Reproductibilité numérique : enjeux de crédibilité pour les expériences de simulation

Paul-Antoine Bisgambiglia [1] et David R.C. Hill

Cet article vise à sensibiliser les utilisateurs d'outils de simulation et à compléter les éléments produits pendant la journée de la Société informatique de France consacrée à la reproductibilité en 2021. Nous proposons de revenir sur les définitions de base, sur la distinction entre répétabilité et reproductibilité numérique et nous apportons un éclairage sur les pratiques permettant de consolider la crédibilité des résultats de simulation informatique.

Introduction

Karl Popper a profondément impacté la production des connaissances dans de nombreux domaines [21], il reste encore aujourd'hui une référence majeure en épistémologie. Les dérives récentes des pratiques scientifiques ont conduit la société à réaliser l'importance des critères de scientificité qu'il avait mis en avant, notamment un critère majeur qui est la reproductibilité d'expériences scientifiques.

Nous pouvions lire, dans un article du journal Le Monde, daté du 2 octobre 2017, qui a pour titre « *La rigueur scientifique à l'épreuve de la reproductibilité* » que : « *dès 2005, John Ioannidis, de l'université Stanford, suggérait de façon provocatrice dans un article de PloS Medicine :* « la plupart des résultats scientifiques sont faux », car impossibles à reproduire.

1. Maître de conférences à l'université de Corse.

DOI:10.48556/SIF.1024.19.137

Dans [11], nous pouvons lire que « *même les estimations les plus prudentes de la recherche en biomédecine placent le taux de reproductibilité à moins de 50 %.* ». Au même moment en 2019, le Guardian faisait un bilan des travaux financés en Grande-Bretagne par le fond pour l'excellence scientifique, et même ici, le taux de reproductibilité pour les travaux scientifiques du domaine médical tombait à 11 %. Dans le secteur du numérique, nous pourrions nous attendre à de bien meilleures « performances » car nos machines et nos piles logicielles sont supposées toujours déterministes, mais une étude poussée [3], datant de la même époque a montré que la reproductibilité des travaux de recherche en informatique ne dépassait guère les 30 %.

Une recherche [2] rapide sur *Web of Science* avec les termes *reproducible research* donne 51 862 résultats, et avec le terme *reproducibility* donne 139 857 résultats. Le domaine qui semble produire le plus d'études en ce sens est la chimie analytique avec 22,565 résultats, soit plus de 16 %. La production par an est croissante, elle a presque doublé en 10 ans passant de 5000 à plus de 9000 articles entre 2012 et 2021. Le français est la quatrième langue de production, mais ne représente que 0.5 % des articles alors que plus de 6 % de ces articles sont produits par des équipes françaises. Ce focus montre l'intérêt croissant du domaine. Nous retrouvons également trois articles significatifs, dont un éditorial, sur le sujet dans la revue Nature fin 2021 [1, 10, 19]. Nous pouvons y lire que reproduire les résultats scientifiques est difficile et chronophage, mais crucial, ainsi l'effort doit être partagé entre les auteurs, les laboratoires et la communauté scientifique : « *L'ensemble de la communauté scientifique doit reconnaître que la réplication d'une observation ou d'un résultat est indispensable, elle permet d'acquérir une assurance essentielle au progrès de la science : elle montre qu'une observation ou un résultat est suffisamment solide pour encourager de futurs travaux* [3]. ».

L'idée première derrière la notion de reproductibilité en informatique est que l'on puisse reproduire une expérience qui a été partagée au sein de la communauté scientifique dans le but d'obtenir strictement les mêmes résultats. Dans la partie qui suit nous aborderons des définitions en se rappelant des pratiques dans les sciences expérimentales plus anciennes, comme la biologie et la physique. Tout le déroulé d'une expérimentation est noté sur un cahier de laboratoire et partagé avec les pairs afin d'en permettre la reproductibilité et d'en valider les conclusions, ces approches sont transposables aux expériences numériques de simulation.

