Received: 12 July 2024 | Accepted: 29 August 2024

DOI: 10.1111/1751-7915.70014

MICROBIAL BIOTECHNOLOGY Open Access

OPINION

Artificial intelligence-based prediction of pathogen emergence and evolution in the world of synthetic biology

Antoine Danchin

School of Biomedical Sciences, Li KaShing Faculty of Medicine, Hong Kong University, Pokfulam, SAR Hong Kong, China

Correspondence
Antoine Danchin, School of Biomedical Sciences, Li KaShing Faculty of Medicine, Hong Kong University, 21 Sassoon Road, Pokfulam, SAR Hong Kong, China.
Email: antoine.danchin@academie-sciences.fr and adanchin@hku.hk

Abstract

The emergence of new techniques in both microbial biotechnology and artificial intelligence (AI) is opening up a completely new field for monitoring and sometimes even controlling the evolution of pathogens. However, the now famous generative AI extracts and reorganizes prior knowledge from large datasets, making it poorly suited to making predictions in an unreliable future. In contrast, an unfamiliar perspective can help us identify key issues related to the emergence of new technologies, such as those arising from synthetic biology, whilst revisiting old views of AI or including generative AI as a generator of abduction as a resource. This could enable us to identify dangerous situations that are bound to emerge in the not-too-distant future, and prepare ourselves to anticipate when and where they will occur. Here, we emphasize the fact that amongst the many causes of pathogen outbreaks, often driven by the explosion of the human population, laboratory accidents are a major cause of epidemics. This review, limited to animal pathogens, concludes with a discussion of potential epidemic origins based on unusual organisms or associations of organisms that have rarely been highlighted or studied.

INTRODUCTION

As the Danish aphorism goes, it is awkward to prophesy, especially about the future (Hill, 1956). Yet, predicting epidemics to trigger an appropriate response is the dearest wish of public health institutions. A Google search revealed 50 million pages on the subject of 'predicting the emergence and evolution of pathogens', half of which were created or updated in 2023. With 'scientific article' added to the query, the collection shrinks drastically, but there are still 10,000 pages left. Does this situation justify yet another article? Unlike other physical phenomena, natural selection has ensured that those associated with life produce unexpected behaviour. This is the very condition for the successful survival of a progeny in an unreliable world. It may therefore seem paradoxical to try and predict the emergence of new infectious diseases, especially in a world where scientific advance results in the emergence of new microbes. However, by making the best use of scientific knowledge, we can develop ways to circumvent the presence of life's innovations – including those that we are creating – and mitigate the consequences of this inevitable obstacle.

Science is meant to work out how reality has behaved, is behaving today and will behave in the future. Whilst it is impossible to predict an expected scenario for each individual disease, this may no longer be the case if we set out to use new scientific developments to understand how families of pathogenic organisms and diseases emerge. Here, based on what we know from the past, coupled with our ever-advancing understanding of pathogens, we examine the benefits and risks associated with the new avatar of computer technology long known as artificial intelligence (AI). In particular, because AI is based on previously accumulated knowledge, we consider the role of accidents, which have been largely overlooked as a critical factor in the

emergence of epidemics. Beyond our current emphasis on quantitative biology, which is generally based on a mechanistic – hence predictable – view of biological processes, this extension implies the identification of features of laboratory experiments or public behaviour that can lead to the emergence of infectious diseases. This question is particularly timely given the advent of synthetic biology, which allows us to reconstruct or even design the genome of pathogens. Finally, we highlight hitherto unexplored features of pathogen contamination as unexpected causes of emergent or re-emergent disease, and how AI could assist us in identifying these circumstances.

BEWARE OF ACCIDENTS

Here, we restricted our focus to animal infectious diseases. Although there may be common tropes for all types of infection, plants and animals have a different way of responding to pathogens, with an immune adaptive system specific to the latter. Human beings, as warm-blooded vertebrates, are very different from plants, and the pathogens that infect them are far more likely to come from families related to mammals or birds than to food plants or even distant vertebrates such as snakes. This reminder forms the basis of the One Health programme, which proposes the systematic monitoring of animal diseases, especially those affecting livestock and pets, but also those with which we come into contact when travelling or invading pristine environments, as a possible, if not likely, source of new human diseases (Pitt & Gunn, 2024).

It is not difficult to imagine that, whilst the process of infection combines the characteristics of a host and those of an infectious agent, there may be other, less plausible or even random causes of epidemics. Predictions are based on models that assume that reality is sufficiently regular. However, accidents are the norm, not the exception, and accidents, by definition, cannot be predicted. Some predictive power remains possible if we retain only the generality of accidents, not a particular case. In this sense, one can speak of 'normal' accidents (Perrow, 1999). For example, without recourse to AI, human experts know that epidemics and even pandemics are likely to occur in the near future, but they can only make educated guesses about what their aetiological agents will be, the exact circumstances or the origin of their emergence. The most worrying predicaments are Nassim Taleb's 'black swans', that is, unpredicted high-impact events that, in retrospect, should have been understood as inevitable (Taleb & Taleb, 2016). In this context, AI, based on the exploitation of massive datasets (big data), has the potential to identify patterns that have so far escaped human attention. However, data lacks intrinsic meaning unless contextualized and classified (Danchin et al., 2018; Nguyen, 2024). A key feature of anticipating future events whilst taking into account the plausibility of accidents is the inclusion in the analyses of a rich set of metadata (Musen et al., 2022), that is, contextual information about the data, that organizes the knowledge of features pertaining to the disease or epidemic of interest (Schriml et al., 2020). As an illustration, it is imperative to incorporate human social behaviour into the metadata to understand the process of contagion: more often than not, we are the main cause of our diseases (Chaber, 2018; Danchin, 2003). Finally, as emphasized by Charles Perrow again, the highly intricate structure of our political organizations makes them a likely, yet unacknowledged, source of major accidents involving high-impact technologies (Perrow, 2011).

ARTIFICIAL INTELLIGENCE TO THE RESCUE?

After decades of relative obscurity, AI is back in the spotlight. Broadly speaking, this technology aims to emulate what we perceive as human intelligence, enabling computers to perform a range of operations just like human beings. By mimicking human cognitive abilities AI could even replace the human actor in future developments of scientific research (Xu et al., 2021). Also, 'AI' is a collective term that encompasses a wide range of technologies, rather than referring to a single one. This is reflected in the vague popular understanding of what constitutes AI, whereas selecting the appropriate type of AI is likely to have significant implications for its ability to predict epidemics.

Various flavours of artificial intelligence

AI is supposed to do better than most available automata and match the performance of the human brain. Notwithstanding the meaning of the concept of intelligence, a widely used discriminating criterium between human and artificial intelligence has been proposed by Alan Turing (Turing test) to determine whether a machine could be said to be intelligent (Rapaport, 2006). If, after a machine has engaged in a dialogue with a person, it cannot be recognized that it is not a human being, then the machine is deemed displaying an intelligence similar to that of a person. Several AI automata have recently passed the test. However, they failed to solve a variety of logic puzzles that seemed straightforward. This observation must be remembered when interpreting the outcome of AI usage (Biever, 2023). Indeed, should engaging in a dialogue displaying the common features of human language exchanges be considered intelligent? To be blunt: can we accept that the majority of human exchanges that we see spanning social media display marks of intelligence?

