Quelques réflexions autour de la notion de bêtise artificielle

Jean Lieber, Jean-Guy Mailly, Pierre Marquis, Henri Prade, François Rollin

Quelques réflexions autour de la notion de bêtise artificielle *

Jean Lieber[1] **Jean-Guy Mailly**[2] **Pierre Marquis**[3,4]
Henri Prade[5] **François Rollin**

[1] Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy
[2] Université Paris Cité, LIPADE, F-75006 Paris
[3] Univ. Artois, CNRS, CRIL, 62300 Lens
[4] Institut Universitaire de France
[5] IRIT, Toulouse
jean.lieber@loria.fr jean-guy.mailly@u-paris.fr
marquis@cril.univ-artois.fr prade@irit.fr francoisrollin3@icloud.com

La bêtise est souvent l'ornement de la beauté ; c'est elle qui donne aux yeux cette limpidité morne des étangs noirâtres et ce calme huileux des heures tropicales.
Charles Baudelaire, *Journaux intimes* (1887)

Résumé

Le professeur Rollin a écrit au sujet de la bêtise artificielle. Pouvions-nous rester indifférent à cela ? Cet article montre le contraire.

Abstract

Professor Rollin wrote about artificial stupidity. Could we remain indifferent to this ? This article shows the contrary.

1 Introduction

Dans un ouvrage paru récemment et consacré à la notion de bêtise, François Rollin consacre un chapitre intitulé « BA ? » et consacré à la *bêtise artificielle* [24]. Si on voit la bêtise comme une « intelligence en creux », cela suggère des liens forts entre BA et IA, d'où l'intérêt pour la communauté de l'IA de se pencher sur cette question.

Mais comment pourrait-on définir la BA ?

La thèse que François Rollin développe est, en substance, qu'alors qu'un système d'IA doit effectuer des opérations de tri dans les données et d'adaptation sur ces données, un système de BA s'abstiendra d'un de ces deux types d'opérations. Il en conclut qu'un système de BA pourrait être un système associant à une question une réponse sans lien avec elle et que la mise en place d'un système de BA serait facile. Il illustre cette idée par des exemples, dont ceux-ci :

> Comment résoudre une équation du second degré ? Réponse de la machine : en faisant manger à tous les centenaires de la région Hauts-de-France six kilos de cœur d'artichaut par jour pendant huit semaines.
>
> Combien d'étoiles y a-t-il au total dans l'Univers à 20% près ? Réponse : 8.
>
> *(extraits cités avec l'aimable autorisation de l'auteur)*

Cette thèse nous a interpellés et cet article présente quelques réflexions à ce sujet. La bêtise est un sujet peut-être encore plus fascinant que l'intelligence, au vu du nombre considérable d'ouvrages célébrant, interrogeant, dénonçant la bêtise. Citons-en tout de même quelques-uns dans différents registres [14, 19, 16, 3, 12, 7].

Cet article est organisé comme suit. La section 2 tente de cerner la notion de bêtise. Puis, quelques bêtises faites par des êtres humains ou des machines seront discutées (section 3). On s'intéressera ensuite à la notion de système de BA. À son sujet, les deux questions suivantes peuvent se poser à des spécialistes de l'IA :

*Les auteurs remercient les rapporteurs anonymes de cet article pour leurs retours encourageants ou critiques (ou les deux). Ils se sont efforcés d'en tenir compte pour la version finale, en espérant n'avoir pas mal interprété ces remarques et tout en sachant que le thème de l'article leur fournissait une excuse toute trouvée le cas échéant. Ils ont également apprécié les suggestions faites par ces rapporteurs anonymes, même quand ils n'ont pas pu les intégrer pleinement dans cet article.

— Pourquoi construire (sciemment) un système de BA ?
— Comment construire un tel système ?

Les sections 4 et 5 abordent ces deux questions. Après une conclusion provisoire (section 6), les quatre premiers auteurs laissent la parole au cinquième : François Rollin *himself*.

2 Cerner la notion de bêtise

Une approche possible pour cerner la « bêtise artificielle » est de procéder par une mise en contraste avec « l'intelligence artificielle (IA) ».

2.1 Un point de vue anthropocentré

Il s'agit alors de commencer par préciser ce qu'il faut comprendre par « intelligence artificielle ». Et dans une telle tentative, on se heurte immédiatement à la question épineuse de devoir définir ce que recouvre « l'intelligence ». Pour ce faire, on contourne le plus souvent la difficulté en faisant *référence à l'humain*, qui présente la forme d'intelligence sans doute la plus aboutie, en tout cas la plus polymorphe.

Le test de Turing [28] va dans le sens d'un tel jeu d'imitation : un processus est vu comme « intelligent » quand celui qui l'observe ne peut pas prédire s'il est le fait d'un être humain ou d'une machine. Ainsi, l'attribut « artificiel » ne pose pas de souci spécifique dans l'élaboration d'une définition (« artificiel » signifie seulement que le processus « intelligent » qui est analysé est produit par une machine). Seule « l'intelligence » pose question.

On notera au passage que la comparaison à l'humain pour tenter de s'en sortir et caractériser « l'intelligence » par une opposition pré-existait au test de Turing. Ainsi, pour René Descartes [8, cinquième partie], « l'intelligence » (au sens de la raison) est l'apanage de l'humain, c'est ce qui sépare l'Homme de l'animal. Le bêtise serait donc *a contrario* le propre de l'animal (ce qui explique l'étymologie du mot « bêtise » et les nombreuses références à l'animal quand il s'agit de parler d'âneries...). La thèse de Descartes est évidemment discutable : la raison n'est qu'une forme d'intelligence parmi d'autres et, d'un point de vue biologique, l'être humain est un animal parmi d'autres. Aujourd'hui, dans le champ disciplinaire qu'est l'IA, quand on met en avant des algorithmes dits bio-inspirés mettant en œuvre une forme d'intelligence collective, on mentionne le plus souvent des sociétés animales (comme c'est le cas pour les colonies de fourmis), mais pas des sociétés humaines. Est-ce parce qu'elles présentent des comportements trop complexes pour être modélisées de façon satisfaisante ou parce que la valeur ajoutée par le collectif est finalement trop limitée pour être notable ? Chacun peut avoir un point de vue sur la question.

2.2 Être bête

La bêtise (au sens de la capacité à être bête ou plutôt de l'incapacité à ne pas être bête) a souvent été définie en creux comme une déficience des capacités intellectuelles (en particulier, comme le veulent cette fois l'étymologie de l'« intelligence » et les tests de QI, le manque d'aptitudes à établir des liens). Sans surprise, la référence au développement intellectuel chez l'Homme a servi de repère pour établir au XIX^e^ siècle une gradation de la bêtise [13], où l'idiot est celui qui ne peut pas communiquer par la parole, l'imbécile celui qui est incapable de lire et d'écrire et le sot, celui qui ne raisonne ni ne se comporte de manière normale. Comme l'écrivait Eugène Marbeau : « La bêtise ne comprend pas ; la sottise comprend de travers » (*in* Livre d'or de la Comtesse Diane, 1886).

