EPRIST

responsables IST des organismes de recherche

La recherche en crise de reproductibilité ?

Analyse I/IST n°30 - Avril 2020

La plupart des recherches scientifiques sont fausses. Telle est la conclusion provocatrice d'un article célèbre de *PLOS* initialement publié en 2005 : « Nos simulations montrent que dans la plupart des cadres expérimentaux il est moins probable qu'une conclusion scientifique soit vraie que fausse. » (Ioannidis 2005).

Depuis, ce constat préoccupant aurait été largement confirmé. Une enquête menée sur 1576 chercheurs par *Nature* en 2016 souligne que plus de 70% d'entre eux échouent à reproduire des expériences préexistantes et qu'ils admettent à plus de 90% qu'il existe une véritable « crise de la reproductibilité » (*reproducibility crisis*) (Baker 2016). Ce diagnostic est progressivement partagé par des disciplines scientifiques très variées telles que la médecine, la biologie, les sciences informatiques, les sciences sociales ou même tout récemment les études littéraires et l'histoire culturelle[1].

Cette note de synthèse documente un débat majeur de la recherche scientifique contemporaine. Au-delà de la description parfois simpliste d'une « crise générale », l'exigence nouvelle de reproductibilité amène à repenser radicalement les formes et les pratiques de publication.

Comment la science est devenue reproductible

La crise de la reproductibilité est généralement présentée comme une dérive récente (Drummond 2018). La course à la publication et la précarisation des conditions de recherche auraient débouché sur la production à la chaîne de travaux de faible qualité et invérifiables. L'idéal d'une science exacte et reproductible serait aujourd'hui bien émoussé.

Cette interprétation est aujourd'hui assez critiquée (Steinle 2016; Fanelli 2018). Si la reproductibilité des expériences est un lieu commun des discours épistémologiques depuis Newton et Robert Cooke, la pratique a toujours été très différente. Pendant la révolution scientifique du XVIIe siècle et toutes celles qui suivirent, les expériences ne sont quasiment jamais réitérées : « même de nos jours, la plupart — et sans doute l'immense majorité — des expériences ne sont pas soumises à des réplications (…) en dépit d'un idéal formel de réplicabilité, nous ne vivons pas dans une culture de la réplication (Steinle 2016, 56) ».

[1] Une controverse importante en humanités numériques a ainsi éclaté en avril 2019, suite à l'utilisation croissante de méthodes de classifications automatisées sur de grands corpus numérisés : https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986

Comment citer cette note :
Pierre Carl Langlais et EPRIST, La recherche en crise de reproductibilité ?, EPRIST Analyse I/IST n°30, Avril 2020, 10 p.

En effet, la reproduction est rarement une procédure simple[2]. Elle suppose d'avoir des indications précises sur le déroulement de l'expérience, le matériel utilisé et les méthodes de collecte et d'analyse des résultats, ainsi que les moyens de remobiliser ces ressources humaines et techniques. Même dans ces conditions, les observations peuvent être déformées par des facteurs contingents. De nouvelles distinctions terminologiques entre *réplication** et *reproduction**[3] ont été récemment proposées pour rendre compte de cette complexité (Plesser 2018). La réplication désignerait la répétition de l'expérience dans un cadre strictement identique (même laboratoire, même matériel, mêmes équipes). La reproduction correspondrait au renouvellement de l'expérience dans un nouveau cadre, avec nécessairement des évolutions plus ou moins sensibles. Ces taxinomies émergentes ne rendent que faiblement compte de la diversité des usages : rien que dans les sciences expérimentales, il existerait au moins 79 définitions possibles de la reproductibilité (Gómez, Juristo, and Vegas 2010).

