffre effectivement. Mais Wittgenstein considère que ce soutien n'est pas de nature inductive. Les raisons qu'il avance reposent pour l'essentiel sur une distinction largement répandue entre deux manières d'utiliser un mot. On peut l'utiliser sur la base de sa signification. Ainsi du mot « pluie » dans la phrase :

(P) *Le baromètre baisse : voilà la pluie*

qui affirme et motive l'imminence de ce que « pluie » désigne. Mais un tel usage, pour être intelligible, suppose que l'on connaisse déjà la manière correcte d'utiliser le mot, c'est-à-dire les usages qui en définissent la signification. C'est précisément de cette manière là que la phrase :

(P') *Je sens du froid et du mouillé : voilà la pluie*

utilise le mot. La sensation de froid et d'humidité ne peut pas être traitée comme un indicateur probabiliste de la pluie, car c'est un phénomène qui tire sa vertu indicatrice non pas d'une régularité de la nature, mais d'une convention de la « grammaire » : si je sens une certaine qualité de froid et d'humidité, je suis justifié immédiatement, et pour ainsi dire par définition, à affirmer qu'il pleut, car c'est précisément cela que signifie « pleuvoir ». Mais la phrase P', bien qu'elle exprime une relation « interne », indépendante du monde empirique, ne peut pas être considérée comme une tautologie, ni la sensation qu'elle rapporte comme une justification catégorique et définitive de « il pleut ». Car cette phrase ne dit la signification de « il pleut » que dans des circonstances normales, c'est-à-dire semblables aux circonstances paradigmatiques dans lesquelles nous avons appris à utiliser le verbe « pleuvoir », et dans lesquelles nous avions au moins la certitude qu'il était d'un emploi correct. Et la sensation en question ne justifie donc l'assertion de « il pleut » que sous réserve de la normalité des circonstances dans lesquelles elle apparaît (quoiqu'elle la justifie, dans ce cas, de manière pleine et non partielle).

À en croire Wittgenstein, la donnée du comportement pathoïde de X apporte exactement le même genre de soutien à l'énoncé « X souffre » : un soutien non pas inductif mais « critériel », c'est-à-dire un soutien fourni par un élément qui est à la fois l'un de ceux qui justifient l'énoncé et l'un de ceux qu'il convient d'invoquer pour en expliquer la signification. Cette requalification a plusieurs conséquences importantes :

— (i) Il existe des circonstances, à savoir celles qui sont paradigmatiquement adéquates pour apprendre l'usage assertorique du mot « souffrir », dans lesquelles le soutien en question est maximal, c'est-à-dire dans lesquelles l'assertion est *entièrement* justifiée. Lorsque le sceptique conteste qu'il y ait une seule circonstance dans laquelle nous puissions parvenir à la certitude en ce domaine, il n'emploie donc pas les mots au même sens que nous (selon Wittgenstein, le scepticisme est avant tout une erreur grammaticale).

— (ii) Notre usage du mot « souffrance » n'est pas assez rigide pour qu'il soit possible d'expliciter toutes les circonstances dans lesquelles son emploi est légitime. Entre les cas paradigmatiques dans lesquels l'emploi du mot est clairement prescrit et ceux dans lesquels il

est, non moins clairement, proscrit, il subsiste donc une variété de cas dans lesquels aucune règle ne s'applique. Cette indétermination, (qui n'empêche nullement le mot d'être, selon l'expression favorite de Wittgenstein, « en ordre[3] », n'est pas d'ordre intensif : il n'y a pas de paramètre ou de « méta-règle » qui pourrait servir à mesurer, depuis les cas d'assertabilité manifeste jusqu'aux cas d'évidente réfutabilité, le degré auquel est justifiée l'assertion de la phrase « X souffre » (l'absence d'un tel paramètre est encore plus claire dans le cas d'un mot comme « chaise[4] » : il n'existe évidemment aucun moyen de déformer continûment une chaise en une non-chaise).

— (iii) Le phénomène auquel le sceptique se réfère en parlant de « l'inaccessibilité de l'esprit d'autrui » se résume à une particularité grammaticale que l'on peut présenter de la manière suivante : une scène qui commence par ressembler exactement à la scène paradigmatique à laquelle est associé l'emploi d'un mot comme « souffrance » peut toujours évoluer, et se transformer pour finir en une scène en tout point semblable à celle dans laquelle cet emploi est formellement déconseillé. Or dans un tel cas de simulation prolongée, nous ne dirions ni que les critères qui fondent l'attribution de la douleur n'étaient pas, au début, satisfaits — car de toute évidence ils l'étaient—, ni (prétextant que c'est justement cela, donner de tels signes, que nous entendons par « souffrir ») que le simulateur souffrait vraiment — car de toute évidence les critères qui excluent l'attribution de la douleur étaient, à la fin, satisfaits, et c'est à la fin que nous jugeons. Nous devons donc nous résigner à admettre à la fois que les critères qui justifient l'assertion étaient satisfaits dans l'état d'information initial, et que l'assertion n'était plus justifiée dans l'état d'information étendu. Les justifications critérielles sont donc *défaisables* : chaque situation dans laquelle un critère justifie une assertion peut être prolongée de façon cohérente en une situation dans laquelle l'assertion n'est plus justifiée. Le relâchement introduit par le programme wittgensteinien dans la notion de justification catégorique et définitive porte précisément sur cette dernière propriété : il renonce à l'un des principes fondamentaux de la sémantique intuitionniste, celui de la clôture épistémique du futur (une chose une fois justifiée est justifiée pour toujours[5]).