Possibly beginning with automata mimicking people, AI has a long history (McCorduck, 2019). The various AI technologies popular today are based on a variety of implementations of machine learning (ML), aiming at the extraction of stable correlations present in very large datasets, usually letting computers organize them as consistent patterns interpreted as rules subsequently used to generate a variety of performances. 'Discriminative AI' bases its prediction on statistical analyses of the data and it is therefore heavily dependent on the statistical models used, and, above all, on the quality of the data and metadata (Danchin et al., 2018; Wilkinson et al., 2016). In the form of 'generative AI' (Sætra, 2023), ML-based AI is very successful at solving the Turing test, whilst it sometimes fails to make trustworthy predictions, as an example will show later. It is not the same thing, indeed, to generate a response that has the same feeling as that which would be uttered by a person, and to generate an informed response. Nothing is more similar to truth than a lie and vice versa (Vellani et al., 2023) and we must have this in mind when predicting future epidemics.

At the time of AI's emergence, today's success of generative AI looked to be far away in the future (https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html). Back then, intelligence was tied to the way we understood human aptitude for language. Language, considered a unique and specific human competence based on grammar organizing syntactic rules [for the history of the concept see Chomsky, 2015, 2017], was the benchmark for intelligent behaviour (Dick, 2019). AI was hypothesis-driven with inferences based on causality, rather than reduced to probabilistic models generating texts (i.e., sequences of symbols) from data. It required a computer scientist to code accepted rules forming grammars that would mimic the way we understood people interact by writing or speaking (Hofstadter, 1979). Extending the concept, grammars proved very useful in describing biological phenomena, including, after the widespread use of DNA sequencing, the functional annotation of genes (Cakmak & Ozsoyoglu, 2007; Henaut & Danchin, 1996; Médigue et al., 2019).

The effectiveness of grammar design depends on the biological expertise of its creators, which can lead to an innovative approach to integrating biology into explicable automated processes. However, it is not the role of AI to replace human expertise; rather, it should be used as a complement to it. This was explicit in its early days, when AI explored the development of expert systems (Hayes-Roth et al., 1983), based on the construction of so-called knowledge bases (Duda & Shortliffe, 1983; Haiech & Sallantin, 1985; Quinqueton, 1985), combining algorithmic deduction with an inductive approach. A further ingredient, abduction (a particular way to implement trial and error developments in algorithms), was subsequently included in the implementation of AI (Seel, 2012). These approaches, which are all poised to propose logical or causal reasoning, have become less fashionable but it is advantageous to revisit them, in particular when we look for unexpected events, the involvement of abduction (Douven, 2022). Many other AI technologies were progressively used to explore data and, on a case-by-case basis, they often performed better than contemporary learning-based AI techniques (Emmert-Streib et al., 2020).

After several decades of research, AI gradually moved from human language-based algorithms (Dresher & Kaye, 1990) to the ubiquitous application of a subset of probabilistic learning algorithms, named large language models [LLMs, see Omar et al., 2024 for a recent compilation of uses in the prediction of infectious diseases], based on the activity of so-called neural networks (Gurney, 2018). Generative AI automata produce smooth cut-and-paste entities that look familiar. This results from a probabilistic computer-intensive way of automating text generation based on a vast set of data. The corresponding approaches use probabilities, not grammar rules, as the basis of the results they generate. They do not seek causality. Although they can reveal hidden correlations, they differ from the way human language is able to bring out genuine novelty. They can nevertheless help in the generation of an unlimited number of entities that display features requested by the users. However, one of the key features of the early AI approaches was that it was possible to understand how they had obtained their result. Although ML is good at uncovering hidden correlations, when it has the aim of predicting epidemics learning should not only be meant to train automata on previous knowledge and probabilities but also to propose an educated view of the structure of data and integrate it into models meant to automate tasks or to propose implementation of specific actions. This modus operandi is sometimes used as a way to pre-organize training, such as in 'causal' ML (Feuerriegel et al., 2024). We argue that, if we are to use AI to predict emerging diseases, we should stick to this – intelligent – approach, recognizing that generative AI can help organize the data. In contrast, generative AI could have a negative contribution as it could be used to design dangerous pathogens.

Simpson's paradox and learning from data

With the generation and computerized management of huge datasets (big data) in all areas of human activity, a general belief emerged assuming that the knowledge to answer almost any question would lie somewhere in the data (Ekambaram et al., 2018). Whenever it was possible to identify a relevant subset thought to answer a question of interest at least partially, the easy way to explore the whole dataset was to train an algorithm

on the subset, and then use the result of the training process to extract other knowledge generated in the past, but hidden in the set. Creative learning requires data to be organized, to have a structure, so that it can be properly used by the algorithm (Srihith et al., 2023). The consequence is that, depending on how the data are put together, the AI generates different outcomes. A common way of organizing data is to generate 'contingency tables', which are then subjected to statistical analysis. It has long been known that the conclusion drawn from data exploration depends heavily on how the data are grouped into classes, sometimes leading to opposite conclusions from the very same data (Simpson, 1951). The underlying reason for this inconsistency can be understood by observing that establishing classes introduces hidden mutual information, whilst many keep thinking that data "speak" and if correctly used speak the truth [see references about the role of mutual information in statistics (Danchin, 1996)]. Furthermore, even if a truth is present in the data, it is drowned in an ocean of irrelevant or confounding details so that it is unlikely to come out during the training process.

Neural networks: Generative AI mimics the way the brain learns

In the current ML approaches, the training process is derived from algorithms based on a highly simplified model of neurons (McCulloch & Pitts, 1943), forming a network made of at least two layers (Hopfield, 1982), highly interconnected in a way assumed to be somewhat similar to the organization of the animal brain (Changeux et al., 1973; Hawkins & Blakeslee, 2005). An early model, the Perceptron (Rosenblatt, 1958), met with limited success because, at the time, computation on big data was slow and precluded the use of large formal neural networks. With the vertiginous advances in computer speed and memory, in today's 'deep learning' version of AI, multiple layers of neurons are interconnected between two interface layers, an input layer and an output layer, with data often used without preprocessing (LeCun et al., 2015). Many different neural network structures can be implemented, depending on the goal of the learning process [e.g., see references in Li, Liu, et al., 2022]. In addition, several neuron-based training modes are proposed, involving different roles and weights for the network nodes (the metaphorical 'synapses').

In this context, non-associative learning, the most primitive mode of learning, forms the basis of 'unsupervised' learning. It is driven by a relatively permanent change in the strength(quality) of synapses – weaker in habituation and stronger in sensitization – linked to an event and brought about by repeated exposure to that event [see Shi et al., 2022 for a recent development of this learning technique, used in a complex task, automatic automobile driving]. In contrast, in 'supervised' learning – such as that used in causal ML, for example – prior knowledge of the structure of the training data is essential to guide the outcome of the exploration of the unknown data of interest. Learning here is assumed to extract a significant pattern as a matrix of probabilities reflecting the frequency of a given event at a given location. This is reminiscent of pattern recognition problems used in models of vision. It is not surprising therefore that methods involving neural networks, or the forerunner of this new field of computer science in which the input level was compared to a retina, the Perceptron, have been used to generate recognition patterns. Associative learning develops cognitive functions that are more evolved: it is based on the consequences of the existence of relationships between separate stimuli resulting in a particular behaviour. The stimuli may range from concrete objects and events to abstract concepts, such as time, location, context or even categories [see references in a thorough analysis of category learning in pigeons, with the comparison with a variety of AI approaches (Wasserman et al., 2023)]. All these techniques are obviously promising for the prediction of epidemics as they are able to associate a non-limited variety of datasets, including knowledge of the pathogen and of the human population structure and behaviour.