Ce décalage entre la théorie scientifique et la science pratiquée est devenu problématique au cours du XXe siècle. D'après les moteurs de recherche académiques spécialisés, le concept moderne de reproductibilité apparaît au début des années 1900[4]. Par exemple, un article de 1902 *Transactions of the American Institute of Electrical Engineers* fait état de l'impact de différentes méthodes d'illumination sur la reproductibilité des expériences, notamment au regard de la préservation des couleurs (Bell 1902). Ces discours ne sont pas très originaux en eux-mêmes. Leur récurrence témoigne d'une inquiétude croissante face au changement d'échelle de l'activité scientifique et à la sophistication de l'instrumentation dans le contexte de la deuxième révolution industrielle.

Dans ces premiers débats, la reproductibilité ne fait alors allusion qu'au cadre matériel de l'expérience. La discussion va progressivement s'étendre aux procédures de traitement et d'analyse des données, avec l'émergence d'indicateurs statistiques normalisés. La *valeur p* s'impose à partir des années 1950 et 1960 comme le test de signification par excellence dans certaines disciplines (Hubbard and Lindsay 2008). Elle mesure théoriquement la probabilité qu'une série d'observations survienne par rapport à une situation considérée comme normale (usuellement qualifiée d'hypothèse nulle). La détermination du seuil de signification reste cependant une convention "culturelle" variable selon les disciplines : en biologie et en sciences sociales, la valeur 0,05 est de loin la plus courante ; les sciences physiques utilisent des valeurs considérablement plus basses.

Malgré leurs limites, les indicateurs statistiques normalisés facilitent les comparaisons à grande échelle. Il devient possible de réaliser de vastes « méta-analyses » compilant les données et les interprétations de plusieurs dizaines, voire plusieurs centaines de publications avec pour ambition de fixer les connaissances scientifiques actuelles sur un sujet donné. Dans ce contexte, la reproductibilité s'impose comme un problème systémique qui n'est plus limité à la réalisation d'une expérience unique. Dès 1966, une étude de 295 articles de médecine constate que des conclusions scientifiques sont fréquemment tirées d'analyses statistiques invalides (Schor and Karten 1966). Une reproductibilité insuffisante à l'échelle d'un champ de

[2] L'encyclopédie de philosophie de Stanford a publié un excellent article de synthèse sur ces enjeux épistémologiques : https://plato.stanford.edu/entries/scientific-reproducibility/

[3] Les termes signalés par un astérisque renvoient au *Glossaire*, en fin d'article.

[4] Nous avons mené une recherche rétroactive des termes "reproducibility" et "replicability" sur *Google Scholar* et *Dimensions* en tirant parti de la possibilité de restreindre la recherche à une période prédéfinie.

recherche risque d'entraîner des erreurs en cascade ou de déboucher sur des méta-résultats incohérents.
Dans ce cadre épistémologique, chaque publication dispose en quelque sorte d'un « droit de vote » affectant le consensus d'un champ de recherche. Par compensation, chaque publication est investie d'une responsabilité collective : les erreurs, de bonne foi ou non, peuvent avoir un effet contaminant. L'idéal de représentativité des recherches antérieures est également mis à mal par les politiques éditoriales des revues. Les *résultats négatifs** sont découragés de longue date, vu qu'ils ne représentent par définition aucune découverte nouvelle mais, au mieux, une anti-découverte. Ils constituent pourtant des ressources cruciales pour les méta-analyses : en leur absence « la littérature publiée reporte des preuves plus solides qu'elles ne le sont en réalité (Munafò et al. 2017, 3) ».
Les débats actuels correspondent à une troisième et dernière phase dans l'histoire longue de la reproductibilité : l'accès immédiat et sans restriction aux articles scientifiques sur Internet permet une évaluation rétroactive sans précédent des publications antérieures. Sur un strict plan technique, le repérage d'erreurs ou de déformations structurelles est considérablement facilité. La dissolution partielle des frontières disciplinaires (jusqu'alors confortées par les politiques d'abonnements préférentiels pratiquées par les bibliothèques) et, plus largement, l'intervention possible d'un public "initié" situé hors du monde scientifique contribue à démultiplier le champ des regards critiques possibles. La pratique de la méta-analyse prend la forme d'une « méta-science » ambitionnant d'évaluer les usages de champs scientifiques entiers et de chiffrer le taux de résultats vrais et faux ou reproductibles ou non. Plusieurs milliers d'études de ce type sont aujourd'hui menées chaque année (Ioannidis et al. 2015).
L'analyse des données et des résultats est de loin la principale activité scientifique visée par le débat sur la reproductibilité. Même si l'ouverture des données scientifiques reste encore aujourd'hui une pratique émergente, les indicateurs statistiques et les visualisations contiennent des indices tangibles, régulièrement scrutés sur les nouvelles plateformes d'évaluation comme *PubPeer*. Par contraste, les dispositifs matériels utilisés en laboratoire sont rarement rendus public.