2. Il n'est pas facile d'évaluer l'efficacité de la stratégie wittgensteinienne contre le scepticisme. Car le remède qu'il propose — introduire une notion défaisable de justification — est étrangement semblable à la maladie elle-même — croire que toutes nos justifications peuvent être défaites. À défaut de contraindre le sceptique à rendre les armes, cette stratégie a en tout cas pour effet de faire apparaître ses convictions comme déraisonnables. Ceci prend deux formes (non clairement distinguées chez Wittgenstein lui-même) :

— (i) Le renversement de l'asymétrie argumentative sur laquelle se fonde le sceptique, renversement que Hacker[6] résume en ces termes : alors que le sceptique affirme qu'une croyance doit être présumée coupable jusqu'à ce qu'elle soit reconnue innocente, Wittgenstein soutient qu'une croyance critériellement justifiée est innocente jusqu'à ce qu'elle soit reconnue coupable. En vertu de ce renversement, les justifications critérielles sont des justifications *prima facie*, qu'il suffit de compléter par l'absence d'évidence contre ce qu'elles justifient, et la charge de la preuve incombe donc à celui qui doute.

— (ii) L'abandon de ce que l'on pourrait nommer le principe d'isotropie des possibles, qui affirme la parité de tous les scénarios cohérents du monde. Wittgenstein considère que le monde dont le sceptique envisage la possibilité avec une telle délectation (un monde dans lequel toutes mes expériences seraient des hallucinations, dans lequel tous ceux qui ont l'air de souffrir seraient des comédiens, et où tous ceux qui semblent en paix seraient des stoïques) est suprêmement déraisonnable, et qu'un monde raisonnable est au contraire un monde dans lequel une assertion critériellement justifiable est satisfaite dès lors que ses critères de justification le sont déjà.

Justifications *prima facie*

Une justification *prima facie* est une justification qui peut être tenue, tant qu'elle n'est pas « défaite », pour une justification tout court. Dans l'exemple suivant, E constitue ainsi un argument *prima facie* en faveur de H, argument qui peut être défait par E' :

(E) L'objet *a* semble ϕ

(H) L'objet *a* est ϕ

(E') L'objet *a* n'est pas ϕ, et ne semble tel que parce que certaines conditions « anormales » sont réalisées.

Défaites individuelles

Admettons pour l'instant que nous saurions rigoureusement caractériser (peut-être en termes d'augmentation de probabilité pour H, mais peut-être autrement) la notion générale d'argument pour ou en faveur de H, c'est-à-dire le genre dont nous essayons de définir une espèce. Alors la définition suivante s'impose d'elle-même :

(PF) Un argument *prima facie* pour H est un argument E pour H possédant la particularité suivante : il existe un énoncé E' cohérent avec E et tel que la conjonction EE' entraîne la négation de H.

On voit immédiatement que cette définition peut être généralisée de diverses manières, et notamment qu'elle peut être affaiblie en :

(PF') Un argument *prima facie* pour H est un argument E pour H possédant la particularité suivante : il existe un énoncé E' cohérent avec E et tel que la conjonction EE' est un argument contre H.

Par ailleurs une analyse plus poussée montre qu'un argument *prima facie* peut être défait de deux façons très distinctes[7] :

— (i) D'une part l'énoncé capable de venir à bout de l'argument, celui qui est, si l'on

ose dire, son « défaiseur » — l'énoncé E' de notre définition — peut tout simplement entraîner à lui seul la négation ¬ H de la conclusion (ou encore, dans une version généralisée, constituer à lui seul un argument contre H). Tous les exemples que nous avons donnés jusqu'ici sont de ce type, dans lequel les arguments *prima facie* sont défaits par annulation de leur conclusion.

— (ii) D'autre part un énoncé peut défaire un argument *prima facie* en affirmant l'occurrence d'une situation dans laquelle la conclusion peut certes être correcte, mais où la vérité des prémisses ne peut plus être considérée comme une indicateur fiable de cette correction. En somme les arguments *prima facie* sont alors défaits non par l'annulation de leur conclusion, mais par l'annulation du lien argumentatif entre les prémisses et la conclusion. Un exemple de cette dernière situation peut être obtenu en prenant pour 'ϕ' le prédicat 'rouge' dans les énoncés E et H ci-dessus, et en considérant l'énoncé :

(E'') L'objet *a* est éclairé en rouge

(un objet éclairé en rouge peut fort bien être rouge, mais la lumière rouge n'est pas une circonstance dans laquelle l'apparence rouge des objets peut être considérée comme un indicateur fiable de leur rougeur « intrinsèque »[8]).

Défaites collectives

Deux arguments *prima facie* peuvent entrer en conflit[9] : il suffit pour cela que la conclusion de l'un « défasse » l'autre (en l'un quelconque des deux sens de ce mot). On a alors affaire à une défaite « collective », qui résulte de l'interaction des arguments.

1. Le cas le plus simple est celui où deux arguments possèdent des prémisses qui peuvent être simultanément vraies mais des conclusions contradictoires l'une de l'autre. Il est habituellement illustré par l'exemple de Nixon, lequel devait être à la fois belliciste — puisque, allègue le premier argument, les Républicains sont, *prima facie*, des fauteurs de guerre ; et pacifiste — puisque, allègue le second, les quakers comme lui sont, *prima facie*, des partisans de la paix. Dans cette situation, où deux propositions cohérentes entre elles sont prémisses d'arguments dont les conclusions sont contradictoires, les arguments *prima facie* sont défaits non parce qu'une proposition avérée vient contredire par après les conclusions que l'on avait tiré d'eux, mais pour ainsi dire *ex ante*, par le simple jeu de leur concurrence dans un système argumentatif.