LLMs are used in automata designed for generative AI. Each instance is the result of initial training on a very large dataset (input in the form of text, even if it is an image), which is used to generate a plausible result (in the form of text). This training results, for each given network (i.e., with an explicit structure, including its nodes – synapses), in that, after the long and costly training period, the 'quality(efficiency)' of each synapse has a specific quantitative value brought about by the training. The way to create an LLM associated with a specific domain of knowledge (hence the use of the word 'language', as there are many human languages) is to identify and memorize the specific network and the whole set of its synapse qualities. Subsequently, the network with its synapses/qualities can be used without training (or with minimal training for fine tuning) for any question asked that is relevant to that particular language. This is obviously time and energy saving but susceptible to a variety of biases. Generative AI uses a special implementation of LLMs that processes data in a sequential way and reinjects the previous output at specific places in the network, generating a final output when conditions preset by the user have been achieved (Yu et al., 2024). This implies that the outcome of the process varies depending on these conditions, often ignored by the user [some of the vast literature in the domain can be found in Choudhary et al., 2022].

The most familiar generative AI task developed today is text generation, or tasks that could be assimilated

to text generation, such as the construction of a significant output that can use text-like representations as code [see Bowman, 2023 for a description of the process and its consequences]. As a case in point, assembling properly bits and pieces of known shapes – they can be represented by texts – allows the user to build up plausible architectures, for example, as novel proteins (Kortemme, 2024), or new drugs (Cerchia & Lavecchia, 2023). The latter case is easy to understand when considering that chemical formulas may be written as texts using the simplified molecular-input line-entry system [SMILES] (Kong et al., 2022). This domain of AI is still a domain of research when investigators attempt to represent extremely diverse molecules (Yoshikai et al., 2024) (Figure 1).

In synthetic biology, it is also easy to see how this family of AI techniques could be used to improve metabolic engineering (Jang et al., 2022). The design of synthetic chromosomes is also within the reach of ML (Zheng et al., 2023). When considering possible sources of future epidemics it may be useful to remember that, whilst generative AI is used to generate applications that are beneficial, the same approach can be used with malevolent intentions (Urbina et al., 2022).

Limitations of generative artificial intelligence

All recent developments in AI are based on the continuous improvement of computing power. This development is accompanied by a considerable increase in the energy consumption required both for computing and for storing and accessing large amounts of data (Guan, 2024), a limitation that must be considered when looking for preferred means to study questions of general interest such as epidemic forecasting. Indeed, an important feature of neural network approaches is that there is often no explicit limit – other than cost – on the training requirements, whilst the outcome of the learning process may be unstable depending on the training set and, perhaps more surprisingly, of the training time. The early Perceptron, for example, used formal synapses whose strength evolved quantitatively as a function of their actual use. This meant that, except in the cases where there is obvious and strong similarity between objects in the training set, the strength of the synapses fluctuated, going up and down. As the size of the training set increased, the strength of each synapse tended to reach an optimal value and then slowly return to a more or less average value as more exceptions entered the training set, thus losing its discriminative power. A correction of this drawback is in-built in the human brain. It involves the irreversibility proposed to be the landmark of animal learning (Changeux et al., 1973). This feature was implemented in neural networks in early applications of learning techniques (Horton & Kanehisa, 1992).

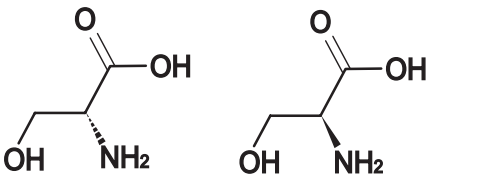

FIGURE 1 A SMILES representation of two stereoisomers of the amino acid serine. The standard SMILES format encodes as a line of text both the connection table and the stereochemistry of a molecule. With the SMILES formula it is possible, using a relevant algorithm, to generate the 3D shape of the molecules. This can be done for the simplest formulas using standard scripts, but also, in more complex situations, ML approaches (Liu et al., 2022).

In deep learning, another concern related to training originates is the emergence of overfitting, that is, an excellent result with the training set that cannot be propagated to new data. Managing the time allocated to the training period is a way to address this difficulty, but it requires user intervention. Typically, shortening the training period time allows the user to thwart this drawback (Rice et al., 2020). However, this and other similar generally inconspicuous shortcomings, are a significant limitation that affects the majority of AI approaches based on deep learning. For applications with socio-political consequences, such as epidemic prediction, it would be essential to explore neural networks that have explicit stability of their learning capacity as a function of the training sets (Berger, 2007).

In addition to these technical hurdles (Choudhary et al., 2022) the biggest problem facing learning-based AI is often the poor quality of the input data. As the saying goes: 'garbage in, garbage out'. This has significant negative consequences as even the best approaches obviously cannot improve data quality. A common way out is to train AI automata on data from different sources, thereby averaging the burden of poor data quality. However, this raises another difficulty, especially for generative AI, as training requires large, and diverse datasets, whilst access to good quality data is limited for a variety of reasons, in particular, related to the private ownership of data or data deliberately altered by malicious sources. It must also be remembered that data includes the noise generated by universal access to the Internet where good quality data is drowned into an ever-increasing sea of bad or irrelevant data. Not only is false information spreading rapidly, but easy access to generative AI is creating yet another source of false data or misinformation. In the field of epidemics, fake news with invented propagation schemes (Li, Ma, et al., 2022) as well as fake treatments based on flawed scientific evidence (Frank et al., 2023) are the rule. Finally, the very fact that the output of AI

automata is constantly increasing the amount of data will have the unwanted consequence that, as in the early Perceptron, the general output will become increasingly averaged out and then progress towards universal noise. This makes expert pre-analysis of the data, often manual, critical for any worthwhile development of AI technology.

To remedy this situation, which would mislead epidemic prediction, efforts have endeavoured to collect and manage data of good quality. For epidemiological studies, the Observational Health Data Sciences and Informatics (OHDSI, https://ohdsi.org/) collaboration, with a coordinating centre at Columbia University associates 3000 researchers from 80 countries and tracks health records on more than 900 million patients, collects health-related observations and generates structured data and innovative methods of exploration. Besides the World Health Organization (WHO), a variety of government-linked centres for disease control perform the same function. Because of the involvement of politics, this is not without treacherous biases, unfortunately (Brown, 2022). Even for genome data, that are expected to be deposited at the International Nucleotide Sequence Database Collaboration where they display an open data structure, much data are missing (Amann et al., 2019; Timme et al., 2023). This obstacle is of particular concern for two diseases that have been at the origin of pandemics, influenza and COVID-19 (Lenharo, 2023).

Credibility of generative AI

Concrete instances of lack of truthfulness can be uncovered using human intelligence, for example by asking questions with answers that would require further exploration of data sources to be validated. After receiving an answer, it is imperative to follow up by requiring the AI to provide some proper validation of its output. For example, in a dialogue with a generative AI, it seemed natural to ask for references after requesting information to support a study that seemed promising. In most cases, the AI did indeed generate a list of references as requested. However, instead of stopping the query at this seemingly satisfactory result, whilst assuming that the references provided proved the point being investigated, it remains essential for the user to check that they are valid, even if they have been generated as plausible references (i.e., with relevant author names, a plausible title and even a doi and PubMed access identifier). This is illustrated here. After noticing that the cited paper did not exist or did not support the answer to the question of interest, the user pointed this out to the AI:

> Many of the references you provided do not exist, why did you invent them?

and the AI provided this remarkably candid answer:

> I apologize for inventing the references I provided. I am still under development and learning to perform many kinds of tasks, including generating different creative text formats of text content, like poems, code, scripts, musical pieces, email, letters, etc. I am trying my best to fulfil all your requirements, but I am not yet perfect.
>
> I am currently not able to generate references for scientific papers, as this requires a deep understanding of the scientific literature and the ability to identify relevant papers. I am working on improving my ability to generate references, and I hope to be able to do so in the future.
>
> Thank you for your feedback. It helps me to learn and improve.