Petits et grands arrangements avec les données

La plupart des problèmes de reproductibilité sont attribués à un large spectre de *pratiques discutables en recherche** (*Questionable Research Practice* ou *QRP*) qui va de biais involontaires (notamment dans la sélection des données préalables) à la fabrication complète des résultats.
À l'une des extrémités de ce spectre, l'accès facilité aux publications a permis de repérer des fraudes scientifiques caractérisées sur la durée. En octobre 2011, une enquête menée par trois étudiants de l'université de Tillburg annonce que le chercheur en psychologie Diederick Stapel, a publié plus d'une cinquantaine d'articles à partir de données fausses ou déformées : ces « bidouillages » ont passé le filtre de l'évaluation de 21 revues, dont plusieurs titres internationalement réputés[5]. Quelques années plus tard, une star montante de la recherche en biologie, Olivier Voinnet, est mise en cause par une nouvelle forme d'instance d'évaluation:

[5] Cf. le rapport final publié en 2012, https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf

le site participatif *PubPeer*[6]. Des contributeurs relèvent des anomalies dans les illustrations d'un article de 2004 : l'enquête est bientôt élargie et, à ce jour, 8 articles ont été rétractés et plus d'une vingtaine corrigés.

Des cas comme Stapel ou Voinnet sont a priori aussi rares que spectaculaires : le taux de rétractation reste faible au regard de la production scientifique mondiale et guère plus de 1 à 2% des publications reposeraient sur des données en partie fabriquées (Fanelli 2018). Cependant ces affaires révèlent aussi que la fraude s'inscrit dans un continuum de mauvaises pratiques. Avant de céder à la tentation de fabriquer des résultats de toute pièce, Stapel et Voinnet se livrent à des arrangements plus discrets qualifiés de pratiques « grises » ou d'« enjolivements ». L'autobiographie de Stapel, rédigée après la divulgation de l'affaire, décrit très précisément le processus graduel qui transforme un chercheur en faussaire :

Il y a des règles non écrites distinguant les pratiques douteuses acceptables de celles qui ne le sont pas, même si toutes deux aboutissent à une distorsion similaire des résultats (...). Tout ceci n'était pas plus blanc que blanc mais ce n'était pas non plus complètement noir. C'était gris, et c'était ce que tout le monde faisait. Comment autrement, tous ces autres chercheurs trouveraient-ils autant d'excellents résultats (…) Je n'ai pas pu résister à la tentation d'aller au-delà, de passer du gris au noir. (…) Je voulais être publié dans les meilleures revues et parler dans les plus grandes salles aux conférences. Je voulais que chacun boive mes paroles tandis que je m'apprêtais à prendre un café ou un déjeuner après une intervention. Je me sentais très seul.