Pour qu'un agent (qu'il soit humain ou artificiel) puisse être qualifié de bête, il ne suffit évidemment pas qu'il produise une bêtise de temps en temps (à ce compte-là, tous les humains seraient stupides), mais que l'agent concerné *persiste dans ses erreurs*, même quand une interaction soutenue avec son environnement lui fournit suffisamment d'arguments valables pour qu'en principe, les erreurs ne soient pas reproduites. Les agents artificiels ont à ce titre une prédisposition plus grande à la bêtise que les agents humains, simplement parce que leur interaction avec leur environnement est typiquement beaucoup plus réduite, voire inexistante (les agents artificiels ne sont pas tous munis de capteurs et d'effecteurs) mais aussi parce que leur force est précisément de pouvoir répéter sans se lasser (et très rapidement) les mêmes traitements et en conséquence de répéter les mêmes erreurs. On peut noter que cette idée de la bêtise liée à la persistance de l'erreur fait partie des points étudiés en détail dans l'ouvrage de François Rollin cité en introduction [24], même s'il n'est guère mis en avant dans le chapitre sur la BA.

La bêtise prenant des formes très variées, il est sans doute illusoire de vouloir caractériser précisément ce qu'est être bête. On peut néanmoins retenir cette idée de persistance dans l'erreur comme un de ses traits marquants. L'immutabilité des croyances nous sort de la condition humaine, et de sa capacité à savoir en interrogeant ses savoirs antérieurs pour les remettre en question chaque fois que nécessaire. Elle nous rapproche une fois de plus de l'animal, en particulier, de l'autruche (dont la politique pourrait être de ne rien voir pour ne rien savoir) voire du sur-humain comme le laisse entendre la célèbre citation latine *errare humanum est, perseverare diabolicum*. Pour laisser le dernier mot à Eugène Marbeau : « La sottise se croit très habile, et ne doute de rien. » (*Pensées et maximes diverses*, 1906).

2.3 Faire des bêtises

Définir ce qu'est une bêtise est sans doute beaucoup plus simple que de définir ce qu'est être bête. Lançons-nous donc dans une tentative de cerner les bêtises.

Une bêtise peut être vue comme le résultat d'un processus « intelligent » qui a mal tourné, qu'il s'agisse d'un processus de raisonnement ou de prise de décision ; ce résultat est simplement faux ou inadapté, parce qu'il s'appuie sur des informations qui le sont tout autant ou parce qu'il est engendré en utilisant des règles de raisonnement fallacieuses ou en passant par des chemins de traverses, ou encore parce qu'il s'appuie sur un modèle décisionnel irrationnel. Supposer au départ l'existence d'un processus « intelligent » qui aurait pu convenir permet de séparer les bêtises des absurdités, qui, elles, seraient produites par des processus totalement non pertinents. Ainsi, à la question « Combien font 12345 fois 67890 ? », une réponse comme « Stéphanie de Monaco » ne serait pas juste bête, mais bien absurde. Et un agent qui répondrait systématiquement « Stéphanie de Monaco » à toute question posée (sauf éventuellement quand la bonne réponse pourrait être « Stéphanie de Monaco »...) [1] serait assurément jugé stupide.

Pour illustrer la production de bêtises comme résultat d'un processus « intelligent » inadapté, considérons les deux exemples suivants (l'un lié au raisonnement, l'autre à la prise de décision) :

> *Tous les chats sont mortels, Socrate est mortel, donc Socrate est un chat.*
>
> Eugène Ionesco, *Rhinocéros*.

Ici, les informations utilisées sont correctes, c'est l'inférence qui ne l'est pas : en l'espèce, elle est abductive, pas déductive, et donc peut aboutir à des conclusions erronées (comme c'est le cas ici).

Voici un autre exemple (un dialogue entre deux amis) :

— *J'ai lu dans un magazine que ma consommation d'alcool (une bouteille de Bordeaux par jour) est excessive et pourrait nuire gravement à ma santé.*
— *Sans doute. Qu'as-tu fait alors ?*
— *J'ai arrêté de lire.*

Dans ce cas, la décision prise est inattendue et sans doute inadaptée.

Décider du caractère erroné ou inadéquat du résultat requiert une *comparaison à une norme*. Selon une certaine « distance » à un point de référence, le résultat en question pourra être jugé approprié ou non, et le processus l'ayant produit sera considéré comme (plus ou moins) « intelligent » ou « bête ». La distance en question peut alors susciter chez celui ou celle qui la mesure des sentiments variés, du courroux, du rire, voire de l'émerveillement.

Notons que la norme employée pour mesurer l'adéquation du résultat produit par le processus de raisonnement ou de prise de décision considéré peut être *universelle* (la logique mathématique permet, par exemple, de séparer les raisonnements valides de ceux qui ne le sont pas) ou ne pas l'être (il n'existe pas de modèle unique et consensuel pour caractériser ce qu'est une prise de décision raisonnable). Cette norme est surtout souvent *relative à un contexte*, en particulier à la personne qui reçoit la bêtise et décide de la classer comme telle ; et cela est fonction de ses aptitudes propres, ce qui fait que ce qui est bêtise pour un individu donné ne l'est pas forcément pour un autre. Prenons un exemple pour lequel une norme universelle existe : il s'agit encore de multiplier 12345 par 67890. Il n'y a qu'un seul vrai résultat possible : le produit recherché vaut 838102050. Le résultat -1 produit par l'individu A pourra être considéré comme stupide par toute personne ayant un petit bagage mathématique, le résultat 838102051 produit par l'individu B, tout aussi erroné, sera peut-être considéré comme une erreur de calcul par certains et comme une bêtise par d'autres (on pourra observer que comme 67890 finit par un 0, le produit recherché doit lui aussi finir par un 0), le résultat 838012050 produit par l'individu C, tout aussi faux, sera sans doute perçu comme moins bête. C'est cette « distance » au résultat escompté (peut-être incorrect lui aussi) qui fait qu'une simple erreur acquiert le statut de bêtise.

On notera au passage que, pour ce genre de tâches, les humains, à cause de leurs limitations cognitives, sont souvent beaucoup plus bêtes que des machines élémentaires comme des calculettes. La capacité des humains à raisonner correctement avec des nombres (et, en particulier, avec des probabilités) est réduite et connue comme telle depuis longtemps. Dans son ouvrage [20], le mathématicien américain John Allen Paulos a introduit le néologisme « *innumeracy* » pour désigner cette incapacité (l'« *innumeracy* » est aux nombres ce que l'« *illiteracy* » – en français, l'analphabétisme – est aux lettres). La période pandémique que nous traversons depuis deux ans a conduit à un florilège d'erreurs de raisonnement, comme celle de refuser la vaccination par le fait (avéré) qu'il y a plus de cas graves de la COVID parmi les vaccinés que parmi les non-vaccinés, mais vu comme un argument (fallacieux) qui soutiendrait le fait que la probabilité de faire une COVID grave en étant vacciné serait supérieur à celui de faire un COVID grave en ne l'étant pas (évidemment, l'erreur vient du fait que le nombre de personnes vaccinées et de personnes non vaccinées n'est pas le même !). De telles erreurs de raisonnement peuvent malheureusement conduire à des issues tragiques, comme cela se produit dans d'autres cadres tel le cadre juridique (voir le passionnant ouvrage [26]).

Plus généralement, il est bien connu en théorie de la décision que des systèmes de postulats de *rationalité* qui pouvaient sembler raisonnables pour décrire le comportement d'un décideur, ont rencontré des situations où l'application du critère de décision qui en découlait ne correspondait pas

1. Comme le personnage de Véronique dans le sketch « Télémagouilles » des Inconnus.

à la conduite de la majorité des décideurs [15].

Evidemment, si les machines dépassent largement les humains quand il s'agit de faire des calculs, leur intelligence est beaucoup moins polymorphe : elles ne savent rien faire d'autre ! Si intelligence il y a, c'est surtout celle des humains qui les conçoivent et les programment qu'il faut mettre en avant. En particulier, une calculette n'a pas plus conscience des opérations qu'elle réalise et de ce qu'est un nombre, qu'une horloge comtoise n'a du temps qui s'égrène, ou qu'un programme de reconnaissance faciale (voir à ce sujet la section 3.3) n'a d'un visage et de ce qui le constitue.