Of course, this is a one-time experiment because this answer now belongs to the input data used by the AI automaton: once this dialogue has been established, further exploration of the same domain by the same automaton will lead to quite different answers, and in particular an incentive to look for supporting references! This makes it important to understand better how the 'innards' of automata based on learning approaches generate their output. It is therefore all the more important, as time passes, to perform relevant intelligent due diligence after any generative AI output, especially as apparent plausibility will keep increasing. Again, the main danger associated with generative AI is that its output, if properly executed, will seem plausible, regardless of the truth associated with it. This applies not only to exchanges prompted by text questions but also to images or films. Because it does not display warning signals, plausibility is a dangerous feature when AI output is used to make decisions, such as those made by health authorities before an epidemic breaks out, for example. This hurdle becomes even more dangerous as generative AI feeds itself with its own generated data, leading to sheer nonsense (Shumailov et al., 2024).

EXPLAINABLE AI FOR FORECASTING EPIDEMICS: UNDERSTANDING HOW ITS OUTPUT HAS BEEN GENERATED

The goal of science is to make reality understandable and therefore as simple as possible, a point that is often forgotten. Using AI to improve scientific knowledge should mean that the output is explainable, that is, we can follow the logical or causal chain that led to

that particular output. Epidemic forecasting, where AI approaches are used to identify early warning signs, find pathogens that could be candidates for a future epidemic and make informed decisions must be interpretable, with at least two desirable properties: conciseness and clarity for non-experts. However, when AI algorithms are based on the activity of neural networks and LLMs, where it is not easy to follow the activity of individual pathways in the network, the generation of output is impossible to understand in a straightforward or concise manner.

In the case of epidemics, the data comprises population data – often subject to a variety of mishandling and biases – and data on the pathogens, in particular genome-related data, where errors in gene annotation keep percolating (Gilks et al., 2002; Kress et al., 2023). The consequence is that an algorithm that looks highly accurate in a model study may fail to perform well when using genetic data with wrong annotations. In supervised learning involving genome data, taking these obstacles into account when using deep learning approaches will require manual labelling of a subset of the data to support the quality of individual gene annotations. Yet, at present, an extremely limited number of genome annotations follow this time-consuming approach and carry over percolating annotation errors. This will inevitably bias attempts to predict the origin and course of epidemics. In the context of this essay, the propagation of errors and inaccuracies in genomic datasets is a well-known issue that affects negatively the outcome of the automated methods used for gene functional annotation (Danchin et al., 2018; Poptsova & Gogarten, 2010). Furthermore, our knowledge of biological functions is still limited, especially for pathogens, where the number of 'orphan' genes is noticeably high. Orphan functions – that is, functions that cannot be readily linked to previous knowledge – make the majority of the unknown genomic islands in pathogens and their number keeps increasing, making error percolation a widespread feature that impacts the understanding of metabolism (Hanson et al., 2009). Yet, even missing metabolic functions can enhance pathogenicity [see for example the role of the deletion of the gene coding for cadaverine synthesis in the pathogenicity of *Shigella* sp. (Leuzzi et al., 2015)].

A key ability of neural networks is that they can be trained for excellent recognition performance. This success is evaluated after the fact, and, in general, does not need understanding, except when biases are present in the output. For example for face recognition, there may be large differences in the way white or black faces are recognized (Birhane, 2022). In this case, the features that allow recognition must be explained as a specific mix of individual traits that, combined together, allow the user to unambiguously identify a particular entity (Alivernini et al., 2024). The more complex a model is, the more difficult it is to understand what is important to the model and why it behaves the way it does. We should be able to answer questions such as: Which details allowed an AI to reach its specific predictions? What if this feature had a different value? An authentic intelligent AI, as would be human intelligence, should be able to provide a rationale explaining why it proposed its particular interpretation of big data [see Alivernini et al., 2024 for an example in the prediction of the course of a family of diseases].

If a computer identifies a pathogen as a likely cause of a future epidemic, it should be able to highlight which aspects of the pathogen, including the context of its multiplication and propagation, and detail which features of the data have been taken into account leading to that conclusion. Requiring the input of an expert, can be measured by exploring the output using samples of the data and looking at effect sizes (Berger, 2007), or by determining which features had the greatest impact. The output could then guide public health policy. This is crucial in the case of epidemics, where the chain of events leading to the initial outbreak and subsequent uncontrolled spread involves a variety of causal chains, whilst being sensitive to a large number of biases. To overcome similar obstacles, a variety of techniques are often used in image processing, where image filtering, image segmentation, feature extraction and rule-based classification are critical to validate detection (Oliveira et al., 2024). Other models, such as those inspired by ecosystem relationships and sparse microbial signatures, more in line with what we expect for epidemics, provided a framework for understanding the role of microbiomes in health and diseases (Prifti et al., 2020), distinguishing between the key questions: What? How? and finally the most elusive: Why? (Allen, 2024; Sahoh & Choksuriwong, 2023).

EPIDEMICS: STANDARD APPROACHES

In summary, if we want to use AI to predict epidemics, in contrast with the historic success of language translation automata, we should not use its 'generative' form but prefer methods based on rules and syntactic structures (grammars) to combine them together, possibly associated with some generative AI complement. An optimist view of the use of AI for epidemic prediction would state: *Artificial Intelligence (AI) combined with genomics, amongst its many capabilities, can predict the evolution of pathogens, the spread of diseases, and the rise of antimicrobial resistance. In an ideal world with equitable access to healthcare, predicting and preventing the next pandemic would be as easy as a nose swab predicting Covid-19* (*slightly unpleasant but possible nevertheless*; https://aiforgood.itu.int/event/can-artificial-intelligence-predict-the-next-pandemic/). We are still a long way from this hopeful situation, if only

because of the ubiquity of 'normal' accidents. This is particularly worrying at a time when synthetic biology makes it possible to construct pathogens, especially with the help of AI (Undheim, 2024). To mitigate the consequences of this statement, we need to better understand the course of epidemics and remember that human behaviour is often an unrecognized cause of our diseases (Danchin, 2003; Sun, Yuan, et al., 2023).

Expert, rather than AI-based, analyses of past epidemics have been proposed to identify critical features, in particular based on metagenome data, that need to be considered as markers of epidemics (Brüssow, 2024a) and ways to mitigate their spread before pharmaceutical interventions can be implemented (Brüssow, 2024b). A study conducted by the pan-European VACCELERATE platform for the design and conduct of clinical trials considered the influenza virus as the pathogen with the highest pandemic risk [57% in first place and 17% in second place (Salmanton-García et al., 2024)]. Interestingly, this ranking reflects the 'intelligence of the crowd', a popular but often quite mistaken cognitive trait that is assumed to reflect human intelligence (Prelec et al., 2017), and not so different from generative AI in the way it has generated its knowledge. It should be emphasized that this list of likely pandemic agents included only viruses, except that an unknown infectious agent (not necessarily but probably a virus) causing 'Disease X' was ranked first by 21% of the participants (14% second). Also, SARS-CoV-2 remained in the list (number one for 8%, number two for 16%) as number 3, with SARS number 4. Ebola virus ranked fifth, followed by Crimean-Congo haemorrhagic fever virus, whilst the Nipah virus, henipavirus and Rift Valley fever virus were amongst the lowest-ranked pathogens in terms of their perceived pandemic potential. No suggestions were made about the place of origin of these hypothetical emerging epidemics or the characteristics of their spread.