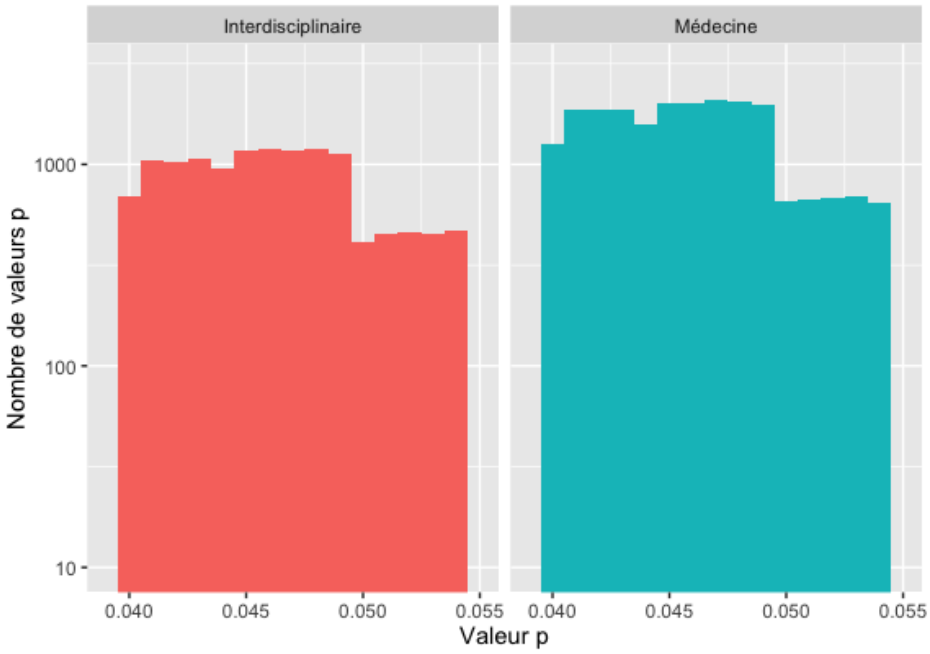
Ces pratiques de recherches discutables sont très variées. Une proposition de typologie en dix catégories pointe aussi bien le non-signalement des résultats négatifs, l'arrondissement des valeurs p ou la sélection des modèles ou des données (Fraser et al. 2018). Le perfectionnement des logiciels de statistiques a vraisemblablement contribué à la diversification des pratiques grises : il est aujourd'hui très facile de procéder à des analyses multiples pour ne retenir que celles qui donnent les résultats les plus positifs.

À la différence de graphes ou de visualisations entièrement bricolés, ces déformations sont souvent indécelables à l'échelon d'une publication individuelle : il n'est pas possible de déduire assurément qu'une valeur p à 0,045 a été arrondie ou arrangée. Les pratiques grises laissent cependant des traces à grande échelle.

Un projet de *text mining* de 2016 est ainsi parvenu à récupérer les valeurs p mentionnées dans des centaines de milliers d'articles de la base *PubMed* (Head et al. 2015). Nous avons effectué de nouveaux calculs à partir de ce jeu de données[7]. La présence régulière de *p-hacking* est signalée par un accroissement soudain du nombre d'occurrences suivi d'une "chute" brutale avant et après la valeur conventionnelle de 0,05.