On pourra enfin garder en tête que les humains peuvent parfaitement percevoir comme bêtes des processus de raisonnement qui ne le sont pas du tout mais au contraire collent parfaitement à la norme universelle quand elle existe. Ainsi, la validité d'un raisonnement ne dépend que de la capacité de sa conclusion à être vraie dans tous les cas de figure où ses prémisses le sont, et absolument pas de la vérité des prémisses et conclusion en présence. En particulier, quand les prémisses sont fausses, toutes les conclusions s'ensuivent. Pour reprendre un exemple célèbre et paraphraser le grand logicien Bertrand Russell : « Si $2+2=5$ alors je suis le Pape. » En effet, on sait que l'on peut retirer des deux membres d'une équation comme $2+2=5$ la même quantité, disons 3, en préservant la validité de l'équation. Ainsi, de l'hypothèse $2+2=5$, on déduit $1=2$. Or, le Pape et moi sommes deux. Mais si $1=2$, alors nous sommes la même personne : je suis le Pape. Ce raisonnement est valide, il respecte les canons de la logique classique et pour autant, il apparaît comme dénué de sens (en particulier, parce que ses prémisses et sa conclusion n'ont aucun lien).

3 La bêtise est la chose du monde la mieux partagée [2]

3.1 Un bel exemple de bêtise humaine

Une histoire qui circule concerne un problème de mathématiques simple qu'un étudiant aurait résolu de façon surprenante, créative, mais fausse (une belle bêtise). Le problème (un peu reformulé) était celui du calcul de

$$\lim_{x\to 0} \frac{1}{5x^2}$$

Il semblerait que l'étudiant se soit rappelé de la solution d'un exercice similaire :

$$\lim_{x\to 0} \frac{1}{8x^2} = +\infty$$

Afin de laisser à la lectrice / au lecteur la possibilité d'imaginer la solution de l'étudiant, nous l'avons mise en fin d'article, en annexe A : la suite de la section fait l'hypothèse que vous êtes allé(e) lire cette annexe.

Maintenant, on peut imaginer ce que cela supposerait comme compétences pour une IA de générer des réponses telles que celle-là (une classe de bêtises, plutôt qu'une bêtise individuelle qu'il est toujours possible d'afficher simplement) et de façon involontaire (i.e., un système d'IA qui ne soit pas conçu pour générer des bêtises, mais qui en génère malgré le concepteur). On imagine qu'un système capable de générer la bêtise de l'étudiant doit à la fois travailler sur une représentation d'expressions mathématiques (par une structure arborescente, par exemple) et sur une représentation visuelle, avec des liens entre ces représentations. De notre point de vue, la construction d'une telle IA n'aurait rien de triviale !

3.2 Des IA bêtes

Un programme d'IA – une IA – ne fait que ce pourquoi il – elle – est programmée. On conçoit facilement qu'un programme sophistiqué puisse faire des choses qu'on jugera remarquables, en matière de résolution de problèmes, ou de reconnaissance de formes, faisant par là montre d'« intelligence » (artificielle). Des programmes très simples peuvent cependant donner le change. C'est le cas en particulier du programme ELIZA [29] qui dès le milieu des années 1960, simulait un entretien avec un thérapeute en reformulant ce que disait le « patient » sous forme de questions et relançait le dialogue avec des phrases toutes faites faisant écho à ce que venait de dire le patient. Il est clair que quand ELIZA déclarait « Je comprends », faute d'être capable d'une réponse plus adaptée au contexte du dialogue, elle bluffait totalement l'interlocuteur, alors qu'il s'agissait d'un système des plus bêtes.

Cette question des IA qui peuvent sembler « réalistes » à leurs utilisateurs, tout en s'appuyant sur des techniques très simples (voire simplistes) s'est longtemps retrouvée dans de nombreux jeux vidéos. Comme précisé dans [31, Section 1.2.2], dans les premiers temps, l'IA des jeux vidéos était généralement fondée sur des scripts pré-établis. Cela limite grandement les comportement des personnages non joueurs (PNJ, c'est-à-dire les personnages du jeu contrôlé par la machine). Le résultat est que, malgré les progrès importants réalisés par l'IA dans le domaine des jeux vidéos, il y a encore aujourd'hui de nombreux joueurs se plaignant de la bêtise des PNJ [3], y compris parmi les joueurs professionnels qui identifient les phases scriptées dans le déroulement du jeu [4]. Une conséquence à cela est qu'il est possible, pour un joueur suffisamment expérimenté, d'anticiper le comportement à venir de l'IA.

2. Ce qui n'empêche pas le bon sens de l'être également.

3. https://tinyurl.com/52mnzf85

4. https://tinyurl.com/yckpy784

3.3 Des IA qui disent des bêtises (ou sont utilisées bêtement)

Le cas Dong Mingzhu, ou l'IA bête et méchante. Dong Mingzhu est une femme d'affaires chinoise, à la tête d'une entreprise importante dans le domaine de la climatisation. Pour promouvoir son entreprise, elle a réalisé une campagne de publicité. Son portrait s'est ainsi retrouvé affiché dans les transports publics, sur des bus urbains. Or, en République Populaire de Chine, la vidéo-surveillance s'est généralisée. Une multitude de caméras et des algorithmes d'IA de reconnaissance de visages permettent à la puissance publique de détecter celles et ceux de leurs citoyens qui commettent des délits ou simplement des incivilités, comme traverser la chaussée quand c'est aux véhicules de passer. Un système de crédit social est en place et quand une incivilité est repérée, la personne l'ayant commise se voit retirer des points (et à un moment, perd par exemple la possibilité d'utiliser les transports publics). Dong Mingzhu a très vite perdu ses points car chaque fois qu'un bus affichant son portait passait devant une caméra de surveillance, le système d'IA en place concluait, à tort, que Dong Mingzhu avait traversé la route alors qu'elle n'en avait pas le droit. Ce scénario illustre bien à la fois la très grande qualité des algorithmes de reconnaissance faciale utilisés (ils ne sont pas intrinsèquement bêtes) mais aussi l'immense bêtise produite ici par le système d'IA utilisé, due à l'absence totale de contextualisation de la prise de décision.

Le traitement automatique de la langue (qui fourche). Pouvoir réaliser automatiquement (par algorithme) et de façon satisfaisante des tâches de traitement de la langue, comme les dialogues à base d'agents conversationnels animés (ou autres *bots*), le résumé ou la traduction d'une langue à une autre requiert une aptitude à comprendre, qui n'est pas acquise par les systèmes d'IA qui restent dénués de connaissances de bon sens sur le monde dans lequel on vit. Les outils de traduction automatique disponibles en ligne et que vous connaissez tous sont de formidables outils (les meilleurs disponibles) pour réaliser des traductions « superficielles ». Ils ne sont pas du tout bêtes mais pour autant, ce sont des générateurs de bêtises dont on peut s'amuser.

La résolution d'anaphores (en particulier, la satisfaction des références pronominales) est un problème difficile qui, pour pouvoir être résolu correctement, demande de disposer (et de savoir raisonner sur) des connaissances du monde. Ce problème est à la base des schémas de Winograd [30], qui peuvent être employés comme variantes au test de Turing.