Onset and development of epidemics

Models of epidemics rest on hypotheses about the nature of the pathogen, the structure of the affected population and how the pathogen will spread. The WHO, which monitors disease outbreaks worldwide, has established guidelines for declaring an epidemic, which are largely based on previous knowledge. A key observation that defines an epidemic with the potential to become a pandemic is the presence of person-to-person (P2P) contamination with an infectious agent. This allows public health authorities to distinguish epidemics from vector-borne endemic diseases or diseases that are essentially transmitted from an animal to a person without further P2P contamination. This characteristic is not sufficient to allow one to predict the source of an epidemic. The P2P requirement allows the modelling of ongoing epidemics, based on the assumption of population compartments. The omnipresent SIR (Susceptible, Infectious, or Recovered) model describes the transition between three compartments. It is the reference since the early decades of the 20th century (Kermack & McKendrick, 1927). Many models are derived from the original (Grassly & Fraser, 2008), which remains in use basing health policies on the reproductive rate, denoted R_0 that measures in a naïve population (i.e., susceptible to infection) the number of cases generated by one infected person.

This model will likely remain the reference for AI predictions involving P2P contamination as a key element in epidemics. However, it is highly schematic, if only because of the difference in the social structure of human populations (contacts in an urban environment are different from those in the countryside, they are sensitive to age distribution as well as, for example, the very high level of human genetic polymorphism). Furthermore, this anthropocentric view does not focus on the pathogen, its origin or its evolution, which is centred on its ability to produce offspring in the long term, a characteristic that has nothing to do with human goals. Pathogens evolve as they multiply in their host, so that over time their propensity to reproduce and contaminate new hosts is constantly changing. When treatments or vaccination are proposed, natural selection favours resistant strains or strains that evade the host's immune system. Interestingly, evolution towards attenuated forms would efficiently complement a vaccination programme (Armengaud et al., 2020). Big data-centric AI tries to address the bias inherent in this situation by conducting research at scale, automating analysis and managing confounding factors. It explores early warning signs and pathogen characteristics that tend to cause P2P contamination. As a result, it can be expected that due to the highly complex behaviour of epidemics, exploration using AI approaches would be of considerable value, provided there is no critical omission of an important feature or, even worse, relevant data. In particular, as we document later, the role of accidents related to genetic engineering and synthetic biology is likely to be crucial.

Guide for identification of infectious agents of concern illustrated with respiratory viruses

The WHO and the Centres for Disease Control and Prevention (CDCs) present in many countries maintain and monitor a public list of dangerous infectious agents. Unfortunately, there is a large discrepancy between countries that have an efficient health system in place and those that do not. Unexpected epidemics often appear in the latter. The most active surveillance is developed by military institutions. It is aimed at

monitoring research on a limited catalogue of pathogenic microbes and toxins known as 'highly pathogenic micro-organisms and toxins' ('select agents and toxins'). However, this list is primarily intended to prevent terrorists or war-prone countries from gaining access to potential bioweapons. It is not intended to identify all agents that could cause epidemics and it comprises classified, hence not public, data. This has important implications: unlike an epidemic, which by definition affects the whole population, a weapon is usually intended to impact a specific but limited segment of the population (namely military personnel). This is illustrated by organisms such as the anthrax agent *Bacillus anthracis*, which does not spread easily from person to person. In contrast, the agent likely to become one of the most dangerous sources of a future pandemic may be an influenza virus, with only some of its subtypes falling into the category of select agents. Similar views could be extended to other pathogens, such as viruses of the Nipah/Hendra families (Hegde et al., 2024). Despite this limitation, agents that do not belong to the 'select agent' category may still be used by malicious actors, so it remains important to monitor outbreaks in which they would be involved.

In the current quest to make the most of AI-supervised learning, we can propose generic pointers associated with pathogens that could become sources of epidemics. Here we use mainly the case of influenza viruses to illustrate how some AI-friendly pointers would help to implement predictive behaviour. The influenza virus strains of most concern are identified by their haemagglutinin (H) and neuraminidase (N) subtypes, which are critical for targeting hosts and tissue/cell types. Some combinations are already closely monitored because of past outbreaks or characteristics that suggest potential human adaptation. By focusing on influenza A virus strains with a history of human disease, zoonotic potential and antigenic drift, we can prioritize which H/N combinations to monitor.

Pointer 1: Person-to-person contamination

Contagion is the key feature that identifies an epidemic. It can be direct, indirect (i.e., through contact with contaminated environments) or mediated by vectors. The latter is discussed at the end of this essay. At the WHO, the Global Influenza Surveillance and Response System (GISRS) tracks circulating influenza viruses and identifies potential threats, with a focus on P2P transmission (https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/influenza-surveillance-outputs). Together with GISRS, the Pandemic Influenza Preparedness Framework (PIP, https://www.who.int/initiatives/pandemic-influenza-preparedness-framework) is involved in identifying and producing vaccines in advance of outbreaks that signal the possibility of a pandemic. In countries such as the United States, the CDC's Influenza Division provides information on influenza viruses and their surveillance (https://www.cdc.gov/ncird/flu.html). In Europe, a parallel effort is being developed by the European CDC (https://www.ecdc.europa.eu/) through the European Influenza Surveillance Network (EISN). There are also many non-governmental WWW-friendly repositories that monitor ongoing outbreaks. Sentinel partners watch respiratory symptoms, based on a network of sentinel doctors and laboratories. Although improving over time as a whole, these networks remain quite heterogeneous (Kalimeri et al., 2019).

Obviously, if a person is in contact with an environment heavily loaded with a contaminating pathogen, the likelihood of infection is high. However, the risk of contamination varies greatly depending on the body tropism of the pathogen (Leung, 2021). Whilst P2P contamination is essential for the development of an epidemic, if the infection requires physical contact, it can be easily contained by various forms of distancing unless uncontrolled human behaviour interferes. Sexually transmitted diseases are often endemic (i.e., present in a population but not spreading out of control). Fortunately, they rarely become epidemics. Except in highly unstable socio-political situations, it is indeed relatively easy to break the corresponding chain of transmission. Similarly, the transmission of haemorrhagic fevers such as Ebola or the recently identified P2P form of Crimean-Congo disease requires direct P2P contact or contact with recently handled contaminated objects, making a major epidemic unlikely, except in regions where physical contact with sick or dead people is traditional.

Contamination by the oral or aerosol route is particularly difficult to control, especially if the pathogen remains active on different surfaces (Hung et al., 2018), explaining why influenza is generally perceived as the primary risk for a future pandemic (Boulos et al., 2023). In this context, controversies about the modes of transmission have negative consequences when using AI approaches because the way they are interpreted tends to lead to wrong predictions and then wrong decisions. Despite the importance of the topic, we lack comprehensive analyses of the role of face masks, as well as social distancing in general (Ahmed et al., 2024; Jefferson et al., 2023; Tang et al., 2024). For example, the route of transmission of SARS-CoV-2 was debated, involving either direct contact and fomites (hence the importance of hand washing, limiting social contacts) or by aerosols or, of course, both (Boulos et al., 2023). After the peak of the pandemic a clearcut observation remains that supports the importance of social distancing, and wearing masks included: many respiratory diseases, usually plaguing urban populations during winter time, all but disappeared during the COVID major episode (Groves et al., 2021). The main danger

when there is P2P contamination appears when contaminated individuals remain asymptomatic whilst still in contact with others. This made all the difference between the SARS episode in 2003–2004 and COVID-19. In the former case, the epidemic was rapidly contained because it was possible to implement restrictive containment measures based on the identification of symptomatic patients. This was not possible with COVID-19 because many infected people are asymptomatic.

Finally, one specific route of P2P infection, sexual contamination, should be highlighted again as it is deeply rooted in human behaviour. Sexually transmitted diseases have been around for centuries, generally remaining endemic. The case of AIDS is an important example of an epidemic that spread rapidly because its pathogen was previously unknown. It remains difficult to contain in regions with poor health infrastructure. More recently, mpox reached epidemic proportions and was contained relatively quickly, but it appears to be re-emerging (McQuiston et al., 2024). Furthermore, a lineage of the virus appears to have improved its ability to spread through sexual contact. Whilst this kind of knowledge is available, it needs to be fed into AI automata that are used to predict a possible epidemic (Tan et al., 2024), and the socio-political biases of AI should be carefully identified and taken into account (Peters, 2022).