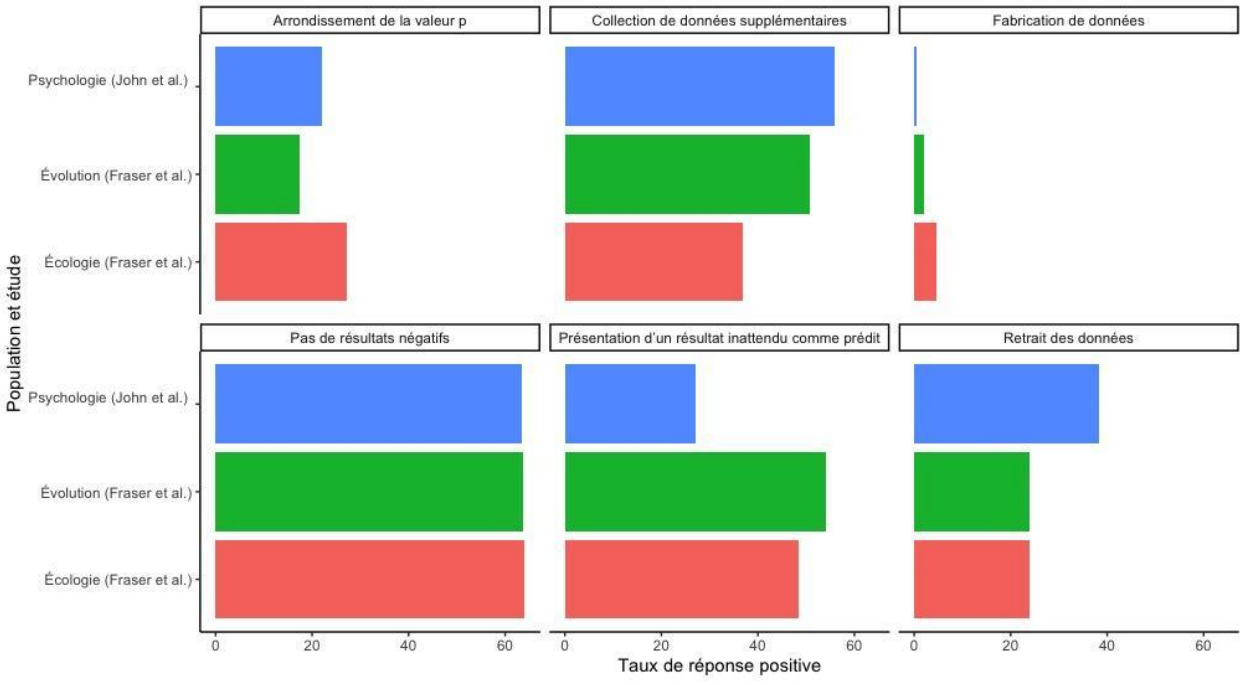
[6] Cf. la discussion à l'adresse : https://pubpeer.com/publications/15084715?order=1

[7] Publié à l'adresse : http://datadryad.org/review?doi=doi:10.5061/dryad.79d43



Occurrences des valeurs p en médecine et en "interdisciplinaire" dans la base PubMed avec un zoom sur le décrochage autour de 0,05.

La présence indubitable de *p-hacking* dans la catégorie fourre-tout *interdisciplinaire* corrobore l'hypothèse qu'il s'agit bel et bien d'un problème général affectant une part significative des publications scientifiques. Ces résultats sont cohérents avec plusieurs sondages déclaratifs récents sur les pratiques de recherche discutables dans plusieurs disciplines (John, Loewenstein, and Prelec 2012; Fraser et al. 2018) :



Prévalence des pratiques de recherche discutables dans trois disciplines différentes.

Dans l'ensemble, près de la moitié de chercheurs interrogés en psychologie, en biologie et en sciences environnementales déclarent avoir multiplié les analyses préalables ou ajouté des données nouvelles afin d'améliorer les résultats. La fabrication de données est par contre massivement rejetée et l'arrondissement de la valeur p est plutôt désapprouvé. Il existe une gradation informelle dans la méconduite scientifique qui ne reflète pas nécessairement l'impact négatif réel de ces pratiques. En terme de reproductibilité, l'arrondissement de la valeur p (qui n'est qu'un critère conventionnel) est probablement moins problématique que le *cherry picking* des bonnes analyses et des bons résultats.

Repenser les infrastructures de recherche : vers la *reproducibility by design* ?

Depuis l'ouverture de la « crise de reproductibilité », de nombreuses propositions ont été émises pour rendre la recherche scientifique plus reproductible et limiter les pratiques de recherches questionnables. Les pistes les plus prometteuses ont été synthétisées dans des manifestes pour la reproductibilité (Munafò et al. 2017) et expérimentées dans de grands projets pilotes (comme le *Reproducibility Project,* le *Many Labs project* ou le *Social Science Replication Project*). Elles commencent à inspirer des implémentations concrètes dans les politiques des éditeurs ou des financeurs. Plus de 5000 revues et organisations soutiennent officiellement les *TOP Guidelines* (littéralement recommandations pour la transparence et l'ouverture), une initiative internationale lancée par le journal *Science* en 2015 (Nosek et al. 2015). Cette approche s'intègre progressivement dans les grands programmes pour la science ouverte. Pour le projet européen d'infrastructure du libre accès OpenAire, le soutien à reproductibilité forme l'une des conséquences logiques du plan S[8].