2. Le cas le plus pur de ces conflits d'arguments est cependant celui où une seule et même proposition est prémisse d'arguments dont les conclusions sont contradictoires les unes des autres, ou qui sont, ensemble, contradictoires avec une proposition d'ores et déjà tenue pour certaine. Il est illustré par le paradoxe de la loterie (Kyburg, 1961) : si l'on sait que, dans une loterie équitable d'un million de billets, un seul est gagnant, on a un argument *prima facie* pour affirmer, à propos du i-ème billet, qu'il n'est pas le billet gagnant. Mais comme un argument semblable peut s'appliquer à chacun des billets, on devrait en conclure qu'aucun billet n'est le bon, ce qui contredit l'information qu'il en existe un : en somme l'énoncé E qui décrit la nature

de la loterie est la prémisse d'arguments *prima facie* dont les conclusions, combinées ensemble, contredisent E.

La description logique des arguments prima facie

Les arguments *prima facie* diffèrent des arguments déductifs qui sont l'objet de la logique usuelle. Ils s'en distinguent par leur « non-monotonie » : alors que l'ensemble des conclusions d'un argument déductif croît avec l'ensemble de ses prémisses (si H est conséquence logique de E, H est encore conséquence logique de tout sur-ensemble de E), l'adjonction d'une prémisse peut conduire à retirer la conclusion d'un argument *prima facie* (il suffit que cette prémisse soit un « défaiseur » de l'argument en question). Mais cette singularité en entraîne d'autres, qui semblent bien hypothéquer la possibilité de décrire systématiquement ces arguments dans la tradition logique issue de Frege et de Hilbert :

— (i) L'admission d'arguments *prima facie* interdit la composition transitive des inférences : la propriété de non-monotonie consiste précisément en ce qu'un argument déductif qui conduit de E à F ne peut pas se composer avec un argument conduisant de F à H pour donner un argument de E à H[10].

— (ii) Cette absence de transitivité semble entraîner l'impossibilité d'effacer les prémisses d'une inférence. Par exemple, il est impossible de conclure de :

(a) Les étudiants de troisième cycle sont, *prima facie*, des adultes

et de :

(b) Les adultes sont, *prima facie*, salariés

à :

(c) Les étudiants de troisième cycle sont, *prima facie*, salariés

(les étudiants typiques sont des adultes atypiques *(Cf. fig. 1)*.

Avant d'affirmer que l'adulte X est un salarié, il convient donc de se souvenir de la raison pour laquelle X a été qualifié d'adulte, car si c'est en raison de sa qualité d'étudiant, il faudra s'abstenir de cette affirmation. Contrairement aux arguments déductifs, qui sont, si l'on ose dire, du type *fire and forget* (on décoche la conclusion et l'on oublie les prémisses), les arguments *prima facie* ne peuvent donc être utilisés sans en garder les prémisses en mémoire.

En somme, dans de tels arguments, la conclusion n'est qu'apparemment détachée : sa généalogie tout entière doit être conservée afin d'être réexaminée lorsqu'il s'agira de l'engager dans une nouvelle inférence. Comme le remarque Girard (1992), la situation est donc à peu près comparable à celle qui régnerait dans les échanges marchands si au moment de la transaction l'acheteur devait se munir non seulement de la somme requise en numéraire, mais aussi de tous les objets qui, jusqu'à la vache primitive, ont été successivement cédés par lui-même et ses ancêtres pour parvenir à la réunir...

— (iii) En présence d'arguments conflictuels (*Défaites collectives* § 1.), les conclusions auxquelles on parvient à partir d'un ensemble donné de propositions dépendent de l'ordre dans lequel ont été invoqués les arguments. Cette sensibilité à l'ordre d'application des règles (Nixon belliciste ou pas selon qu'il est d'abord appréhendé comme républicain ou comme quaker) est évidemment tout à fait étrangère à la logique déductive : l'ensemble des théorèmes d'un système formel, quant à lui, ne dépend pas de l'ordre dans lequel ont été appelées les règles d'inférence dans les preuves.

Devant des différences aussi fondamentales entre les arguments déductifs et les arguments *prima facie*, on peut être tenté de souscrire à l'opinion de Girard, pour lequel une description prétendument logique de ces derniers ne saurait être qu'une « paralogique », une fausse science entretenant avec la logique usuelle à peu près le même rapport que celui de l'astrologie à l'astronomie *(loc. cit.)*. Pour juger du bien-fondé de cette opinion, il suffit au fond d'examiner deux questions :

— 1°) Notre usage des arguments défaisables est-il tout à fait erratique, ou bien manifeste-t-il certaines régularités capables d'être précisément décrites ? Dans la premier cas, l'entreprise est désespérée, car on ne saurait formaliser sans témérité que les activités intellectuelles de nature relativement routinière, possèdent déjà une certaine rigueur « informelle ».

— 2°) Un système se présentant comme une logique des arguments défaisables peut-il fournir un mécanisme inférentiel permettant de dériver de manière effective l'ensemble des énoncés que ces arguments justifient ? Dans le cas contraire, un tel système, qui n'a aucune chance d'être réalisé dans un dispositif informatique comme peuvent l'être les systèmes formels de la logique classique, est dénué d'utilité.

Les règles de l'argumentation défaisable

La question de savoir si un énoncé est ou non un argument en faveur d'un autre énoncé est en général une question non formelle, c'est-à-dire une question qui n'a aucun sens pour des énoncés ininterprétés, et qui ne peut être tranchée que par un « expert » du domaine auquel les énoncés interprétés font référence : c'est au médecin et non au logicien de dire si un symptôme donné doit évoquer telle ou telle étiologie, et combien elle est accréditée par sa manifestation. En revanche, la question de savoir comment ces arguments se combinent les uns avec les autres, cette question-là est une question qui *peut* être une question formelle[11], et qui peut l'être en deux sens. D'une part si nous supposons déjà disponible une analyse technique de ce qu'est pour E que d'être argument pour H (par exemple : la probabilité de H sachant E excède un certain seuil, ou bien excède la probabilité absolue de H), on peut aisément définir les règles de combinaison qui préservent la relation argumentative ainsi analysée, sur le modèle des règles familières qui préservent la relation de conséquence logique (par exemple si nous identifions « E est argument pour H » à « en présence de E, la probabilité de H est haute », nous devrons

assurément considérer comme une règle formelle que si E est argument pour H, alors E est encore argument pour H V K).