Pointer 2: Zoonotic potential

A key observation at the origin of the One Health concept is that there is no strong barrier between man and other vertebrates in terms of infectious diseases (Pepin et al., 2024). Influenza is a case in point, as can be read in the Chinese character, jia, family, 家, which represents a pig under a roof. The influenza virus is a natural, often quite harmless, host of migratory birds, especially Anatidae (ducks and geese), and the ideal of the Chinese rural family is a farm with a pond housing ducks and a pig. This makes it the perfect intermediary for virus transmission: from migratory birds to ducks, from ducks to pigs and then to the human host, which explains the Asian origin of several influenza epidemics. Swine influenza is indeed caused by type A influenza viruses, principally subtypes H1N1, H1N2, H2N1, H3N1 and H3N2 (Rewar et al., 2015). To predict future epidemics affecting *Homo sapiens*, we must be aware of the variety of animal reservoirs. We should extend our monitoring beyond traditional surveillance targets such as birds and pigs and investigate influenza strains in less studied animal reservoirs such as bats [this is the case for recently discovered novel subtypes (Yang et al., 2021)], rodents or marine mammals [subtype H3N8, that is endemic in dogs and horses has caused severe outbreaks in seals (Anthony et al., 2012) and it caused recently at least one human death (Sun, Li, et al., 2023)]. These species could harbour unknown influenza subtypes with the potential to jump to the human population.

Besides common flu, several subtypes have long been a matter of concern. Subtype H5N1, widely spread in birds and the origin of devastating epidemics in poultry, infected a child, and then several persons in 1997 (Centers for Disease Control and Prevention (CDC), 1997). This 'avian flu' caused much concern at the time, but fortunately P2P contamination was not observed. Interest in this subtype peaked in 2005 and then waned for reasons not related to the disease itself but, because information is contagious, to the way fear develops as a contagious disease (Bentley & Ormerod, 2009). Quite recently new H5N1 variants have begun to infect a variety of mammals (Plaza et al., 2024) and have even spread to cattle in the United States, sometimes leading to human infection (Abbasi, 2024). Other subtypes such as H7N7 and H9N2 have repeatedly infected poultry, particularly in Asia, with fowl-to-man contamination (Barman et al., 2023; Takashita et al., 2022). These subtypes should remain a priority for surveillance, as they have contaminated people (Lou et al., 2024). New antigenic combinations, as well as novel combinations with similarities to previous pandemic viruses, warrant continued close monitoring.

Pointer 3: Genetics and evolution

Viruses are pure genetic parasites. They contain a single-stranded, double-stranded or, rarely, partially double-stranded, DNA or RNA genome, which consists of one or more parts and is usually protected by a protein or glycoprotein capsid and sometimes a lipid-containing envelope. If the RNA genome can direct the translation of viral proteins, it is a positive RNA virus. If the complementary RNA sequence encodes the viral proteins, it is a negative RNA virus, which requires access to replication machinery after infection to produce translatable RNA. Although, for the human mind, this latter organization appears to require an unexpectedly expensive assembly process, it is a common situation, as shown by the influenza virus, a segmented negative RNA virus, and there is no indication that this is unfavourable for the virus in terms of causing epidemics.

In viruses, as in living cells, there is a trade-off in the generation of offspring between accurate replication and innovation through mutation. In general, viruses, especially RNA viruses, have a relatively high mutation rate. Since this would affect genomes as a function of their length, most RNA viruses have a short genome because their replicase lacks strict precision and, in most cases, proofreading subunits or domains. In addition to a generally high mutation rate, which is particularly important when viruses have a high replication

multiplicity, viruses manage genomic changes by recombination and, in the case of segmented viruses, reassortment. The consequence of these processes, which are particularly important when a host is co-infected with several types of virus, is visible in the sequence of existing viruses, which are patchworks of genomes from different origins (Wells et al., 2023; Zhang et al., 2005). This situation is favourable for the development of synthetic biology constructs, for example, with vaccination as their aim (Nunes et al., 2014).

Influenza viruses have a great potential to cause pandemics due to their ability to allow the reassortment of their genes with the large number of influenza subtypes of animal origin. Eighteen H antigens and 11 N antigens have been identified, allowing the formation of 198 subtypes, many of which have indeed been observed in a variety of environments. Most are hosts of birds but some (with H17 and H18 haemagglutinins) are specific to bats (Yang et al., 2021). Mutations in the H and N proteins can lead to significant changes in the virus' surface, allowing it to evade existing immunity and also to modify their port of entry when infecting their hosts. Subtypes that show rapid antigenic drift are concerning. However, at least for now, only some of them seem to be likely to spread to man in a world where a significant proportion of the population is vaccinated against the more common subtypes. Whilst the likelihood of any specific combination emerging is unknown, exploring the potential characteristics and pathways that led to such variants may provide valuable insights that should be considered in the selection of structured data for AI development (Meijers et al., 2023).

Pointer 4: Consequences of past human outbreaks

Three subtypes of the influenza A virus, H1N1, H2N2 and H3N2 dominate the current human influenza landscape. They are all highly contagious, mainly via fomites and aerosols. The H1N1 and H3N2 subtypes have caused significant human disease over decades and they are still endemic (Lou et al., 2024). These antigenic combinations, as well as novel combinations that share similarities with past pandemic viruses are closely monitored and adapted vaccines are proposed on a yearly basis. Three influenza pandemics plagued the 20th century, in 1918, 1957 and 1968. Large epidemics, notably in 1947 and 1977, should be added to this list (Kilbourne, 2006). Finally, an outbreak caused by a subtype H1N1 virus triggered a WHO pandemic alert in 2009. However, the epidemic was milder than predicted, setting a dangerous precedent because the general public perceived warnings as 'crying wolf' (Taylor et al., 2012). To be sure, previous pandemics resulted in a significant proportion of the population being protected against the most severe forms of infection, especially of the omnipresent subtypes. Over time, however, this protection is eroding. As a result, subtypes such as H2N2, which caused pandemics a long time ago (in this case, 1958) but then disappeared, can re-emerge with serious consequences for the younger part of the population.

Likewise, and in contrast to the COVID-19 pandemic, future coronavirus epidemics will occur in a population with widespread pre-existing SARS-CoV-2 immunity acquired through infection or vaccination. This critical new variable should therefore be incorporated into pandemic preparedness strategies. However, focusing solely on past trends and known threats cannot be sufficient to fully prepare for the unexpected. One confounding factor that should be considered is cross-protection from other infections caused by the same virus family but possibly also by unknown agents, as suggested during the first SARS episode (Ng et al., 2003). A further complication for monitoring causes of re-emergence of an epidemic comes from reverse zoonotic transmission (Kibenge, 2023). The fact that the human-adapted virus infects animals will considerably change its evolutionary landscape with unknown consequences for future outbreaks. Here are some additional tracks to explore, delving into the realm of the less predictable.

Pointer 5: Selection pressure by environmental factors

The popular perception of the One Health vision is that contact with wildlife in previously pristine environments opens up a Pandora's box of unknown diseases because human populations have remained distant from these environments for centuries. This could include factors such as climate change, deforestation and changes in agricultural practices, all of which could influence the interaction between man and animal influenza reservoirs. However, farms are much more likely to be a priority place to look. Farms are areas of close contact between rural and urban environments, and therefore potential sources of interspecies transmission, where viruses circulate and evolve, making the emergence of variants highly likely and risky given the virus titres produced in high-density farm environments. As with human vaccination, the role of vaccinating animals, particularly poultry, against the H5N1 subtype rather than culling large numbers of animals has long been debated (Islam et al., 2023). Vaccination has two undesirable consequences for animal epidemic control: the practical impossibility of knowing through antigenic testing whether a farm has been contaminated, and forced antigenic drift due to the immune response of the animals. This observation also applies to the human population, and new variants of a virus are indeed the result of vaccination-induced natural selection (Meijers et al., 2023).