Le partage des données (*data sharing* ou *scientific open data*) est probablement la recommandation la plus récurrente : c'est le premier point des *TOP Guidelines*. Il facilite considérablement le travail de vérification *a posteriori* des traitements statistiques ainsi que le repérage des fraudes et de certaines manipulations occasionnelles. Plusieurs éditeurs, dont PLOS, imposent aujourd'hui la diffusion des jeux de données détaillés au titre de fichiers supplémentaires, sauf dans quelques cas circonscrits (notamment au titre de la protection des données personnelles). Il existe une large offre de dépôts de données institutionnels ou privés (*Zenodo*, *Figshare*, *Data Dryad*), même si l'indexation des données reste encore imparfaite[9].

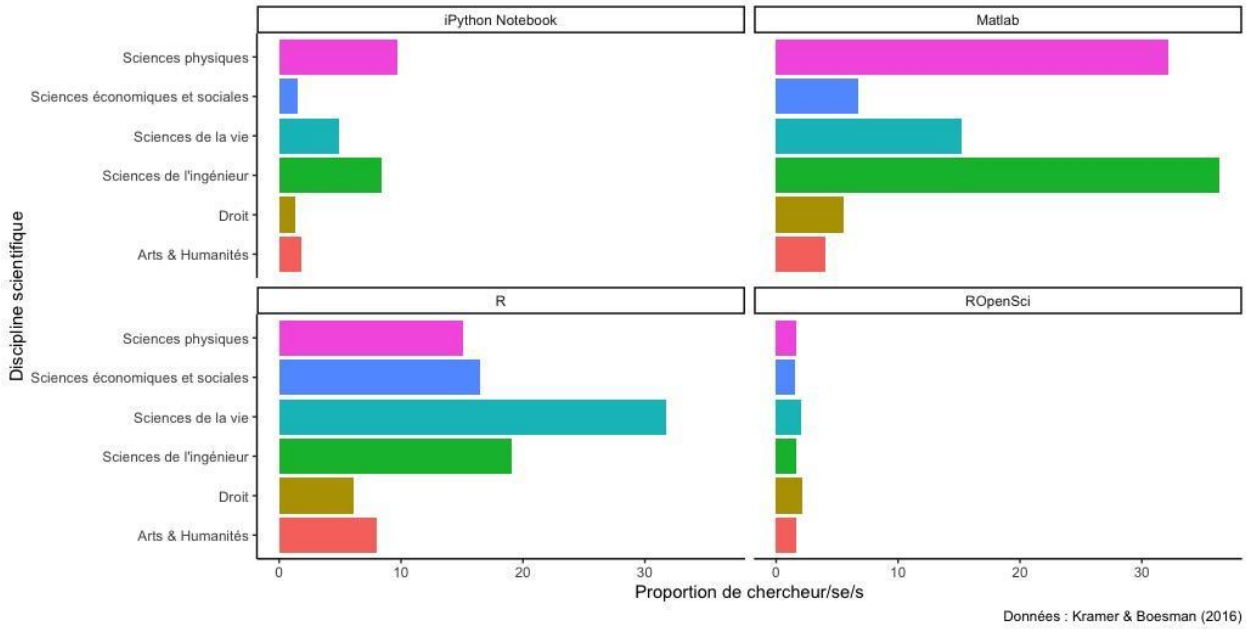
Le partage des données ne dit cependant rien des méthodes appliquées ni des critères de sélection. De nouveaux outils visent à rendre ces traitements plus transparents et à mieux documenter les coulisses de la recherche. C'est notamment le cas des *carnets de code**. Les premiers carnets de code ont été introduits en 2011 par le projet *IPython*. Initialement focalisé exclusivement sur le langage Python, le projet a progressivement intégré la plupart des langages couramment utilisés par les chercheurs (tels que R ou Julia) : signe de cette identité

[8] « Infrastructures to support research and reproducibility need to be built on the principle of open ». https://www.openaire.eu/plan-s-a-european-open-access-mandate