Mais il n'est nullement nécessaire de posséder déjà une telle analyse, qui risque en outre de biaiser la description de notre pratique argumentative (après tout, il n'est pas du tout certain que cette pratique puisse être décrite dans un cadre probabiliste). Car pour que l'on puisse parler ici de règles formelles, il suffit que notre manière de combiner les arguments entre eux soit uniforme, c'est-à-dire à la fois à la fois stable et indépendante du domaine particulier auquel se réfèrent les énoncés en jeu dans ces arguments : bien que l'évaluation des arguments primitifs (« atomiques ») soit indiscutablement de la compétence du médecin ou du garagiste, leur mode d'articulation et d'agencement en arguments complexes devrait être *schématique*, et ne dépendre que de la forme des éléments manipulés. Or tel semble bien être le cas, comme il apparaît au travers des exemples suivants :

1. Il existe des variétés de conflits argumentatifs que nous arbitrons systématiquement dans le même sens. C'est le cas lorsqu'il est possible de distinguer, à l'intérieur d'une classe d'objets défaisablement tenus de satisfaire un certain prédicat (comme les adultes, dans l'exemple (ii) de la *description logique des arguments prima facie*, le sont d'être salariés), une sous-classe qui en représente une exception « nomique », c'est-à-dire dont les membres sont défaisablement tenus de ne pas satisfaire le prédicat en question (dans notre exemple : les étudiants de troisième cycle). Dans ces conditions, deux énoncés contradictoires sont *prima facie* justifiés si X est un étudiant : d'une part « X n'est pas salarié » (si l'on applique d'abord l'argument qui concerne la sous-classe), et d'autre part « X est salarié » (si l'on commence par déduire « X est un adulte » et que l'on applique ensuite l'argument qui concerne la classe entière). Mais contrairement à la situation visiblement indécidable décrite en 1. de *Défaites collectives* (le « diagramme de Nixon »), il y a évidemment ici une bonne manière de procéder : c'est uniquement au titre de membre de la sous-classe qu'un individu doit être appréhendé. En d'autres termes, nous appliquons régulièrement, dans des cas de ce type, un principe de priorité qui favorise les arguments *prima facie* dont les prémisses ont la spécificité la plus grande[12].

2. Bien qu'il soit généralement impossible de renforcer les prémisses d'un argument défaisable (cf le 1° de *Description logique des arguments prima facie*), il y a au moins un cas où nous nous autorisons systématiquement à le faire, celui dans lequel deux arguments se confortent l'un l'autre pour donner naissance à un argument *a fortiori*. Supposons en effet que les ϕ sont, *prima facie*, des ψ, et que d'autre part ils sont, *prima facie*, des χ. Alors nous serions désappointés d'apprendre que certains d'entre eux ne sont pas des ψ. Mais nous serions encore plus désappointés d'apprendre que ces ϕ inattendus appartiennent précisément à l'espèce la plus typique de ϕ, à savoir $\phi\chi$: *a fortiori* les $\phi\chi$ sont donc encore, *prima facie*, des ψ. En d'autres termes, nous admettons régulièrement que les prémisses d'un argument *prima facie* soient renforcées par la conclusion de tout argument *prima facie* possédant les mêmes prémisses (renforcement restreint des prémisses[13]).

3. De même, bien qu'il soit généralement impossible de composer transitivement deux arguments défaisables, il y a des cas où nous le faisons tout aussi systématiquement. Car les

entorses à la transitivité argumentative proviennent régulièrement d'une infraction à une clause *ceteris paribus* implicite dans les prémisses de l'un des deux arguments composés, comme il apparaît dans l'exemple suivant, de :

(a') *Les gens qui travaillent moins sont plus détendus*

et de :

(b') *Les gens qui sont au chômage travaillent moins*

ne découle nullement que :

(c') *Les gens qui sont au chômage sont plus détendus.*

Mais si la transitivité est ici en défaut, c'est que (a') est un énoncé elliptique dont la forme développée serait :

(a") *Les gens qui travaillent moins* — toutes choses égales par ailleurs — *sont plus détendus*,

et que la clause « *toutes choses égales par ailleurs* » n'est précisément pas satisfaite dans la situation décrite par l'antécédent de (b'). Cette clause contient — mais de manière virtuelle, non explicite — les conditions N_a' de « normalité » (X n'est pas chômeur, etc.) dans lesquelles la prémisse « X travaille moins » pourrait être considérée comme un argument indéfaisable en faveur de la conclusion « X est plus détendu ». Dès lors, et compte tenu de la façon implicite dont nous sont données les conditions N_a' :

— 1°) En renforçant la prémisse « X travaille moins », on risque de lui adjoindre un énoncé qui viendrait contredire l'une de ces conditions cachées, transformant ainsi (a") — et donc (a') — en trivialité du type *ex falso quodlibet* (telle est la vraie raison de la « non-monotonie » des arguments défaisables).