Essayez par exemple d'utiliser un outil de traduction automatique disponible en ligne pour traduire de l'anglais vers le français la phrase suivante :

`"An AI algorithm cannot produce such a sillyness since` **`it`** `is too comic."`

La traduction résultante (via un outil bien connu et largement utilisé) est :

« `Un algorithme d'IA ne peut pas produire une telle bêtise car` **`il`** `est trop comique.` »

Ici, l'erreur est sur la référence du pronom « *it* ». Au départ, c'est la bêtise qui se veut trop comique, au final c'est l'algorithme d'IA qui l'est. Ce qui n'est pas tout à fait faux s'il s'agit de l'algorithme utilisé pour réaliser cette traduction.

Voici un autre exemple. Essayez de traduire en anglais la phrase suivante (qui présente un argument dont la justification est assez discutable) avec un outil de traduction automatique :

`La bêtise précède souvent l'intelligence. C'est normal, « bêtise » commence par un « b » et « intelligence » par un « i ».`

Une fois traduite en anglais, cela donne :

`Stupidity often precedes intelligence. This is normal, "stupidity" begins with a "b" and "intelligence" with an "i".`

La justification de l'argument traduit est encore plus discutable... L'algorithme d'IA ne sépare pas ici les lectures *de re* et *de dicto*, en dépit des guillemets utilisés.

Les outils de traduction automatique fournissent aussi des résultats amusants quand ils s'essaient à corriger nos erreurs... en en produisant à la place. Ainsi, la suite de mots `Veille technologique.`
qui, une fois traduite automatiquement en anglais, devient : `Old technology.`

Statistiquement, les mots « vieille » et « technologie » devraient apparaître plus souvent ensemble que les mots « veille » et « technologique », d'où l'erreur produite. Elle est cocasse car typiquement, quand on se lance dans une « veille technologique », c'est parce qu'on est en quête d'une technologie qui serait tellement récente qu'on ne la connaît pas encore, et donc pas d'une technologie qui serait désuète.

Bien entendu, la traduction est une tâche ardue ; il n'y a en règle générale, pour un texte un peu complexe, pas de traduction idéale, parfaite et ainsi les bêtises qui sont produites ne reflètent pas une distance trop importante par rapport à une norme qui serait universelle [5], mais plutôt un écart à ce qui est généralement attendu. Parmi les difficultés rencontrées figurent les ambiguïtés inhérentes au langage naturel qui peuvent se manifester dans la syntaxe même des phrases. Ainsi, la phrase (un classique du genre) « La petite brise la glace » est ambiguë puisque sa signification n'est pas du tout la même selon que l'on considère que le

5. Le problème de l'absence de norme universelle se pose, bien entendu, aussi pour les systèmes de reconnaissance faciale. Il n'existe pas une image de référence du visage de Dong Mingzhu (ou de n'importe qui d'autre).

verbe est « brise » ou que le verbe est « glace ». Prise en dehors d'un contexte qui permettrait de supprimer une des deux hypothèses possibles, cette phrase n'est pas traduisible puisque son propre sens dans sa langue d'origine n'est pas déterminé. Il est amusant d'observer comment les outils de traduction automatique s'en sortent avec ce type de phrases. Pour celle-ci, le résultat est surprenant :
`The little icebreaker`
qui signifie littéralement « Le petit brise-glace ». On se retrouve immédiatement transportés en hiver sur les bords de la Volga !

On pourrait multiplier les exemples. Les algorithmes d'IA ne comprenant pas les textes qu'ils traitent, le sens caché éventuel de ces textes reste inaccessible, même quand leur prise en compte est indispensable, comme dans les jeux de mots, calembours ou autres contrepèteries. Ainsi, « l'art de décaler les sons » une fois traduite en anglais devient "*the art of shifting sounds*", perdant au passage son statut de contrepèterie. La phrase devient clairement moins drôle ! Les chansons à double écoute, comme « Mon père et ses verres » de Boby Lapointe ou « La jeune fille du métro » de Colette Renard sont évidemment hors de portée d'une traduction automatique qui préserverait la double écoute possible.

4 Pourquoi construire un système de BA ?

La construction effective d'un système de BA peut être vue comme un exercice assez vain. Néanmoins, cette section réunit quelques raisons pour lesquelles on pourrait vouloir se lancer dans une telle aventure.

4.1 Dans l'optique de la réussite du test de Turing

Comme on l'a vu, le test de Turing vise à vérifier qu'un système automatique pourrait tromper des humains en leur faisant croire qu'il est humain lui-même, sous certaines conditions d'interaction [28]. Si un tel système répond trop intelligemment à une question, il risque d'échouer à ce test. Pour reprendre un exemple de la section 2.3, si un interlocuteur-machine répondant à la question « $12345 \times 67890 = \ ?$ » très rapidement et de façon correcte, il devrait échouer dans sa tromperie. Une réponse suffisamment éloignée du résultat sans être complètement absurde pourrait être une bêtise qui tromperait l'examinateur humain [6].

Ainsi, avec le but de la tromperie qu'induit le test de Turing, si l'on veut faire un système informatique réalisant ce test, celui-ci devrait relever de l'IA *et* de la BA. Et si on accepte le test de Turing comme une caractérisation de l'intelligence, cela conduit à accepter la bêtise comme une part de l'intelligence.

6. Une autre réponse de la machine pourrait être : « Tu ne serais pas un ordinateur qui essaie de passer son test de Turing, toi, des fois ? »

4.2 Bêtises délibérées

On pourrait imaginer un système d'IA qui produirait des bêtises de manière délibérée, peut-être dans l'esprit de [17]. On peut imaginer au moins deux raisons à cela :

— pour détendre l'atmosphère, pour faire rire l'interlocuteur, s'il s'agit d'un robot de compagnie.
— pour tromper, ou pour décontenancer un interlocuteur qui, par exemple, s'énerverait.

Il s'agit évidemment de défis très durs pour l'IA ! La compréhension et la génération automatique d'histoires drôles sont des problèmes très compliqués [23, 10].

4.3 La bêtise créative

Si séparer le vrai du faux est le Graal de la démarche scientifique et est indispensable « pour marcher avec assurance en cette vie » (pour paraphraser René Descartes), il ne faut pas avoir peur de se tromper, de faire des bêtises. La bêtise est féconde. C'est en en faisant que l'on apprend.

L'erreur est, en effet, source de création. L'Histoire regorge de scénarios où des maladresses ont conduit à des réussites, que l'on pense aux bêtises de Cambrai, à la tarte Tatin, ou pour quitter le domaine culinaire tout en restant dans la sérendipité, à la découverte du procédé de la vulcanisation par Charles Goodyear au XIX^{e} siècle. Bien sûr, pour qu'une erreur soit fructueuse, il faut ensuite de l'intelligence pour reconnaître l'existence d'un potentiel à exploiter dans la situation créée par l'erreur. Ainsi, une erreur dans une preuve mathématique se révélera féconde si sa découverte peut conduire à une approche nouvelle du problème.

Chez l'humain, depuis la petite enfance, l'apprentissage passe ainsi par des phases d'essais et erreurs, par des généralisations abusives que l'éducation, l'enseignement, la raison permettent (parfois) de corriger. L'apprentissage par renforcement implémente cette idée en IA : elle passe par l'exploration de situations qui sont jugées insatisfaisantes ou non. Observons que pour être créatif, un apprentissage par renforcement devrait sans doute autoriser une certaine permissivité dans l'estimation du caractère insatisfaisant d'une situation.