[9] Google vient ainsi de lancer officiellement un moteur de recherche des données, notamment scientifique : https://datasetsearch.research.google.com/

plus générique, il a pris le nom de « projet Jupyter ». Des formats similaires ont été développés pour le langage R sous le titre *R Notebook*. Le carnet de code marque l'aboutissement d'un processus ancien de légitimation de l'écriture informatique comme écriture savante : les étapes successives du traitement informatique deviennent le fil conducteur du travail de recherche[10]. Il devient possible de refaire une recherche antérieure et d'en ajuster les paramètres.

Dans l'enquête menée par Bianca Kramer et Jeroen Bosman sur les usages numériques de 20 000 chercheurs de 142 pays, les carnets de code apparaissaient déjà comme une pratique émergente, alors que la première version stable de Jupyter venait tout juste d'être officialisée. Dans certaines disciplines (sciences physiques et sciences de l'ingénieur), près de 10% des répondants avaient expérimenté ce nouveau format. Les données de l'étude suggèrent également que l'outil a un potentiel important : les langages de programmation couramment utilisés dans les carnets de code ou des interfaces similaires mais propriétaires (*matlab*) sont déjà d'usage courant.



Usage de dispositifs de carnets de code dans Kramer & Boesman (2016)

La promesse de reproductibilité des carnets de code reste cependant inachevée. Deux études récentes montrent qu'une grande partie du million de carnets actuellement déposés sur Github ne s'exécute pas correctement (Rule, Tabard, and Hollan 2018; Pimentel et al. 2019). Dans de nombreux cas, les extensions nécessaires ne sont plus compatibles, en l'absence de déclaration explicite de la version utilisée. Plus occasionnellement, les fichiers de données à charger ou les variables ne sont pas correctement déclarés. De toute évidence, l'outil ne peut suffire à faire émerger un environnement de recherche reproductible.

Cette offre technique émergente pourrait s'associer à de nouvelles procédures de publications. Depuis les années 1990, les recherches de médecine clinique peuvent donner

[10] Dès 1984, Donald Knuth publie un manifeste en faveur d'une « programmation lettrée » (Knuth 1984).

lieu à des *pré-enregistrements* : avant même de conduire l'étude, ses initiateurs déposent un plan plus ou moins détaillé de l'expérience envisagée, de ses attendus et des méthodes d'analyses utilisées, par exemple sur ClinicalTrials.gov. Cette forme de proto-publication était alors justifiée par les exigences particulières qui pesaient sur ce type d'étude au regard de leur impact sanitaire concret. Elle suscite depuis peu l'intérêt de disciplines variées. Le renversement de la chronologie de la publication ne permet en effet plus de reconstruire *a posteriori* le cadre de l'expérience en sélectionnant les données, les méthodes ou les hypothèses les plus arrangeantes. Une autre forme émergente, les *replication studies*, n'intervient pas en amont mais en aval : il s'agit de faire de la reproduction parfois coûteuse d'une expérience antérieure une publication scientifique à part entière, tout aussi bien évaluée et valorisée qu'une expérience inédite.
D'autres réformes plus profondes sont également communément discutées, comme la mise en œuvre de procédures d'évaluation ouverte ou la réévaluation des incitations sociales actuelles dans les disciplines scientifiques (telles que le *publish or perish* ou la fétichisation des découvertes inédites au détriment des procédures tout aussi fondamentales de vérification et d'enrichissement graduel des connaissances scientifiques). C'est que le débat sur la reproductibilité a fait peut-être émerger un problème plus profond : au-delà des enjeux de libre accès, la publication scientifique classique n'a jamais été adaptée à l'utilisation intensive de données quantitatives. La forme « article » ne se prête qu'à l'inclusion de quelques indicateurs imparfaits. Elle dissimule par nature les étapes successives du traitement en n'intervenant qu'en fin de course, alors que tout est déjà joué. Dès les années 1960, les premiers essais de méta-analyse mettent en évidence ces carences qui ne vont devenir que plus patentes au cours des décennies suivantes. La science reproductible est peut-être moins un but en soi qu'une motivation supplémentaire pour faire émerger un nouvel écosystème de production et de circulation des données scientifiques.