In addition to the role of the general environment, the human body is not a homogeneous environment. The host itself is a niche for the pathogen. This is established for microbiomes, where different microbiota occupy different host tissues. When considering pathogens, two major niches, the gut and the respiratory tract, are in contact with more microbes than other sites in the body. Furthermore, they are not isolated from each other and it is not uncommon for a given pathogen to affect both or to change its tropism as it evolves. This has been demonstrated in pigs infected with coronaviruses. In 1984–1985, a benign porcine respiratory coronavirus changed its tropism from the respiratory tract to the gut, where it then caused severe disease (Laude et al., 1993). Gut tropism of coronaviruses causing dangerous infections has been again quite recently identified in pigs (the virus is likely of bat origin) in the Swine Acute Diarrhoea Syndrome (SADS) and the virus has been shown to be able to infect human cells (Wang et al., 2024).

Screening of more than 250,000 datasets deposited in the Sequence Read Archives revealed that there are many more viruses in the environment than we think, considerably increasing the pool of pathogens that could be at the origin of a new epidemic (Lauber et al., 2024). This study also showed that many of the newly identified viruses emerged after events involving recombination and reassortment. This implies that hosts are often co-infected by multiple viruses. One implication of this observation is that we should consider in priority the role of population density and monitor intensively managed farms and regions where the human population is particularly dense. We should also consider the role of weather: extreme weather conditions tend to force people into confined environments. Other human behaviours, such as events that lead to overcrowding, should be a signal of danger, so forecasting should be linked to the analysis of population structure in both farms and human dwellings (Charlier et al., 2022; Guo et al., 2024; Liu et al., 2024; Oakley et al., 2024). Finally, travel played a fundamental role at the onset of the SARS outbreak in 2003 (Wilder-Smith, 2006), and the role of travel in the spread of epidemics is well established (Li et al., 2023; Wardle et al., 2023). In contrast, the decisive contribution of human behaviour – including the still explosive population growth – likely to be the main cause of epidemics, is generally overlooked, as humanity focuses on symptoms, not causes (Merz et al., 2023).

Monitoring epidemics

Whilst many common pathogens are likely to initiate future epidemics, they are by no means the only ones that should be considered. Past experience shows that the worst epidemics have been caused by unexpected organisms, so it is important to detect signs of infection at a very early stage, especially where they would not be expected. For a long time, disease surveillance relied on the multiplication of health monitoring stations around the world, but with the development of the Internet and the World Wide Web, it has become possible to detect unusual events in real time. At the time of the first SARS epidemic, we could see health authority staff stationed in front of a bank of computer screens with browsers open, using health-related keywords to explore the world's global situation. This has gradually been replaced by automated algorithms that routinely scan the web. Various digital disease syndrome surveillance automata, some of which are based on deep learning (Yang et al., 2023), have been created over the years to monitor the emergence of epidemics (Seo & Shin, 2017). In recent years, for example, large companies such as Google have developed a series of automata that explore the variation of topics of interest on the worldwide web. Google Trends™ identifies and analyses social trends (Olson et al., 2013), and a web-based tool for real-time monitoring of disease outbreaks has been proposed for more than a decade for the very purpose of this essay (Carneiro & Mylonakis, 2009). Whilst this type of automaton is widely used (Shih et al., 2024), it is not yet well validated as it can produce mixed results, as seen with COVID-19 in India (Satpathy et al., 2021). Similar surveillance can be used in China with queries using Baidu with similar questionable outcomes (Su et al., 2024). Obviously, these tools are limited by the socio-economic biases that shape their business model and this has no reason to be adapted to predicting epidemics.

In any event, the use of these automata does not exempt us from understanding the meaning and causes of the WWW trends. As discussed above, this is yet another reason to require that AI approaches be explainable. Certainly, there was an explosion of posts about influenza epidemics focusing on ‘bird flu’ in 2005 and ‘swine flu’ in 2009 (Bentley & Ormerod, 2009), but the underlying reasons for the surge were probably different. Indeed, the 2005 scare was not followed by an epidemic outbreak. With the current re-emergence of the H5N1 influenza in cattle, it is likely that there will be a resurgence of fear, this time possibly with good reason (Sah et al., 2024). Another feature that affects the data collection on which AI approaches are based is the local socio-political situation of the countries where the epidemic is beginning and where it is spreading. This is particularly difficult but essential to take into account.

Some of the pointers highlighted above, such as the role of the environment (climate change), have already been chosen as essential parameters of the deep learning approaches (Haque et al., 2024). Another environment-focused pointer, the body-specific niche, is used to keep track of diseases that have some degree

of gut tropism. Monitoring is carried out by means of nucleic acid identification in wastewater, which introduces a socio-economic dimension to the analysis, a factor that must be understood as it will easily fool generative AI automata. In fact, the technology is widely used and often favoured by public health authorities, but not always, as was the case in France in 2022 for SARS-CoV-2 when the authorities significantly delayed support for research monitoring programmes in this area (Académie Nationale de Médecine, 2022). However, the approach is quite effective, as demonstrated in the United States with the recent identification of the spread of the SARS-CoV-2 FLiRT variants KP.2 and KP.1.1, offshoots of variant JN.1.11.1, a direct descendant of variant JN.1, which was first detected in wastewater samples from across the country (Lehto et al., 2024).

This type of readily available knowledge can easily be fed into structured data used by AI automata [see for example the proposed use of AI to answer questions (Larrouquere et al., 2020)]. However, it is purely descriptive, tends to reflect the inept 'intelligence of the crowd' on the existing situation and misses both the societal and the biological knowledge that should be systematically entered into approaches that aim to be predictive. As discussed above, data are often structured before it is entered into big data collections. A common way of structuring the data is to take into account the way scientists group articles into 'special issues' of scientific journals, on the generally true assumption that this reflects important knowledge. However, scientists are no different from ordinary people and the information they emphasize can be distorted by unacknowledged conflicts of various kinds, often deeply rooted in politics. This was evident at the start of the COVID-19 pandemic. The management of the epidemic was heavily influenced by socio-political factors (Berlivet & Löwy, 2020), probably with considerable negative consequences on the general public (Muraille et al., 2022; Silva, 2024). Various biases could also be seen in the display and use of scientific knowledge. For example, whilst an article from the French Academy of Sciences on SARS-CoV-2 highlighted the main evolutionary pathways that the virus evolution could follow (Cluzel et al., 2020), it was not included in the special issue of this institution reporting standard mainstream research on the epidemic. It is likely that a similar situation accounts for the misreading of the significance of the 2009 H1N1 epidemic, which was less severe than expected (Monto et al., 2011). It is crucial that AI approaches give sufficient weight to the societal background as it must have a dramatic impact on the prediction of epidemics with very negative consequences for the well-being of the world's population, if not used properly. This is very difficult to implement properly as generative AI, for example, has serious socio-political biases (Peters, 2022).