Il en va de même chez la machine : la généralisation inductive, tout comme l'abduction ou l'analogie ou encore d'autres formes de raisonnement de sens commun, ne préserve pas la vérité. Un apprentissage, quand il ne se limite pas à un apprentissage par cœur (qui, lui, n'offre pas la capacité de prendre en compte des situations nouvelles) présuppose le risque de se tromper. Les algorithmes mis en place dans les machines pour apprendre visent à minimiser ce risque de façon empirique, en s'appuyant sur les situations disponibles. Ils ne peuvent pas l'éliminer totalement. Chez la machine, il faut noter que l'erreur commise peut être importante quand les données sont trop peu massives ou de mauvaise qualité. C'est la question des biais dans

les données utilisées pour apprendre, qui peuvent conduire des IA à reproduire des comportements répréhensibles [7] ou à générer des prédictions qui ne sont rien d'autres que des prophéties auto-réalisatrices, comme dans le cas du prédicteur *PredPol*, utilisé par le département de police de Los Angeles pour déterminer les zones à patrouiller ; évidemment, le LAPD pouvait observer plus de crimes, délits et infractions là où *PredPol* l'envoyait patrouiller que là où elle n'était pas (et n'était donc pas en situation de les observer) !

Pour que la bêtise soit créative, chez l'humain comme chez la machine, le risque de se tromper doit être accepté, mais il doit aussi être modulé et contrôlé par la mise en œuvre de mécanismes permettant de remettre en cause ses croyances et les décisions auxquelles elles conduisent. C'est indispensable pour éviter le *perseverare diabolicum* déjà mentionné et s'efforcer ainsi de ne pas rester bêtes. Ainsi, de nombreux travaux en IA se sont-ils tournés depuis une quarantaine d'années vers cette problématique à facettes multiples, incluant le raisonnement non monotone (où les conclusions changent selon les hypothèses faites, qu'elles soient implicites ou pas) ou encore la révision de croyances (grosso modo, comment remettre en cause ses croyances pour en changer suffisamment sans en changer trop ?).

5 Concevoir un système de BA

Supposons que l'on veuille construire un système de BA. Pour ce faire, nous considérons d'une part les systèmes d'intelligence/bêtise artificielle fondées sur des raisonnements et d'autre part des systèmes d'[I/B]A fondés sur l'apprentissage supervisé, étant bien entendu que cela ne couvre pas toute l'IA ni toute la BA, et que cela ne constitue pas des ensembles de systèmes disjoints : en particulier, le raisonnement à partir de cas (RàPC), peut être considéré à l'intersection des deux. Le RàPC et l'argumentation sont d'ailleurs les deux problématiques plus spécifiques étudiées en fin de section.

5.1 Raisonner bêtement

Commettre des erreurs de raisonnement, c'est sans doute raisonner bêtement (même si c'est à la portée de tout le monde). Le raisonnement déductif repose largement sur le schéma suivant $\frac{p \to q \quad p}{q}$ (si p est vrai, et si p implique q, alors q est vrai). Autoriser d'autres schémas, comme par exemple $\frac{p \to q \quad q}{p}$ ou $\frac{p \to q \quad p}{\neg q}$ est plus risqué ! Selon le premier, qui n'offre pas les garanties de la déduction, l'observation d'une conséquence de p suggère que p est vrai, ce qui n'est pas bête ! Tandis que le second schéma mène à coup sûr à l'incohérence. On pourra consulter [5] pour une étude des arguments fallacieux. Une mauvaise analogie, reposant sur un parallèle trop hasardeux, mène de même à l'absurdité : « Le hérisson est comme la brosse à dents, il a le poil raide. »

Utiliser improprement la déduction conduit aussi à des absurdités :

— *Les appartements bon marché sont rares.*
— *Tout ce qui est rare est cher.*
— *Donc les appartements bon marché sont chers.*

Le problème provient ici de ce que le second énoncé a des exceptions (comme l'indique d'ailleurs le premier), alors qu'il est traité comme s'il n'en avait pas. La vraie bêtise peut alors être de ne pas démordre d'une conclusion fausse ou absurde.

Si un système de BA veut paraître bête, produire des conclusions fausses ne suffit pas ! Il doit rendre perceptible la manière dont il les obtient. Le comble du raisonnement bête est peut-être d'arriver à une conclusion exacte par des inférences fausses... à moins que cela ne soit un signe de virtuosité intellectuelle !

De façon similaire, un processus de raisonnement parfaitement correct pourra être perçu comme bête simplement parce qu'il n'est pas astucieux, qu'il est trop complexe, qu'il passe par des calculs compliqués alors qu'il existe des approches beaucoup plus simples (à titre d'exemple, on pensera au fameux problème des cyclistes et de la mouche, cf. `https://membres-ljk.imag.fr/Bernard.Ycart/mel/sn/node16.html`).

Enfin, si la bêtise est d'abord, comme il a été dit, une incapacité à voir, à créer des liens entre des faits ou des énoncés de ces faits, alors, on peut utiliser des idées de systèmes à capacité de raisonnement limitée [11] pour concevoir un système de BA. Par exemple, le système peut être « conscient » que p et $p \to q$ sont vrais, sans pouvoir former leur conjonction et conclure q [22].

5.2 Apprendre des bêtises

Un système d'apprentissage supervisé en IA s'appuie sur des exemples sous la forme de couples $(\mathtt{x}, \mathtt{y}) \in \mathcal{P} \times \mathcal{S}$ où $\mathcal{P}$ est l'espace des problèmes (ou des entrées) et $\mathcal{S}$ est l'espace des solutions (ou des sorties), tel que $\mathtt{x}$ a pour solution $\mathtt{y}$. Pour simplifier la discussion qui suit, on supposera que la relation « a pour solution » est fonctionnelle et on notera $\mathtt{F}$ cette fonction : les exemples non bruités $(\mathtt{x}, \mathtt{y})$ d'un système d'IA d'apprentissage supervisé vérifient $\mathtt{y} = \mathtt{F}(\mathtt{x})$. L'objectif d'un tel système est donc d'approcher la fonction $\mathtt{F}$: si on note $\mathtt{EA}$ l'ensemble d'apprentissage et $\mathtt{F}_{\mathtt{EA}}$ une fonction apprise par un système d'apprentissage supervisé donné, l'erreur sera mesurée par une estimation de la différence entre les fonctions $\mathtt{F}$ et $\mathtt{F}_{\mathtt{EA}}$.

Pour construire un système d'apprentissage supervisé en BA, une idée naturelle est d'appliquer une technique d'ap-

7. Voir par exemple `https://tinyurl.com/yvaj4beh`.

prentissage supervisée d'IA qui s'appuie sur des exemples de bêtises. La question à laquelle il faut répondre alors est celle de la constitution de $\mathtt{EAB}$ (l'ensemble d'apprentissage bête). Partant de l'idée qu'une condition nécessaire pour que $(\mathtt{x}, \mathtt{y})$ soit un exemple de bêtise est que $(\mathtt{x}, \mathtt{y})$ soit une erreur, il faut donc que $\mathtt{y} \neq \mathtt{F}(\mathtt{x})$. Cette condition nécessaire est-elle suffisante pour faire un apprentissage bête ? Si non, quelle autre condition serait nécessaire ? Voilà deux questions auxquelles nous n'apporterons pas de réponse ici : nous nous contenterons de la condition nécessaire et allons examiner ce qu'implique le fait de la considérer comme suffisante dans quelques cadres.

