Publier ou Périr (ou Payer) : Les dérives de la publication académique et du monde de la recherche

La vie d'un chercheur dans le monde académique est rythmée par un processus relativement peu connu pour les non-initiés : la publication. Un chercheur passe en effet une partie de son temps à faire de la recherche (jusque là OK), mais consacre aussi énormément de temps à rédiger des papiers de recherche et à essayer de publier ses papiers dans des revues académiques. Si vous n'êtes pas dans ce petit monde de la recherche, vous n'avez sûrement jamais entendu parler du "Journal of Finance", de "Management Science" ou de l'"American Economic Review" (non non, publier dans "Capital", ça ne compte pas vraiment...). Mais si vous êtes chercheurs en économie / gestion/ finance, ces revues représentent pour vous le Graal. Un article dans un journal de top qualité, et hop votre carrière est lancée ! Pas de publication (ou bien uniquement dans des "petites revues"), et c'est la mort assurée. Le fameux "Publish or Perish" ! Bien que le système de publication académique ait sur le papier de nombreux avantages (évaluation de la qualité de la recherche par les pairs, structuration de la recherche...), la "course à la publication" peut malheureusement engendrer certaines dérives et faire oublier ce qui devrait être l'objectif numéro 1 : faire avancer la recherche.

Le fond du problème n'est pas spécifique à la recherche académique. En fait, à partir du moment où un indicateur quelconque, censé permettre de répondre à une problématique donnée, prend une place prépondérante dans un système, alors l'Homme peut avoir tendance à optimiser son travail pour satisfaire à cet indicateur en oubliant la problématique sous-jacente. Cela ne veut cependant pas dire que l'indicateur est mauvais ! Mais un indicateur doit être considéré avec toutes ses limites, et non pas comme une "mesure quantitative et 100% objective" permettant de répondre parfaitement à une problématique. Par exemple, si vous mettez en place une norme anti-pollution basée sur une mesure spécifique, alors les constructeurs automobiles vont optimiser ce paramètre pour répondre à un test donné (voire même tricher... le Captain' ne vise personne), et votre objectif initial de baisse de la pollution à long-terme ne sera pas forcément rempli. Et bien pour la recherche académique, c'est la même chose : à partir du moment où "la publication" devient au centre du système, l'objectif de "faire avancer" la recherche peut en partie disparaître.

Attention : il ne s'agit absolument pas de dire que tous les chercheurs ne pensent qu'à "publier pour publier", en réfléchissant uniquement à leurs petites carrières personnelles. Loin de là ! Mais à l'inverse, il ne faut pas non plus faire preuve d'angélisme : comme dans de nombreux secteurs, il existe des fraudes, des manipulations et des tricheries dans le monde académique. Le dernier scandale à ce sujet a eu lieu il y a quelques semaines semaines, avec l'identification d'un vaste schéma de fraude à la publication concernant 64 "faux papiers" de recherche (source : Washington Post "Major publisher retracts 64 scientific papers in fake peer review outbreak"). Selon un article publié en 2010 dans "Nature" (justement un "top-journal / Graal" du chercheur), "Publish or perish in China", un chercheur sur trois dans les grandes universités chinoises serait coupable de plagiat, de falsification ou de "fabrication de données". Un chercheur sur trois !!!

"In other studies, one in three researchers surveyed at major universities and research institutions admitted to committing plagiarism, falsification or fabrication of data [..] However, several sources revealed to Nature that roughly one-third of more than 6,000 surveyed across six top institutions admitted to plagiarism, falsification or fabrication." - Nature

Plus récemment, une enquête de "Science" (top-journal again) "China’s Publication Bazaar" a montré

l'existence d'un véritable "marché noir de la publication académique" en Chine, où des chercheurs payent plusieurs milliers d'euros pour ajouter leurs noms sur des papiers de recherche et ainsi être publié sans même connaître le sujet, et ce histoire de rajouter une ligne importante sur un CV. Pourquoi spécialement en Chine ? Et bien simplement car, face à une forte concurrence, les jeunes chercheurs chinois cherchent par tous les moyens à "percer dans le système". Et pour percer, il n'y a pas 10.000 solutions : il faut publier dans des top-revues ! De plus, de nombreuses universités chinoises offrent des primes colossales aux chercheurs publiant dans les meilleurs journaux : selon certains chiffres, un chercheur chinois recevrait une prime d'environ 30.000 dollars s'il publie un papier dans "Nature" ou "Science" ! Autant vous dire qu'étant donné le salaire moyen d'un chercheur en Chine, l'incitation financière n'est pas négligeable... Pour finir, les sanctions sont encore très "lights" : alors que l'on pourrait imaginer une interdiction de publier ou d'enseigner pendant plusieurs années et/ou des sanctions financières/pénales, certains cas de fraudes avérés en Chine se sont simplement terminés par un renvoi du chercheur. High-Reward / Low-Risk : et hop le système devient rapidement un peu bancal.