Dans son exposé pour la Société informatique de France, Christophe Pouzat, définit la recherche reproductible comme : « *une démarche qui consiste à fournir aux*

2. Recherche effectuée le 26 Novembre 2021.

3. Traduction de « *The entire scientific community must recognize that replication is not for replication's sake, but to gain an assurance central to the progress of science : that an observation or result is sturdy enough to spur future work* ».

lecteurs d'articles, d'ouvrages, etc., l'ensemble des données et des programmes accompagnés d'une description algorithmique de la façon dont les programmes ont été appliqués aux données pour obtenir les résultats présentés. ». Obtenir les résultats présentés est bien souvent problématique, ce phénomène est accentué en informatique par la prolifération de codes de calcul ou de simulation, faciles à utiliser et mis à disposition sans contexte d'utilisation ou avec un manque de documentation. Cependant, une raison très fréquente qui empêche d'obtenir les résultats produits dans une étude publiée est clairement l'absence d'accès au code informatique. La simulation informatique s'est développée au point de devenir un outil précieux de production de connaissances et d'aide à la décision, et c'est même le seul outil, indispensable, pour l'exploration des systèmes dits complexes. De nombreux processus de décision utilisent des résultats de simulation informatique pour guider les décideurs : aérodynamisme de voiture, structure de bâtiment, propagation de virus, etc. Dans bien des cas, les outils ou codes de simulation qui ont fait d'énormes progrès en termes d'ergonomie et de facilité d'utilisation. Il arrive alors qu'ils soient utilisés sans assez de recul sur les mécanismes sous-jacents : méthodes de résolution ou de discrétisation, générateurs de nombres pseudo-aléatoires, ordonnanceurs d'événements, mécanismes de gestion de la concurrence, mécanismes de compilation, etc. La pile logicielle qui permet d'utiliser une application est supposée toujours maîtrisée, ce qui n'est pas toujours le cas pour des applications de calcul souvent sophistiquées, et également dans bien des chaînes logicielles modernes ou il n'est plus « évident » d'avoir deux logiciels identiques lors de deux *builds* successifs [18].

Les causes de non-reproductibilité sont nombreuses et complexes : développement trop rapide de codes de simulation personnels, complexité des codes, difficultés et coûts de maintenance des codes existants, mauvais usage des générateurs de nombres pseudo-aléatoires ou tout simplement manque de rigueur, ou encore et surtout de temps, etc.

Ces éléments sont abordés dans [2], nous les avons traités dans le cadre de la thèse de [12] qui s'intéressait à un aspect facilitant la reproductibilité des expériences de simulation, à savoir la formalisation du processus de modélisation et l'usage de code de simulation univoque. Nos propositions avaient pour but de faciliter la reproductibilité de simulations multi-agents, mais peuvent compliquer l'usage du paradigme agent et ne concernent qu'une partie limitée de la problématique de la science reproductible. Nous avions pris soin des aspects stochastiques des simulations depuis 2014 dans [8], puis nous nous sommes intéressés aux problèmes de reproductibilité des simulations stochastiques distribuées dans [15, 17]. Une rapide association de ces différents travaux est proposée dans [13]. Forts de ces différentes expériences, nous proposons dans cet article de revenir sur les définitions du domaine et proposons de lister et résumer quelques bonnes pratiques.

Définitions

Dans le cadre d'expériences numériques, [7, 6, 17, 11] proposent un état de l'art et des définitions issues de la littérature. Deux nuances sont dégagées dans ces études sur la reproductibilité : (1) la reproduction de l'expérience numérique et (2) la reproduction exacte des résultats numériques obtenus. Cette reproductibilité numérique — parfois appelée *bitwise reproducibility* en anglais — est nécessaire ne serait-ce que pour mettre au point les programmes numériques sur ordinateur (sans débogage, plus de logiciels).

La nuance entre reproductibilité et répétabilité est également mise en avant : la répétabilité consiste à retrouver les mêmes résultats lorsque deux expériences sont menées avec les mêmes paramètres d'entrées, avec des matériels, des méthodes et des contextes identiques. On parle aussi en anglais de *run to run reproducibuility*. À grande échelle sur des supercalculateurs, il arrive que cette répétabilité attendue soit perdue après le lancement de deux ou plusieurs expériences de simulations en tout point identiques. L'utilisation d'optimisations au sein des nouveaux microprocesseurs (*out of order execution*) ou des nouvelles instructions machines fusionnées du type FMA(*Fused Mutiply Add*) ou de type vectorielles SIMD (*Single Instruction Multiple Data*) sont des causes de non-reproductibilité qui deviennent fréquentes [16]. Il devient alors difficile de mettre au point les programmes à grande échelle que l'on fait tourner sur des clusters plus lents que les supercalculateurs en production sur lesquels les erreurs de répétabilité ont été constatées.