Considérons dans un premier temps la classification binaire (i.e., $|\mathcal{S}| = 2$) et notons $\overline{\mathtt{y}}$ l'élément de $\mathcal{S}$ différent de $\mathtt{y} \in \mathcal{S}$. On peut donc constituer $\mathtt{EAB}$ en partant d'un ensemble d'apprentissage $\mathtt{EA}$ (destiné à une IA) : $\mathtt{EAB}$ serait alors l'ensemble des $(\mathtt{x}, \overline{\mathtt{y}})$ où $(\mathtt{x}, \mathtt{y}) \in \mathtt{EA}$. L'apprentissage mènerait alors à une fonction $\mathtt{F}_{\mathtt{EAB}}$ qui, si l'apprentissage est réussi, se tromperait la plupart du temps, ce qui signifie que la fonction $\overline{\mathtt{F}_{\mathtt{EAB}}} : \mathtt{x} \in \mathcal{P} \mapsto \overline{\mathtt{F}_{\mathtt{EAB}}(\mathtt{x})} \in \mathcal{S}$ donnerait une fonction d'apprentissage correct [8] : demander l'avis aux personnes qui sont bêtes dans ce sens-là est donc utile pour savoir ce qu'on ne devrait pas faire ! Par exemple, si on a un système de classification d'images permettant de distinguer les photos de rats-taupes glabres des photos de bicyclettes, un classifieur binaire efficace construit sur un ensemble d'apprentissage bête (au sens précédemment décrit), il prendra, la plupart du temps, une photo de rat-taupe glabre, pour celle d'un vélo et inversement.

Quand la taille de $\mathcal{S}$ augmente, les choses peuvent devenir différentes. Prenons le cas de la régression où $\mathcal{P} = \mathbb{R}^m$ et $\mathcal{S} = \mathbb{R}^n$, étant donné $(\mathtt{x}, \mathtt{y}) \in \mathtt{EA}$, comment choisir $(\mathtt{x}, \mathtt{y}') \in \mathtt{EAB}$ avec la seule condition $\mathtt{y}' \neq \mathtt{y}$? Si c'est par une bijection $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, on se ramène à une situation similaire à la précédente ($\varphi^{-1} \circ \mathtt{F}_{\mathtt{EAB}}$ pourrait être une aussi bonne approximation de $\mathtt{F}$ que $\mathtt{F}_{\mathtt{EA}}$). Si le calcul de $\mathtt{y}'$ se fait de façon aléatoire (et sans tenir compte de $\mathtt{y}$), alors $\mathtt{F}_{\mathtt{EAB}}$ ne serait représentatif que de la distribution de probabilité du générateur aléatoire sur $\mathcal{S}$: on retombe sur l'hypothèse rollinesque d'un système de BA qui répond n'importe comment à la question posée (cf. introduction). Si le calcul de $\mathtt{y}'$ se fait de façon aléatoire mais proche de $\mathtt{y}$ (p. ex., pour $n = 1$, $\mathtt{y}'$ choisi aléatoirement dans $[\mathtt{y} - C, \mathtt{y} + C] \setminus \{\mathtt{y}\}$, où $C > 0$ est une constante relativement « petite »), on se retrouve dans le cadre de l'apprentissage avec des exemples bruités, avec un bruit additif.

Ces quelques exemples abstraits illustrent le fait que la condition nécessaire « être bête, c'est se tromper » n'est pas toujours suffisante pour faire un système d'apprentissage en BA qui ne soit ni un système d'apprentissage en IA « en creux » ni un système qui se contente de répondre n'importe comment.

5.3 Le RàPB

Le RàPB serait le raisonnement à partir de cas (RàPC) produisant des bêtises. On rappelle que le RàPC consiste à (tenter de) résoudre un problème cible en s'appuyant sur une base de cas où un cas est une représentation d'un épisode de résolution de problèmes : un cas est souvent la donnée d'un couple $(\mathtt{x}, \mathtt{y})$ (i.e., ce qu'on appelle un exemple en apprentissage supervisé), parfois agrémenté par des explications sur des liens entre $\mathtt{x}$ et $\mathtt{y}$. Une session de RàPC consiste souvent en la sélection dans la base de cas d'un cas jugé similaire au problème cible (étape de remémoration) et en la modification de ce cas dans l'optique de la résolution du problème cible (étape d'adaptation).

Pour construire un système de RàPB, on pourrait s'inspirer du principe de l'apprentissage de bêtises évoqué ci-dessus, à savoir la constitution d'une base de cas-bêtises. Cela rencontre la problématique récente du RàPC utilisant des cas positifs et des cas négatifs (ces derniers peuvent être rapprochés — voire assimilés ? — à des cas-bêtises) et qui a donné de bons résultats pour l'apprentissage de connaissances d'adaptation [18], ce qui tend à dire que si on mémorise à la fois les expériences positives et les expériences de bêtises, en distinguant les deux, cela peut être utile pour mieux adapter en RàPC.

Une autre voie, éventuellement complémentaire, pour le RàPB, consisterait à changer les étapes du raisonnement (comme c'est le cas dans le cadre général de la section 5.1), par exemple en sélectionnant le cas le moins similaire au problème cible : si l'adaptation parvient à modifier le cas-recette du gratin dauphinois afin de résoudre le problème « J'aimerais une recette de sorbet au cassis. », le résultat pourrait être surprenant (et utile quand on a des invités qui s'attardent trop). On peut aussi imaginer une remémoration pertinente suivie d'une adaptation qui ne l'est pas : l'exemple de la section 3.1 illustre cela (pour un système de RàPC ou de RàPB résolvant des problèmes de calcul de limites).

On notera que ces idées rejoignent celle présentées par François Rollin d'un système de BA qui s'abstiendrait de trier (ici : de faire une remémoration bête) ou d'adapter (ici : de faire une adaptation bête).

5.4 Bêtise argumentée

En IA, l'argumentation formelle est le domaine qui s'intéresse à la représentation d'informations conflictuelles, et à la définition de méthodes de raisonnement non triviales à partir de ces informations, en se basant sur des notions de logique et de rhétorique : un argument est composé d'un ensemble d'informations considéré comme fiables (le support) et d'une information qui peut être déduite de ce support (la conclusion) ; si un argument a contredit un argu-

8. Sous certaines hypothèses de symétrie entre l'appréhension des deux classes par le classifieur, il devrait être possible de montrer que $\overline{\mathtt{F}_{\mathtt{EAB}}}$ et $\mathtt{F}_{\mathtt{EA}}$ ont la même espérance d'erreur.

ment b, mais que rien ne vient contredire a, alors on considère que a est acceptable et que b ne l'est pas. Différents principes gouvernent les modèles de raisonnement argumentatif, il semble donc relativement aisé de définir un système de bêtise argumentée (BArg), via un système d'argumentation qui faillirait à satisfaire certains de ces principes de base.