— 2°) En composant avec (a") — et donc avec (a') — un argument dont la conclusion serait « X travaille moins » on risque que la prémisse E de ce nouvel argument vienne contredire l'une de ces conditions cachées, transformant à nouveau (a') en trivialité selon le schéma :

Si E, alors X travaille moins
Si X travaille moins (et si N_a'), alors X est plus détendu

Si E (et si N_a'), alors X est plus détendu

(telle est la vraie raison de l'absence de transitivité des arguments défaisables).

Mais il n'y a aucun risque de cet ordre lorsque E, conjointement à « X travaille moins », est aussi un argument (non trivial) en faveur de « X est plus détendu ». Car dans ces conditions E ne contredit certainement aucune des conditions implicites sous lesquelles « X travaille moins » est un argument indéfaisable pour « X est plus détendu ». On obtient alors une forme restreinte de transitivité qui est un mode de combinaison indiscutablement correct des arguments défaisables et qui, appliquée à notre exemple, est la suivante :

Si X est chômeur, alors X travaille moins
Si X est chômeur et que X travaille moins, alors X est plus détendu

---

Si X est chômeur, alors X est plus détendu

(l'incorrection de la conclusion ne provient plus de l'incorrection du schéma d'inférence, mais simplement de l'incorrection de la seconde prémisse[14]).

Les exemples 1., 2. et 3. montrent que nous faisons des arguments défaisables un usage cohérent et régulier, et qu'il est possible de codifier cet usage par des principes généraux (priorité aux arguments dotés des prémisses les plus spécifiques) ou des règles d'inférence (renforcement restreint des prémisses, transitivité restreinte) qui sont adéquates relativement à la notion informelle d'argument défaisable correct. Reste à savoir si ces régularités peuvent être systématiquement décrites dans un système formel comparable à ceux qui codifient la pratique déductive. Nous allons voir que la manière dont les spécialistes de l'intelligence artificielle ont abordé cette question dans les années 1980 conduit à une impasse. (Cette erreur de perspective est à la source d'un scepticisme général — et à mon avis injustifié — sur la possibilité de jamais résoudre la question).

L'anisotropie des possibles

Le système de Reiter (1980) est l'une des premières tentatives pour décrire formellement la défaisabilité. Son objectif est de spécifier, étant donnés un ensemble $\mathcal{A}$ de propositions avérées (les « faits ») et un ensemble Δ de « règles par défaut » (contreparties formelles des arguments *prima facie*) du type (E : H/H) (« si E, et si l'on peut supposer H sans introduire de contradiction, alors H »), l'ensemble des énoncés dont l'assertion peut être considérée comme justifiée. Ces énoncés constituent une « extension » de $< \mathcal{A}, \Delta >$, c'est-à-dire un ensemble cohérent qui contient $\mathcal{A}$ et qui est, en un certain sens, saturé, c'est-à-dire qui contient toutes les conclusions que l'on peut tirer de ses membres, aussi bien par des arguments déductifs que par les règles « par défaut » contenues dans Δ. On montre aisément qu'une telle extension, censée représenter un « ensemble cohérent de croyances sur le monde » (Reiter, p.194) existe pour n'importe quelle paire $< \mathcal{A}, \Delta >$.

La notion d'extension est de nature sémantique, et son usage est à rapprocher de l'idée wittgensteinienne d'anisotropie des possibles (cf. *L'objection wittgensteinienne* : 2. [ii]). En un mot, tous les mondes possibles compatibles avec les informations détenues ne sont pas à parité : certains (les « extensions », justement) sont plus « normaux » ou plus « raisonnables » que d'autres. Par exemple si E est le seul fait connu, et que la seule règle « par défaut » est (E : H) / H, la seule extension est la clôture déductive Cl ({E,H}) de E et de H, qui est un monde possible plus « normal » que ceux qui contiennent E et $\neg$ H. Mais la difficulté rédhibitoire de cette « logique des défauts » est de ne nous donner aucun moyen général de construire effectivement les extensions. Une extension de $< \mathcal{A}, \Delta >$ ne peut en effet être obtenue que comme la réunion $\mathcal{E}$ de la suite infinie définie par :

$$\mathcal{E}_0 = \mathcal{A} \text{ et } \mathcal{E}_{n+1} = \mathrm{Cl}(\mathcal{E}_n) \cup \{ H ; [(E : H) / H] \in \Delta, E \in \mathcal{E}_n \text{ et } \mathcal{E} \dashv \neg H \},$$

ce qui n'a aucune espèce d'utilité pratique, puisque la construction de la $(n+1)$-ème étape de $\mathcal{E}$ requiert un test d'indémontrabilité (donc un test dont on sait qu'il ne peut être effectué par aucune machine) relatif au résultat final de la construction tout entière ! S'il existe une description logique correcte de l'argumentation défaisable, le système de Reiter ne la donne pas[15].

Arguments défaisables, enthymèmes et contextes argumentatifs

La grossière inadéquation du système de Reiter est le symptôme d'une erreur dans l'analyse conceptuelle des arguments défaisables, et il convient de la reprendre à nouveaux frais. L'objectif est de définir le statut logique de l'expression *prima facie* dans un argument comme :

(1) *a* semble rouge, donc *(prima facie) a* est rouge.

L'interprétation :

(2) *a* semble rouge, donc (sauf preuve du contraire) *a* est rouge

est incorrecte, car un agent qui entendrait la phrase (1) de cette façon devrait posséder des capacités de calcul exorbitantes (cf. § *L'anisotropie des possibles*, ci-dessus). L'interprétation à retenir est donc plutôt

(3) Dans les conditions normales, si *a* semble rouge, alors *a* est rouge

Ces « conditions normales » sont celles dans lesquelles la rougeur apparente de *a* est une garantie indéfaisable de sa rougeur réelle. (3) s'apparente donc à un argument déductif dont les prémisses sont d'une part « si *a* semble rouge » et d'autre part « si les conditions sont normales ». Mais l'expression « si les conditions sont normales » possède elle-même une signification qui peut varier d'un argument à l'autre (les conditions normales pour conclure à la rougeur ne sont pas celles qui permettent de conclure à la présence de CO_2). Donc cette expression n'est qu'un marque-place (d'ailleurs facultatif) destiné à désigner implicitement les conditions dans lesquelles la prémisse explicite de l'argument permet d'en déduire la conclusion (cf. *Les règles de l'argumentation défaisable*, exemple § 3.). (3) — et donc (1) — est un argument qui serait valide si une certaine prémisse était explicitée : c'est un enthymème.