La reproductibilité se veut donc plus générale que la répétabilité. Parmi les apports de la reproductibilité, [7] nous rappellent qu'elle « constitue une méthode et un standard pour juger de la pertinence d'une expérience numérique publiée et donc des conclusions qui en découlent ». Entre la notion de répétabilité et de reproductibilité, différents degrés sont suggérés dans la littérature. La reproductibilité implique des changements, mais l'obtention de la même conclusion scientifique. C'est un critère essentiel pour la scientificité d'une étude selon Karl Popper. Des équipes différentes, des méthodes différentes, des instruments différents donnent (heureusement) les mêmes conclusions scientifiques. Dans la phrase précédente, le terme « conclusion scientifique » est préféré au terme « résultat ». Dans un contexte numérique, il faut dans bien des cas complexes savoir se contenter de résultats « seulement » similaires, donnant la même conclusion scientifique. C'est un critère de validation, etc.

Obtenir les mêmes résultats numériques lorsque quelque chose a changé dans le contexte d'exécution revient à parler de portabilité. Le standard de calcul numérique flottant IEEE 754 a été élaboré à ce propos. Nous pouvons aussi citer l'importance du respect de l'ordre d'exécution des calculs flottants au sein des compilateurs. Nous sommes proches de la répétabilité par le fait que le même résultat numérique est

obtenu — se répète —, mais comme le contexte a changé nous sommes dans le cadre large de la reproductibilité dite *bitwise reproducibility*.

Pour asseoir les définitions, il faut se replacer au niveau épistémologique — dans ce cadre la répétabilité suppose que nous n'avons aucun changement dans la réalisation de l'expérience. La répétabilité n'est donc pas un élément de la démarche scientifique au sens de Karl Popper [20].

Chris Drumond rappelle même que ce n'est pas ce que l'on attend dans une démarche scientifique [9]. Mais il se trouve que cette caractéristique est essentielle et attendue dans le contexte de développements informatiques purement déterministes. Sans cette propriété comment mettre au point des programmes, comment les déboguer [17]. Dans [5], nous trouvons quatre niveaux de reproductibilité : (1) la disponibilité du code source, (2) les aspects non-déterministes du calcul, (3) la similitude du plan d'expérience, et (4) les sorties du modèle.

Dans leurs travaux, les auteurs de [22] définissent également plusieurs niveaux de reproductibilité :

(1) l'examen par les pairs, qui est la méthode traditionnelle de publication (les travaux et résultats décrits sont jugés crédibles par la communauté scientifique) ;

(2) la recherche réplicable, où les outils permettant de reproduire les mêmes résultats sont fournis ;

(3) la recherche vérifiable, où les mêmes conclusions peuvent être atteintes indépendamment du code source fourni par l'auteur ;

(4) la recherche validable, où suffisamment de ressources (sources et données) sont archivées afin de permettre de défendre les résultats fournis ;

(5) la recherche ouverte, où tous les éléments utilisés pour arriver aux résultats présentés sont fournis en accès libre et documentés.

Il est important de préciser le cas des calculs stochastiques. Soit la source de hasard utilisée est maîtrisée : générateurs pseudo-aléatoires ou quasi-aléatoires et dans ce cas, nous avons un modèle maîtrisé (et déterministe !) du hasard et nous sommes à même de répéter numériquement les expériences des simulations stochastiques. Soit, nous avons une source non déterministe telle que celle utilisée pour du calcul quantique (avec un algorithme et des circuits quantiques) et dans ce cas nous attendons la reproductibilité des conclusions scientifiques, mais nous n'aurons jamais de répétabilité numérique.