Prenons d'abord le cas de l'argumentation abstraite [9]. Dans ce contexte, on s'intéresse aux relations entre les arguments pour déterminer lesquels sont acceptables, sans se soucier de leur nature précise (en particulier, de leur structure logique interne). Un système d'argumentation abstrait est donc un graphe $\mathtt{F} = \langle \mathtt{A}, \mathtt{R} \rangle$, dont les noeuds $\mathtt{A}$ sont les arguments, et les arcs $\mathtt{R} \subseteq \mathtt{A} \times \mathtt{A}$ sont les attaques, c'est-à-dire la représentation de la notion de « contre-argument ». Le raisonnement à partir d'un tel graphe se fait au moyen du concept d'extension, c'est-à-dire d'ensemble d'arguments conjointement acceptable, représentant une solution potentielle du problème représenté par le graphe argumentatif. Il existe différentes façon de définir une extension, selon les propriétés attendues de cet ensemble d'arguments. Parmi les approches classiques, nous parlerons uniquement (à titre illustratif) de la sémantique stable : un ensemble d'arguments $\mathtt{E} \subseteq \mathtt{A}$ est une extension stable si et seulement si 1) c'est un ensemble sans conflit ($\forall a, b \in \mathtt{E}$, $(a, b) \notin \mathtt{R}$), et 2) c'est un ensemble qui attaque son complémentaire ($\forall a \in \mathtt{A} \setminus \mathtt{E}$, $\exists b \in \mathtt{E}$ tel que $(b, a) \in \mathtt{R}$). Il est donc possible de définir un système de BArg comme un système qui retourne des ensembles d'arguments qui attaquent leur complémentaire, mais ne satisfont pas le principe d'absence de conflit. Prenons l'exemple d'un scénario très simple, où le dialogue suivant [9] :
– *C'est Charles Lytton qui a volé la pierre précieuse !* (a)
– *C'est impossible, il a une jambe dans le plâtre.* (b)
– *Il feint probablement d'être blessé.* (c)
est modélisé par un système d'argumentation $\mathtt{F} = \langle \mathtt{A}, \mathtt{R} \rangle$, où $\mathtt{A} = \{a, b, c\}$ et $\mathtt{R} = \{(c, b), (b, a)\}$. Alors que la sémantique stable requiert d'accepter $\{a, c\}$, une version bête peut mener à l'acceptabilité de $\{b, c\}$ (l'ensemble attaque bien son complémentaire, mais n'est pas sans conflit). Des versions bêtes des autres sémantiques de raisonnement argumentatif peuvent être définies de façon similaire, en choisissant des ensembles d'arguments qui ne satisfont qu'une partie des propriétés requises. Notons tout de même que dans certains contextes, comme l'existence d'informations au sujet de priorités à appliquer entre les arguments, certaines propriétés de base comme l'absence de conflit peuvent être violées [1, 2, 25]. Cela veut-il dire que ces cadres étendus sont bêtes ? Nous laissons le lecteur en juger.

Si on ne s'intéresse pas uniquement aux relations entre les arguments, mais également à la façon de les former à partir de connaissances structurées (généralement avec un formalisme logique [4]), il existe également plusieurs façons d'argumenter bêtement. En résumé, il y a deux étapes dans la conception d'un système d'argumentation structuré à partir de connaissances logiques : l'identification des arguments, et l'identification des attaques entre eux. Ces deux étapes peuvent être sources d'importantes bêtises. Supposons qu'on ait affaire à un individu énonçant la phrase suivante : « J'aime le chocolat, donc je suis la reine d'Angleterre. ». Cela revient à considérer $(\{a\}, b)$ (où a représente le fait d'aimer le chocolat, et b le fait d'être la reine d'Angleterre) comme un argument, alors qu'il n'y a aucune raison rationnelle de supposer que b soit une conséquence de a. S'il est déjà un peu bête de prétendre au trône britannique en raison de son amour du chocolat, ce genre de raisonnement peut naturellement poser d'autres problèmes si on applique l'argumentation à des sujets plus sensibles. Même dans un cas où les arguments sont bien construits, une piste pour concevoir un système de BArg se situe au niveau de l'étape d'identification des attaques. Ignorer certaines attaques, ou au contraire en ajouter certaines qui ne devraient pas l'être, va naturellement changer l'issue du raisonnement.

On peut raisonnablement [10] se demander si un système de BArg « sans limite » ne ressemblerait pas à un raisonnement chaotique, sans intérêt particulier. Comme on l'a mentionné en introduction, la bêtise d'un résultat peut se définir en fonction d'une distance par rapport au résultat correct attendu. Un « bon » système de BArg serait alors un système qui identifie incorrectement un ratio pré-défini des arguments et des attaques, et qui sélectionne les arguments acceptables en maintenant également une certaine proximité avec une solution correcte.

5.5 D'autres BA

Chaque domaine et problématique de l'IA pourrait être examiné sous le prisme de la BA. Voici un court inventaire non exhaustif de tels exercices.

Pour la prise de décision, une question serait : « Comment décider bêtement ? » Par exemple, en décision multi-critères, on peut décider finalement à partir d'un critère non pertinent : acheter une voiture parce qu'elle est rouge alors qu'on avait décidé de l'acheter sur des critères écologiques, de prix et de performance, peut apparaître comme bête.

Pour l'apprentissage non supervisé, une question serait : « Comment catégoriser bêtement un domaine ? » À titre d'exemple, on peut séparer les humains en deux catégories, en fonction de la parité du nombre de voyelles dans le prénom de leurs grands-mères paternelles.

Dans le cadre des systèmes multi-agents, une question serait : « Comment faire émerger de la bêtise d'un groupe d'agents individuellement pas bêtes ? » Dans un

9. Toute ressemblance avec une comédie policière des années 1960 est absolument non fortuite.

10. Est-ce vraiment si raisonnable ?

tel contexte, on notera au passage l'identification de cinq lois fondamentales de la stupidité humaine, proposée par Carlo M. Cipolla [6]. Cette théorie pose, en particulier (la troisième loi énoncée par l'auteur), qu'un agent (humain) stupide est un agent qui nuit à un autre (ou à un groupe d'autres agents) sans en tirer aucun bénéfice, voire en étant nuisible à lui-même. En quelque sorte, tirer sur quelqu'un d'autre en se mettant une balle dans le pied ! Ces lois ont donné lieu à des expérimentations visant à estimer la validité de la théorie de Cipolla dans des contextes de simulation à base d'agents artificiels [27], mais aussi à des travaux visant à fournir une interprétation biophysique à la théorie [21].

6 Conclusion provisoire

Cet article est né d'une réflexion autour de la notion de BA qu'a introduite François Rollin dans un ouvrage récent. L'objectif poursuivi ici est de réunir quelques réflexions sur la question de la bêtise vue du point de vue de l'IA. L'article ne prétend pas couvrir toutes les problématiques que cette notion de bêtise artificielle suscite, mais pose les questions suivantes sur les liens entre ces notions encore mal définies d'intelligence [artificielle] et de bêtise [artificielle] : la seconde peut-elle être vue comme le complémentaire de la première ? ou, à l'inverse, considère-t-on qu'il ne peut y avoir de bêtise [artificielle] sans (un peu d')intelligence [artificielle] ? Une autre question ou, peut-être, une autre façon de poser cette question est celle de ce que serait un système de BA : un système d'IA quand il se trompe ? ou un système destiné à proposer des bêtises ? La deuxième réponse pourrait poser problème si l'on comprend la bêtise comme étant involontaire, mais on peut contourner cette difficulté en affirmant que la volonté (que le système soit bête) est celle du concepteur de ce système, alors que le système n'aurait pas plus conscience de dire des bêtises qu'un système d'IA n'a (jusqu'à preuve du contraire) de conscience d'être (parfois) intelligent. Enfin, nous ne nous sentons pas exemptés d'être bêtes par le simple fait d'avoir écrit cet article (qui contient au moins 7 bêtises : saurez-vous les découvrir ?), mais nous nous consolons en appréciant le charme, sinon la beauté de la bêtise.

Ce travail offre plusieurs perspectives d'études, en partie évoquées dans l'article. Une d'entre elles est la tentative de modélisation formelle de la bêtise. La bêtise ne se réduit pas à l'erreur, elle doit être considérée comme un écart vis-à-vis d'une norme qu'il conviendra de caractériser. Cette norme peut être celle d'un observateur jugeant s'il y a bêtise. Elle peut également être une estimation des conséquences de la bêtise (par exemple par une fonction d'utilité).