À la lumière d'une telle interprétation, le contraste entre la « détachabilité » *(fire and forget)* des conclusions déductives et l'« adhérence » des conclusions défaisables, engluées dans les prémisses dont elles sont issues et dont il faut toujours, semble-t-il, garder la mémoire, ce contraste apparaît sous un jour nouveau et passablement moins inquiétant. Il s'agit simplement, on va le voir, du contraste entre des arguments qui dépendent de leur contexte argumentatif et des arguments qui n'en dépendent pas.

— 1°) Si l'on peut conclure déductivement à E dans le contexte Γ, et que d'autre part H peut être déduit de E dans le contexte Γ', alors les deux contextes Γ et Γ' peuvent être « fusionnés » pour déduire H. Tel est à peu près le sens de la transitivité des arguments déductifs, transcrite par ce que les logiciens appellent la règle de « coupure », et qui autorise à combiner deux déductions en une seule en « oubliant » le moyen-terme qui est à la fois conclusion de la première et prémisse de la seconde (la figure 2, en annexe, rend manifeste cette fusion des contextes argumentatifs dans le cas de la déduction).

— 2°) Une telle licence pour fusionner les contextes argumentatifs transformerait immédiatement les arguments défaisables en trivialités du type *ex falso quodlibet* : compte-tenu du caractère implicite des « conditions de normalité » dans lesquelles la prémisse y valide la conclusion, on se heurterait en effet au phénomène qu'illustre la figure 3 *(en annexe)* à propos de notre exemple-type (§ 3.). Reste alors, puisque l'on cherche des mécanismes qui donnent à l'argumentation défaisable une certaine « automaticité », et que l'on se refuse à juste titre à garder la mémoire de toute la généalogie des conclusions, à déterminer les formes « licites » de combinaison des arguments défaisables. Un mécanisme de ce genre ne peut être qu'un mécanisme requérant l'identité des contextes argumentatifs, c'est-à-dire un mécanisme capable seulement d'assurer que si l'on peut conclure à E dans le contexte Γ, et que d'autre part E permet de conclure à H dans ce *même* contexte Γ, alors il est permis de conclure à H (et donc d'oublier E !) dans le contexte Γ. Telle est, au fond, la vraie signification de la transitivité restreinte (§ *L'anisotropie des possibles.*) des arguments défaisables (cf. fig. 4, *en annexe).*

On restaure donc par ce biais, mais sous une forme locale, sensible au contexte, les mécanismes formels de combinaison argumentative qui caractérisent la logique usuelle. Sauf à repousser l'idée de règles logiques contextuelles[16], cet exercice ne transforme pas nécessairement la description des arguments défaisables en astrologique...

Jacques DUBUCS

NOTES

1. « *Le résultat de l'examen inductif ne peut pas consister dans la mention de l'énoncé seul : la valeur trouvée pour le degré de confirmation est une part essentielle du résultat* » (Carnap, 1950, p. 206). Cf. sur ce point Dubucs, 1989, p. 85-88.

2. Considéré comme une tentative de construire un concept de justification adéquat dans le domaine empirique, le programme carnapien soulève plusieurs objections, qui concernent notamment l'usage, en ce domaine, du calcul usuel des probabilités : toute mesure additive (vérifiant, pour tout A, $\mu(A) + \mu(\neg A) = 1$) du degré de justification nous contraint en effet à admettre que toute l'évidence favorable à un énoncé peut provenir de la seule faiblesse de l'évidence en faveur de sa négation, ce qui ne paraît pas représenter correctement les situations d'ignorance complète, dans lesquelles nous ne disposons d'aucune donnée pertinente à propos d'une hypothèse. Cette critique,

qui a donné naissance à la construction de métriques non additives pour les degrés de justification (par exemple la théorie de l'évidence de Shafer (1976)), ne met cependant pas en cause la possibilité même de s'en tenir en toute circonstance à une notion graduelle de justification.

3. Un signe « est en ordre si, dans les circonstances normales, il remplit sa fonction. » (Wittgenstein, 1958, § 87).

4. *Op. cit..*, § 80.

5. Dans tout modèle $M = < M, \leq, V >$, si $\models_w A$ et si $w \leq w'$, alors $\models_{w'} A$.

6. Cf. Hacker,1972, p. 303.

7. Cf. la distinction de Pollock (1987, p. 484-485) entre *rebutting defeaters* et *undercutting defeaters*.

8. Les arguments inductifs peuvent également subir les deux types de défaites :
   — le syllogisme

   Tous les ϕ de l'échantillon observé sont des ψ
   L'objet *a* (hors de l'échantillon) est ϕ

   L'objet *a* est ϕ

   peut être défait aussi bien par l'observation que *a* n'est pas ψ (annulation de la conclusion) que par l'annonce que l'échantillon est biaisé (mise en cause de la fiabilité du lien argumentatif entre les prémisses et la conclusion).

9. Les épigones de Wittgenstein avaient fort bien perçu la possibilité pour les « critères » d'entrer en conflit, mais ils s'étaient contentés d'en tirer un argument supplémentaire en faveur de la défaisabilité de la justification critérielle (cf. Baker 1974, p. 162).