Dans les travaux cités, nous voyons aussi apparaître les notions de vérification et de validation. Ce sont des notions importantes pour la communauté des chercheurs en simulation et plus largement en ingénierie logicielle. Elles sont employées dans cette communauté avec un sens précis et sont les piliers de la crédibilité que l'on peut accorder aux études de simulation. Le travail de vérification vise à montrer qu'un modèle informatisé (opérationnel et exécutable) représente bien un modèle

conceptuel dans des limites de précisions spécifiées. En ce qui concerne la validation des modèles, il s'agit de montrer qu'un modèle opérationnel informatisé possède une plage de précision satisfaisante compatible avec l'application prévue du modèle dans son domaine d'applicabilité (son cadre expérimental). Pour plus de détails sur les différentes techniques de vérification et de validation, le lecteur intéressé pourra consulter le chapitre VII de l'ouvrage suivant [4].

C'est en se souciant de ces aspects tout au long de la conception et de l'utilisation des simulations que notre domaine scientifique pourra gagner en crédibilité.

Discussion

Les conclusions établies par [5, 22, 7, 17] se rejoignent, et nous pouvons les résumer de la manière suivante : pour permettre un degré de reproductibilité se rapprochant de la « recherche ouverte », un changement dans la culture de publication est nécessaire. Ce changement doit être impulsé non seulement par les auteurs, mais également par les éditeurs. Il est recommandé aux premiers de fournir les éléments permettant de reproduire l'expérience de simulation (données, code source, etc.), et aux seconds d'encourager cet effort durant le processus de publication en offrant les supports numériques d'archivage et de diffusion de ce contenu.

Des efforts sont faits, par exemple dans le formulaire d'examen des articles des conférences SCS est apparu en 2020 un item d'évaluation portant sur la reproductibilité des résultats de simulation, c'est également le cas maintenant pour de nombreux journaux. Il y a également de nombreuses initiatives de la communauté comme un MOOC [4] d'Inria « recherche reproductible : principes méthodologiques pour une science transparente » qui met en avant l'utilisation d'outils tels que :

— markdown pour la prise de notes structurées ;

— des outils d'indexation (DocFetcher et ExifTool) ;

— gitlab / github pour le suivi de version et le travail collaboratif ;

— notebooks (Jupyter, Rstudio ou Org-mode) ;

— ou encore, la journée de la SIF [5].

Il convient de citer également : le projet Software Heritage d'Inria pour l'archivage de code source ; les ReproHackathon du GDR MaDICS qui visent à tester les capacités des systèmes de workflows disponibles à reproduire une expérience scientifique, etc.

Ce processus d'échange des méthodes, codes et protocoles d'expériences augmente la confiance dans les résultats. Il est donc en partie garant du respect et de

4. https://www.fun-mooc.fr/courses/course-v1:inria+41016+self-paced/about.
5. https://www.societe-informatique-de-france.fr/journee-reproductibilite.

la crédibilité de la méthode scientifique en cette période où les controverses scientifiques montrent plus souvent les limites de la méthode et l'impact des conflits d'intérêts sur la production de résultats qui peuvent être mis à jour grâce à l'absence de reproductibilité.

Dans son dernier ouvrage [14], Stephen Grossberg, pionnier des modèles cognitifs utilisés aujourd'hui dans l'apprentissage profond, précise que nous ne pouvons pas nous fier de façon aveugle à ces techniques. Les principales raisons sont d'une part, la non explicabilité, et d'autre part son inadéquation à plusieurs types de domaines applicatifs où la technique peut engendrer des « oublis catastrophiques ». Nous constatons également que lorsque ces algorithmes fonctionnent, il peut être impossible de comprendre pleinement pourquoi et que lors d'apprentissages sur de nouvelles bases de données, des souvenirs construits précédemment disparaissent de façon arbitraire.

Si la reproductibilité des algorithmes est assez simple à obtenir (la classification a cette souplesse), la possibilité de rejouer ou de répéter strictement les expériences est nécessaire pour progresser dans la connaissance (dans la science). Restons prudent dans l'usage de ces techniques pour toutes applications où les conséquences peuvent être lourdes – notamment dans le cadre d'applications médicales par exemple. Il est fréquent que ces modèles utilisent des simulations stochastiques en nombre massif pour la phase d'apprentissage. C'était notamment le cas pour l'entrainement d'Alpha Go qui avait battu Lee Sedol le champion du monde de Go. Pour éviter les écueils de ces nouvelles méthodes, Stephen Grossberg propose d'approfondir les travaux en lien avec la logique floue à partir d'un modèle qu'il appelle ART (*Adaptive Resonance Theory*). Ce type de modèle permettrait d'éviter les problèmes évoqués et on peut espérer valider des modèles de ce type couplés aux simulations, pour explorer et entraîner les modèles.