Pour diminuer le risque de "fabrication de données" (et de "fabrication de beaux résultats") et permettre une meilleure réplicabilité, certains journaux exigent que l'ensemble des données, programmes et scripts utilisés dans le cadre d'un papier de recherche soit publique. C'est le cas par exemple pour l'"American Economic Review" (source : "Data Availability Policy") ou "PLOS" (source : "Data Policy"). Mais c'est une pratique qui est (malheureusement) encore trop peu répandue. Avec le déluge de nouvelles données "big-data", le problème de "fabrication des données" sera de plus en plus présent : plus de données = plus de facilité à truquer ses datasets pour un chercheur et plus de difficulté lors du processus d'évaluation pour vérifier la véracité des résultats avant publication (temps et puissance de calcul, connaissance spécifique en big-data...).

De nombreuses études sont désormais basées sur des données privées ou payantes, ce qui implique par définition moins de transparence et une quasi-absence de réplicabilité. Par exemple, le Captain' travaille sur la thématique "Twitter et Marchés Financiers" (sentiment des investisseurs et détection d'évènements), et a créé un robot qui tourne en permanence depuis des mois pour aller extraire des données en temps-réel sur Twitter. La base de données de plusieurs millions de tweets est donc privée, et selon les Termes d'Utilisation de Twitter, le Captain' n'a pas le droit de transmettre ou de revendre cette base. Dans une configuration telle que celle-ci, le risque (sur le papier) de falsification de données est énorme. Autant vous le dire tout de suite, ayant un minimum d'éthique et aimant réussir à se regarder correctement dans sa glace le matin, le Captain' ne va bien évidemment pas "truquer ses données" ni aller consciemment vers "l'overfitting". Mais face à une pression extérieure croissante poussant à la "publication à tout prix" (ce qui n'est pas le cas pour le Captain', merci @LaSorbonne et @IESEG de me laisser le temps de mener ma recherche sérieusement et consciencieusement), la balance entre "éthique" et "si je publie pas je suis mort" ne penchera malheureusement pas pour tout le monde du bon côté.

Il y aura forcément, parmi l'énorme majorité de chercheurs "honnêtes", quelques tricheurs... et malheureusement les tricheurs risquent de gagner à ce petit jeu, car la probabilité de se faire "prendre" est très faible. Même si, comme dirait Patrick Bruel, "That's poker", il est important pour éviter autant que possible cela (1) que le processus de "peer-review" (évaluation par les pairs) avant publication soit irréprochable, (2) que, dès que cela est possible, les données soient disponibles afin de permettre une réplicabilité (et vérification) après publication par d'autres chercheurs, (3) que les sanctions en cas de fraude soient bien plus importantes (4) que les chercheurs déclarent qui a financé la recherche et s'ils ont été payés pour écrire ce papier ("Disclosure Policy" déjà imposée dans pas mal de journaux) et (5) (avis personnel) que des critères autres que la "publication académique pure" soient davantage pris en compte pour une évolution de carrière dans le monde académique, comme par exemple la participation au débat public, la présence dans les médias, la qualité de l'enseignement, la publication d'ouvrages, et même, soyons fous, la création d'un blog (le Captain' ne vise personne... ;)).

Conclusion : Le système de "publication à comité de lecture" et l'évaluation de la qualité de la recherche en fonction des publications a de nombreux avantages, et l'objectif de cet article n'est absolument pas de remettre en cause tout le système. Cependant, et étant donné les dérives et fraudes avérées (qui ne sont d'ailleurs sûrement que la partie émergé de l'iceberg), il est important de voir s'il est possible d'améliorer le système, même de manière infinitésimale, pour essayer de revenir vers l'objectif de base : "faire avancer la recherche". C'est un peu #Bisounours comme vision il est vrai, mais si chaque chercheur essayait par exemple d'améliorer la transparence et de faciliter la réplicabilité en mettant à disposition simplement et gratuitement l'ensemble des données utilisées lorsqu'un papier est publié, alors cela serait déjà une belle avancée. Allez pour la peine, le Captain' s'engage à publier toutes les données et indicateurs utilisés lors de sa thèse (lorsque cela est légal) sur son site de recherche http://www.thomas-renault.com (site en construction). Dictionnaire et scripts pour l'analyse de sentiment, liste d'experts au sein d'un réseau, séries temporelles de tweets (agrégé, car sinon pas légal)... Tout sera disponible gratuitement une fois mes papiers publiés. Sur ce, en route vers le Graal !