10. En revanche, on a de bonnes raisons de penser qu'un argument *prima facie* doit pouvoir se composer à droite avec un argument déductif, c'est-à-dire qu'à défaut de pouvoir en renforcer les prémisses on devrait au moins pouvoir en affaiblir les conclusions. Cf. Dubucs, 1993, p. 91-92.

11. Cf. sur ce point Dubucs, 1974, p. 23, sq.

12. Cf. sur ce point Etherington et Reiter, 1983.

13. On notera que cette règle est incompatible avec l'interprétation des arguments *prima facie* en termes de haute probabilité de la conclusion connaissant les prémisses (Dubucs, 1993, p. 96 sq).

14. Adams (1975, p. 21 sq), qui est l'un des premiers à avoir proposé d'expliquer l'absence de transitivité des arguments par le caractère elliptique de leurs prémisses, est en tout cas le premier à avoir suggéré que l'on pouvait y remédier en exigeant que l'antécédent de l'un des deux arguments-prémisses puisse être conjoint à l'antécédent de l'autre. Son exemple de base est le suivant :

   Si Smith meurt avant l'élection, alors Jones gagnera
   Si Jones gagne, Smith prendra sa retraite après l'élection

   ---

   Si Smith meurt avant l'élection, il prendra sa retraite après,

   dans lequel il convient de remplacer la seconde ligne par :
   Si Jones gagne (toujours supposé que Smith meure avant l'élection), alors Smith prendra sa retraite après l'élection.

15. En outre, le système de Reiter ne permet pas de trancher entre plusieurs extensions dans les cas où le principe de priorité (*Les règles de l'argumentation défaisable*, 1.) nous donne pourtant d'impérieuses raisons d'en préférer certaines. Car dans la généralisation qui serait exigible pour le faire (système de défauts « semi-normaux »), le théorème assurant l'existence des extensions ne s'applique plus...

16. Les logiques « sub-structurales » (parmi lesquelles la logique linéaire !), dans lesquelles on restreint l'application des règles structurales (échange, affaiblissement ou contraction), donnent précisément la possibilité de contextualiser les règles logiques (Cf. par exemple, Troelstra, 1992).

RÉFÉRENCES BIBLIOGRAPHIQUES

ADAMS, Ernest W., *The Logic of Conditionals*. Reidel Publ., Dordrecht, 1975.

BAKER, Gordon P., « Criteria : a New Foundation for Semantics », *Ratio*, 1974, vol. 16, p. 156-189.

CARNAP, Rudolf, *Logical Foundations of Probability*. University of Chicago Press (2e édition), 1971.

CROCCO, Gabriella, *Fondements logiques du raisonnement contextuel*, Thèse d'Informatique de l'Université de Toulouse, 1993.

DUBUCS, Jacques, « Sur la logique des arguments plausibles », *Philosophie*, 1987, vol. 14, p. 15-37.
— « Probabilité et objectivité », *L'Âge de la science*, 1989, vol. 2, p. 83-98.
— Justification, probabilité, défaisabilité. *Cahiers de l'Université de Montréal*, n° 92 (hors série), 1992.
— « Inductive Logic Revisited », *in* J. Dubucs (ed.), *Philosophy of Probability*. Dordrecht, Kluwer, 1993, p. 79-108.

ETHERINGTON, David W., REITER Raymond, *On Inheritance Hierarchies with Exceptions*. IJCAI, Karlsruhe, 1983, p. 104-108.

GIRARD, Jean-Yves, *La logique au milieu du gué* (manuscrit), 1992.

KYBURG, Henry E., Jr, *Probability and the Logic of Rational Belief, Middletown.*. Conn., Wesleyan University Press, 1961.

HACKER, P.M.S., *Insight and Illusion. Londres*. Oxford University Press, 1972.

POLLOCK, John L., « Defeasible Reasoning », *Cognitive Science* 1987, vol. 11, p. 481-518.

REITER, Raymond, « A Logic for Default Reasoning », *Artificial Intelligence* 1980, vol. 13, p. 81-132.

SHAFER, Glenn., *A Mathematical Theory of Evidence*. Princeton U.P., 1976.

TROELSTRA, Anne S., *Lectures on Linear Logic*. Stanford, C.S.L.I. Lectures Notes 1992, vol. 29.

WITTGENSTEIN, Ludwig, *Philosophische Untersuchungen (Philosophical Investigations)*. Londres, Blackwell, 1958.