Références

[1] O. B. Amaral and K. Neves. Reproducibility : expect less of the scientific paper. 597, 2021/09.

[2] P.-A. Bisgambiglia. *Habilitation à diriger des recherches : Les expériences virtuelles de simulation comme outils d'aide à la prise de décisions des données au processus de décision*. PhD thesis, Université de Corse, Université de Corse, 07 2021.

[3] C. Collberg and T. A. Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 59(3) :62–69, 2016.

[4] P. Coquillard and D. R. C. Hill. *Modélisation et simulation d'écosystèmes : Des modèles déterministes aux simulations à événements discrets*. Masson, Recherche en Écologie, 1997. ISBN 2-225-85363-0.

[5] O. Dalle. On reproducibility and traceability of simulations. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12, Dec. 2012.

[6] V. T. Dao. *Calcul à haute performance et simulations stochastiques. Etude de la reproductibilité numérique sur architectures multicore et manycore*. PhD thesis, LIMOS – UMR CNRS 6158, Université Clermont Auvergne, Clermont, Mar. 2017. 00000.

[7] V. T. Dao, V. Breton, H. Nguyen, and D. R. C. Hill. La reproductibilité des simulations stochastiques parallèles et distribuées utilisant le calcul à haute performance. *Journées DEVS Francophone*, pages 109–117, 2016.

[8] V. T. Dao, L. Maigne, V. Breton, H. Nguyen, and D. R. C. Hill. Numerical reproducibility, portability and performance of modern pseudo random number generators : Preliminary study for parallel stochastic simulations using hybrid xeon phi computing processors. In *European Simulation And Modelling Conference*, pages 80–87, 2014.

[9] C. Drummond. Replicability is not reproducibility : nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning workshop 26th International Conference for Machine Learning 2009*, 2009.

[10] N. Editorials. Replicating scientific results is tough — but crucial. 600, 2021/12.

[11] B. G. Fitzpatrick. Issues in reproducible simulation research. *Bulletin of mathematical biology*, 81(1) :1–6, 2019.

[12] R. Franceschini. *Approche formelle pour la modélisation et la simulation à évènements discrets de systèmes multi-agents*. phdthesis, Université de Corse Pasquale Paoli, Dec. 2017.

[13] R. Franceschini, P.-A. Bisgambiglia, and D. R. C. Hill. Reproducibility Study of a PDEVS Model Application to Fire Spreading. In *Proceedings of the 50th Computer Simulation Conference*, SummerSim '18, pages 29 :1–29 :11, San Diego, CA, USA, 2018. Society for Computer Simulation International.

[14] S. Grossberg. *Conscious Mind, Resonant Brain : How Each Brain Makes a Mind*. Oxford University Press, 2021.

[15] D. R. C. Hill. Parallel Random Numbers, Simulation, and Reproducible Research. *Computing in Science & Engineering*, 17(4) :66–71, July 2015. 00001.

[16] D. R. C. Hill. Numerical reproducibility of parallel and distributed stochastic simulation using high-performance computing. In *Computational Frameworks*, pages 95–109. Elsevier, 2017.

[17] D. R. C. Hill, V. T. DAO, C. Mazel, and V. Breton. Reproductibilité et répétabilité numérique. constats, conseils et bonnes pratiques pour le cas des simulations stochastiques parallèles et distribuées. *Technique et Science Informatiques*, 36(3-6) :243, 2017.

[18] C. Lamb and S. Zacchiroli. Reproducible builds : Increasing the integrity of software supply chains. *IEEE Software*, 2021.

[19] N. C. Nelson et al. Understand the real reasons reproducibility reform fails. *Nature*, 600(7888) :191–191, 2021.

[20] A. O'Hear. Karl popper : Philosophy and problems. *Royal Institute of Philosophy*, 1995.

[21] K. Popper. *Logique de la découverte scientifique*. édition originale : Logic der Forschung, Springer, Wien, 1934, payot edition, 1973.

[22] V. Stodden, J. Borwein, and D. H. Bailey. Setting the default to reproducible. *computational science research. SIAM News*, 46(5) :4–6, 2013. 00030.