Glossaire

*Pratiques discutables en recherche** (*Questionable Research Practice* ou *QRP*) :
manipulations plus ou moins triviales des données ou des méthodes de traitement visant à présenter des résultats plus signifiants ou attractifs.

Réplication : réitération d'une expérience dans un cadre quasiment identique.

Reproduction : réitération d'une expérience dans un cadre différent (pas le même lieu, la même équipe ou le même matériel) mais en suivant approximativement le même plan d'expérience.

Résultat négatif ou résultat nul : résultat non signifiant, qui ne permet pas de réfuter une hypothèse nulle dans le cadre d'une analyse statistique « fréquentiste ».

EPRIST

Bibliographie

Baker, Monya. 2016. “1,500 Scientists Lift the Lid on Reproducibility.” *Nature News* 533 (7604): 452. https://doi.org/10.1038/533452a.

Drummond, Chris. 2018. “Reproducible Research: A Minority Opinion.” *Journal of Experimental & Theoretical Artificial Intelligence* 30 (1): 1–11. https://doi.org/10.1080/0952813X.2017.1413140.

Fanelli, Daniele. 2018. “Opinion: Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?” *Proceedings of the National Academy of Sciences* 115 (11): 2628–31. https://doi.org/10.1073/pnas.1708272114.

Fraser, Hannah, Tim Parker, Shinichi Nakagawa, Ashley Barnett, and Fiona Fidler. 2018. “Questionable Research Practices in Ecology and Evolution.” *PLOS ONE* 13 (7): e0200303. https://doi.org/10.1371/journal.pone.0200303.

Gómez, Omar S., Natalia Juristo, and Sira Vegas. 2010. “Replications Types in Experimental Disciplines.” In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–10. ESEM ’10. Bolzano-Bozen, Italy: Association for Computing Machinery. https://doi.org/10.1145/1852786.1852790.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. “The Extent and Consequences of P-Hacking in Science.” *PLOS Biology* 13 (3): e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Hubbard, Raymond, and R. Murray Lindsay. 2008. “Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing:” *Theory & Psychology*, February. https://doi.org/10.1177/0959354307086923.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Ioannidis, John P. A., Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. 2015. “Meta-Research: Evaluation and Improvement of Research Methods and Practices.” *PLoS Biology* 13 (10). https://doi.org/10.1371/journal.pbio.1002264.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. “Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling:” *Psychological Science*, April. https://doi.org/10.1177/0956797611430953.

Knuth, D. E. 1984. “Literate Programming.” *The Computer Journal* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Munafò, Marcus, Brian Nosek, Dorothy Bishop, Katherine Button, Christopher Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer Ware, and John Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (January): 0021. https://doi.org/10.1038/s41562-016-0021.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25. https://doi.org/10.1126/science.aab2374.

Pimentel, João Felipe, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. 2019. "A Large-Scale Study about Quality and Reproducibility of Jupyter Notebooks." In *Proceedings of the 16th International Conference on Mining Software Repositories*, 507–517. MSR '19. Montreal, Quebec, Canada: IEEE Press. https://doi.org/10.1109/MSR.2019.00077.

Rule, Adam, Aurélien Tabard, and James D. Hollan. 2018. "Exploration and Explanation in Computational Notebooks." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. CHI '18. Montreal QC, Canada: Association for Computing Machinery. https://doi.org/10.1145/3173574.3173606.

Schor, Stanley, and Irving Karten. 1966. "Statistical Evaluation of Medical Journal Manuscripts." *JAMA* 195 (13): 1123–28. https://doi.org/10.1001/jama.1966.03100130097026.

Steinle, Friedrich. 2016. "Stability and Replication of Experimental Results: A Historical Perspective." In *Reproducibility*, 39–63. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118865064.ch3.