Cet article pose plus de problèmes qu'il n'en résout (en résout-il un seul ?). En fait, on peut le voir comme une invitation à la communauté de l'IA de considérer ce champ de recherche sous un angle inhabituel, avec l'idée que ce pas de côté pourrait se révéler déambulatoire [11]. Et n'oublions pas Gustave Flaubert qui a écrit « Oui, la bêtise consiste à vouloir conclure » (Lettre à Louis Bouilhet, 4 septembre 1850). Sachons, en toutes choses, nous garder des conclusions définitives !

Toutes les sections de cet article, à l'exception de la dernière, ont été rédigées par ses quatre premiers auteurs (qui ont néanmoins cité le cinquième en introduction). La dernière section est rédigée par le cinquième auteur.

7 Le mot de François Rollin

Des réflexions et questionnements exposés ci-dessus, je crains de devoir déduire qu'il est encore plus difficile de produire de la Bêtise Artificielle que de produire de l'Intelligence Artificielle, dès lors que l'objectif est de générer de l'authentique bêtise, de la bonne vraie bêtise bien de chez nous, et non pas simplement de l'absurdité ou du non-sens.

Cette conclusion serait contre-intuitive si on voyait la bêtise comme un simple ratage. De fait, il est plus facile de rater que de réussir : en cuisine, en musique, en littérature, en médecine, au jeu d'échecs, en amour, et j'en oublie, il faut une vraie compétence pour bien faire, tandis qu'une simple incompétence suffit à mal faire. En matière d'intelligence, il semble que ce soit le contraire, du moins si on veut produire une bonne vraie bêtise.

Cette surprenante difficulté ne peut qu'appuyer la nécessité et l'urgence de travaux savants et studieux sur la BA. L'aventure a démarré, c'est une bonne chose, il faut maintenant la mener à son terme, et ce n'est pas à moi que la tâche incombe : je suis bien trop bête pour ça.

A Réponse à l'exercice de la section 3.1

La solution proposée par l'étudiant était $+\backsim$.

Références

[1] Amgoud, L. et C. Cayrol: *A Reasoning Model Based on the Production of Acceptable Arguments*. Ann. Math. Artif. Intell., 34(1-3) :197–215, 2002.

[2] Atkinson, K. et T. J. M. Bench-Capon: *Value-based Argumentation*. FLAP, 8(6) :1543–1588, 2021.

[3] Bechtel, G. et J.-C. Carrière: *Dictionnaire de la Bêtise et des erreurs de jugement, suivi du Livre des Bizarres*. Robert Laffont, Bouquins, 1965, 1991, 2014.

11. Une légende raconte que le choix de cet adjectif a été proposé par un système de BA. L'origine de cette légende se perd dans la nuit d'étang (noirâtre ?).

[4] Besnard, P. et A. Hunter: *Elements of Argumentation*. MIT Press, 2008, ISBN 978-0-262-02643-7.

[5] Bisquert, P., F. Dupin de Saint-Cyr et P. Besnard: *Assessing arguments with schemes and fallacies*. Dans Balduccini, M., Y. Lierler et S. Woltran (rédacteurs) : *Proc. 15th Int. Conf. on Logic Programming and Nonmonotonic Reasoning (LPNMR'19), Philadelphia, June 3-7*, tome 11481 de *LNCS*, pages 61–74. Springer, 2019.

[6] Cipolla, C. M.: *The basic laws of human stupidity*. Bologna : il Mulino, 2011.

[7] Denis, P.: *Éloge de la Bêtise*. Presses Universitaires de France - PUF, 2001.

[8] Descartes, R.: *Discours de la méthode*. 1637.

[9] Dung, P. M.: *On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games*. Artif. Intell., 77(2) :321–358, 1995.

[10] Dupin de Saint-Cyr, F. et H. Prade: *La compréhension des histoires drôles : une affaire de révision de croyances*. Revue Ouverte d'Intelligence Artificielle, 2022.

[11] Fagin, R. et J. Y. Halpern: *Belief, Awareness, and Limited Reasoning.* Artif. Intell., 34(1) :39–76, 1987.

[12] Frankfurt, H.: *On Bullshit*. Princeton University Press, 2005. Traduction française : De l'Art de Dire des Conneries, 10/18, 2006.

[13] Godin, C.: *La bêtise existe-t-elle ?* Le Philosophoire, 42 :27–38, 2014. L'auteur développe son propos dans son Encyclopédie Conceptuelle et Thématique de la Philosophie (Editions Champ Vallon, 2018) au chapitre 96 sur l'intelligence qui comporte une section III sur la bêtise et une section IV sur l'intelligence artificielle.

[14] Jean-Paul: *Éloge de la Bêtise*. Corti, Domaine Romantique, 1782. Traduit de l'allemand par N. Briand, 1993.

[15] Kast, R.: *La Théorie de la Décision*. La Découverte, 1993.

[16] Latzarus, L.: *Éloge de la Bêtise* . Hachette, 1925.

[17] Lidén, L.: *Artificial stupidity : The art of intentional mistakes*. Dans Rabin, S. (rédacteur) : *AI Game Programming Wisdom*, tome 2, pages 41–48. Charles River Media Rockland, MA, 2003.

[18] Lieber, J. et E. Nauer: *Adaptation knowledge discovery using positive and negative cases*. Dans *Proc. of ICCBR 2021*, 2021.

[19] Musil, R.: *De la Bêtise*. Allia, 1937. Traduit de l'allemand par Ph. Jaccottet, 2000.

[20] Paulos, J. A.: *Innumeracy : Mathematical Illiteracy and Its Consequences*. Hill and Wang, 1988.

[21] Perissi, I. et U. Bardi: *The Sixth Law of Stupidity : A Biophysical Interpretation of Carlo Cipolla's Stupidity Laws*, mars 2021.

[22] Prade, H.: *Handling (un)awareness and related issues in possibilistic logic : A preliminary discussion*. Dans *Proc. of NMR 2006*, pages 219–225, 2006.

[23] Ritchie, G.: *The Comprehension of Jokes : A Cognitive Science Framework*. CRC Press, 2018.

[24] Rollin, F.: *Suis-je bête ! — L'héroïque Professeur Rollin foudroie la bêtise avec ruse et modestie*. Presses Universitaires de France, 2020. Illustrations de D. Gooseens, préface d'A. Astier.

[25] Rossit, J., J. G. Mailly, Y. Dimopoulos et P. Moraitis: *United we stand : Accruals in strength-based argumentation*. Argument Comput., 12(1) :87–113, 2021.

[26] Schneps, L. et C. Colmez: *Les Maths au tribunal. Quand les erreurs de calcul font les erreurs judiciaires*. Seuil, 1995.

[27] Tettamanzi, A. G. B. et C. da Costa Pereira: *Testing Carlo Cipolla's Laws of Human Stupidity with Agent-Based Modeling*. Dans *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, August 11-14, 2014 - Volume III*, pages 246–253. IEEE Computer Society, 2014. `https://doi.org/10.1109/WI-IAT.2014.174`.

[28] Turing, A.: *Computing Machinery and Intelligence*. Mind, LIX(236) :433–460, octobre 1950.

[29] Weizenbaum, J.: *ELIZA - A computer program for the study of natural language communication between man and machine*. Communications of the ACM, 9(1) :36–45, 1966.

[30] Winograd, T.: *Understanding natural language*. Cognitive psychology, 3(1) :1–191, 1972.

[31] Yannakakis, G. N. et J. Togelius: *Artificial Intelligence and Games*. Springer, 2018.