ANNEXES

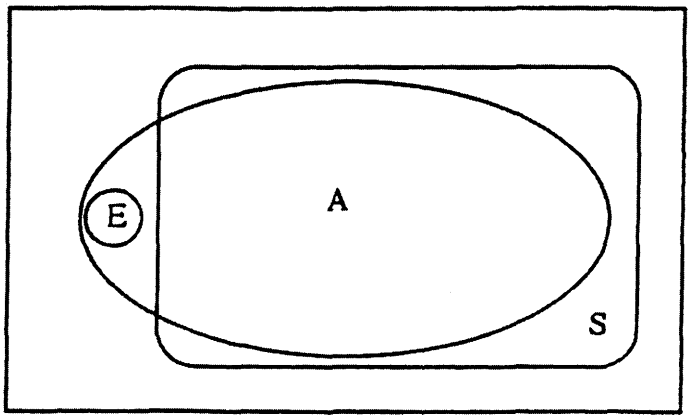

Fig. 1. *Les arguments défaisables ne sont pas transitifs*

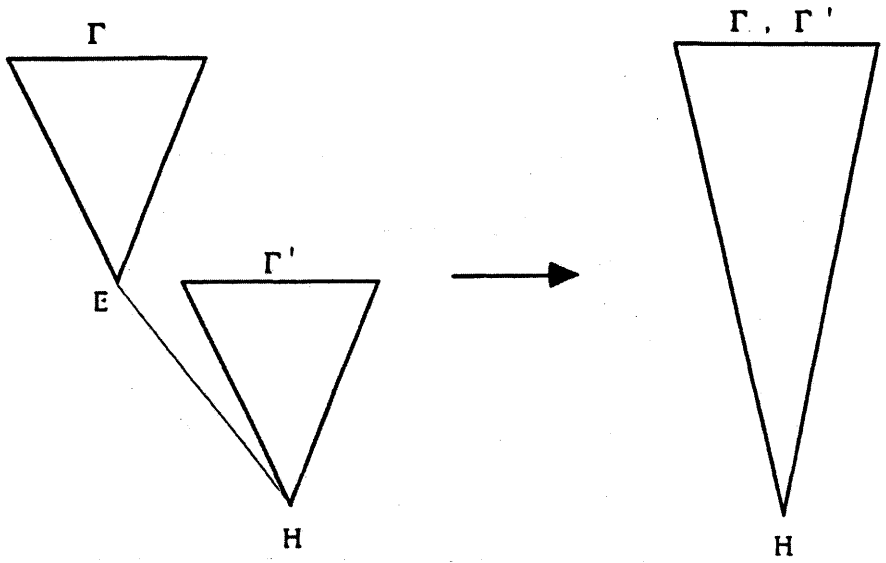

Fig. 2. *Fusion des contextes dans l'argumentation déductive*

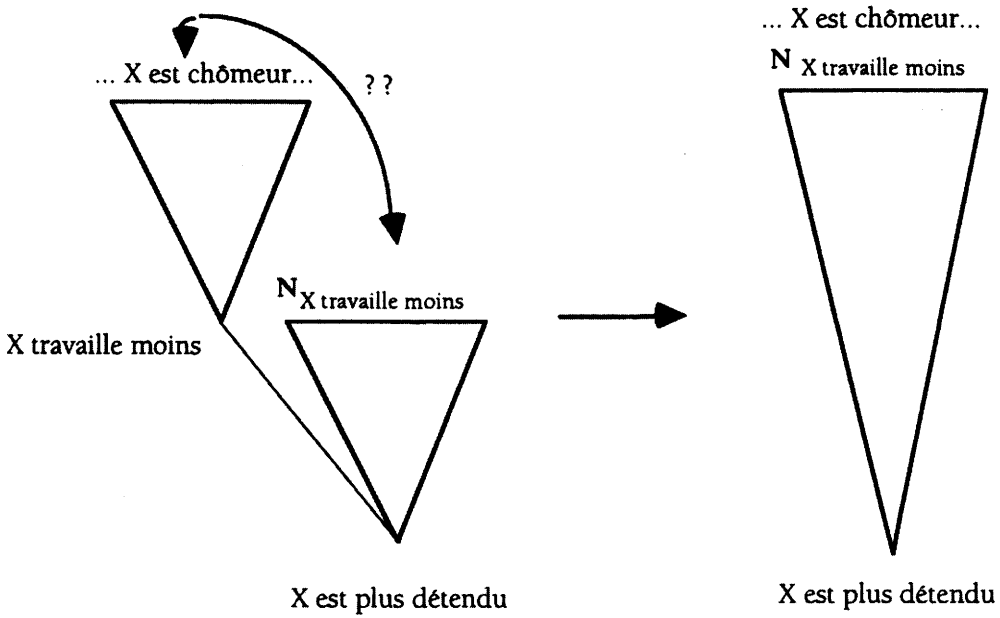

Fig. 3. *Absurdité de la fusion des contextes dans l'argumentation défaisable*

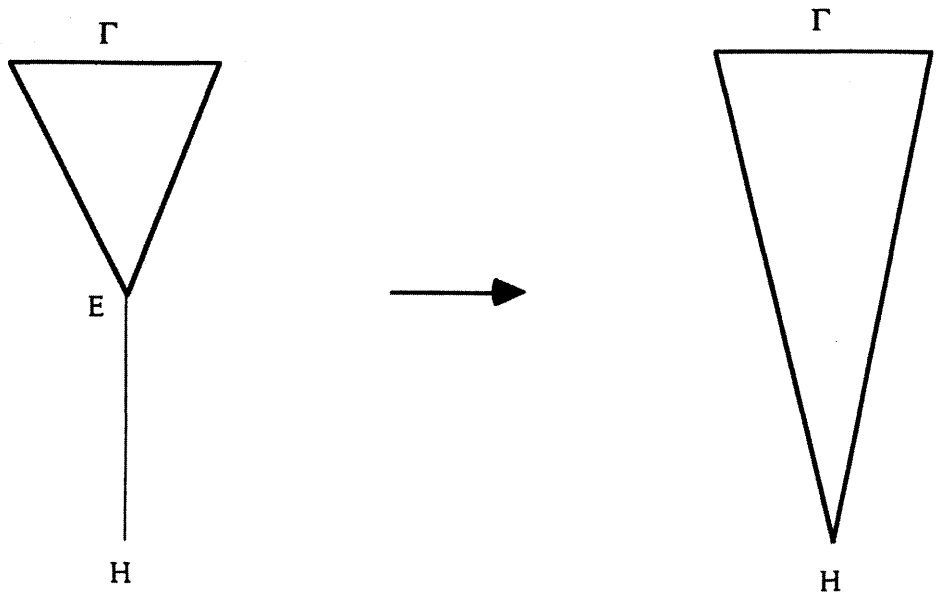

Fig. 4. *Transitivité restreinte (contextuelle) des arguments défaisables*