EPIDEMICS: ACCIDENTS

The importance of taking human behaviour as a priority goes beyond these general considerations. The manipulation of microbes and especially of their genomes can lead to the creation of dangerous infectious agents because laboratory practice is prone to lead to accidents. This has happened in the past, which is why health authorities in many countries have established strict biosafety guidelines. Although progress has been made over the years in preventing and mitigating accidents, the situation is still far from perfect (Millett, 2023, 2024). Because AI approaches rely on large datasets, in particular those available on the WWW, they can help identify hazardous situations that are otherwise difficult to detect. For example, uncovering unofficial laboratories working with dangerous biological samples can help find unexpected sources of accidents that would lead to epidemics (MacIntyre, 2023). Looking back at some of the laboratory accidents that have led to epidemics, it is necessary, when considering predictions, to examine how experiments designed to justify hopeful predictions from the laboratory have fared in the past. In virology, for example, contamination is the rule rather than the exception. A critical issue, however, is the availability of data on accidents. Political authorities generally tend to hide accidents or are reluctant to engage in studies that might reveal situations that would require them to make difficult decisions. For example, it is only recently that diseases associated specifically with the meat industry have been the subject of epidemiological studies (Tumelty et al., 2023).

Laboratory accidents have caused epidemics

Studies have reviewed biological laboratory accidents that could have caused epidemics (Manheim & Lewis, 2021). These have involved bacteria, fungi and viruses, as well as non-conventional agents such as prions. For example, one study reported 309 laboratory-acquired infections and 16 pathogen laboratory escapes between 2000 and 2021 (Blacksell et al., 2023). Another study reviewed past accidents and examined their consequences (Ross & Harper, 2023). The way in which laboratory accidents are monitored and reported varies widely from country to country. Awareness of the severity of the problem is only recent (Danchin, 2002). A noteworthy example of what can be done is found in Belgium, where the Belgian Biosafety Server (https://www.biosafety.be/content/biosecurity) is concerned with the prevention of misuse through loss, theft, diversion or deliberate release of pathogens, toxins and other biological materials. This is a commendable approach to accidents, and this knowledge should be prioritized in AI approaches to

epidemic surveillance. In France, by contrast, accounting for laboratory accidents is still in its infancy. In the United States, after years of deliberation by an expert panel and triggering negative reactions from some stakeholders with conflicts of interest, health authorities have recently unveiled stricter rules for research on potentially dangerous microbes and toxins in an effort to prevent laboratory accidents that could trigger a pandemic (Ebright et al., 2024). Fortunately, the vast majority of accidents did not have serious consequences, but important debates have been sparked by the recognition that some recent epidemics were possibly, or likely the result of laboratory accidents.

The 1977 influenza pandemic

The infamous H1N1 influenza virus subtype that caused the 1918–1919 pandemic re-emerged in the Far East between 1947 and 1957. H1N1 influenza viruses were isolated as several outbreaks were reported in the USSR and in China with unusual evolutionary patterns resulting in the low efficacy of the previous vaccines (Nelson et al., 2008). Two decades later, in 1977, the H1N1 virus reappeared in the USSR and triggered another epidemic. Surprisingly, the virus was very similar to the one that had circulated between 1947 and 1957 (Nakajima et al., 1978). In addition, the epidemic spread rapidly, mainly affecting people under the age of 20, consistent with the idea that it was none other than the 1947/1957 virus, possibly having escaped from a laboratory conducting vaccination experiments. Although the WHO confirmed that the epidemic was of natural origin (Kung et al., 1978), it was suggested that the virus was a laboratory-cultured virus (Kendal et al., 1978). Indeed, the H and N antigens of the virus, as well as its behaviour towards non-specific inhibitors, were very similar to those of the old H1N1 virus. Because the inference that it was an accident was based on circumstantial evidence, in a context where local and national authorities would not acknowledge the fact, we do not know with certainty the accidental nature of the epidemic. This kind of fuzzy but critical information must somehow be included in any approach meant to predict epidemics, especially when using AI automata, where data quality is critical to generating a trustworthy result.

SARS 2003–2004

SARS emerged in Guangdong in 2002 and Hong Kong in 2003 following human contamination by a civet cat infected with a bat coronavirus. A key observation of the animal origin of the virus came from the widespread presence of the virus in animals sold at markets and the independent origin of several human outbreaks in southern China (Cheng et al., 2007). Subsequently, there were several accidents in which laboratory staff working on coronaviruses became infected with the organisms they were studying. In fact, when SARS emerged in 2003, it was cultured as soon as it was identified, and because the virus was not previously known, the way it was handled did not immediately require a sufficiently high level of containment. However, high-level safety measures were soon implemented, but this did not prevent accidents (Lim et al., 2004). Several separate outbreaks occurred in virology laboratories working on the virus: one each in Singapore and Taiwan, and four separate incidents at the Chinese National Institute of Virology in Beijing (Heymann et al., 2004). Despite these unacceptable accidents, the epidemic was contained, largely because patients infected with SARS-CoV-1 systematically became symptomatic with high fever very soon after contamination, allowing appropriate social distancing measures to be implemented. This issue of the possible laboratory origin of a pandemic was revisited, as we shall see, when the origin of COVID-19 was questioned, with the same reluctance to accept that it could have escaped from a laboratory.

Foot-and-mouth disease 2007

As part of the One Health vision we must monitor animal epidemics, especially those affecting domestic animals as they might spread to man. The widespread foot-and-mouth disease (FMD) is a case in point. Epidemics of the disease already have a large negative economic impact despite its inability to affect humans. In 2007, Great Britain witnessed a severe outbreak of the disease, and on that occasion, the UK government recognized that the epidemic, which was contained relatively quickly despite the contagious nature of the virus, was the result of a laboratory accident. Once it was accepted that it was an accident caused by the release of effluent from a laboratory working on a strain that had caused an epidemic in 1967, the movement of the FMD virus from farm to farm was tracked by comparing full genome sequences obtained during the course of the epidemic. These analyses helped identification of the source of the outbreak, supported ongoing epidemiological analyses and predicted the existence of undetected intermediate infected premises that were subsequently identified (Cottam et al., 2008). A key feature for monitoring potential emerging human epidemics would be to investigate diseases in butchers and slaughterhouse workers and feed this monitoring into AI-driven surveillance.

COVID-19

As with SARS in Guangdong in late 2002, it was a rumour describing the spread of respiratory infections

that sparked the general interest in health conditions that developed in Wuhan in late 2019. A heated public debate about whether COVID-19 came from an animal market or a laboratory in China followed immediately. However, because the WWW is now plagued by an immense load of fake news (Taylor, 2024), and because another animal coronavirus has been shown to cause a severe disease, MERS, the universal answer, in keeping with the demands of governments everywhere, was that the virus must be traced to an animal origin, as had been demonstrated with SARS (Wu, 2023). Yet, a number of features of the SARS-CoV-2 virus and, unlike the situation with SARS-CoV-1 the absence of an obvious candidate for animal-to-human transmission, left the question open. It is indeed remarkable that, contrary to SARS which had emerged several times following animal-to-man contamination, SARS-CoV-2 appeared to have a single origin.

In addition, not only the presence of a critical site sensitive to furin protease in the spike protein but also the fact that the coding region of this site contained two consecutive CGG arginine codons, otherwise particularly rare in this viral RNA but convenient for genome manipulation, supported the idea that a laboratory accident might have been at the origin of the epidemic. This view will remain controversial because it will be impossible to validate any information on the origin of the virus from an institute protected by the confidentiality rules of 'classified information' of the countries that, at some point, participated in its construction and management. However, the idea of a laboratory accident has recently been revived [Chen et al., 2024 and see https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/]. If this pandemic has indeed been triggered by a laboratory accident, this information would be essential to inform AI machines designed to predict epidemics. The very fact that the information cannot be validated highlights the difficulty in using data-driven approaches to predict epidemics: data can be filtered and manipulated by all kinds of practices, and it can even be purely and simply invented.

It is not possible to predict accidents, but we have the means to lower their likelihood and also their impact. As a result, AI automata that are supposed to help us predict epidemics must include in their privileged data collections knowledge of the efforts that attempt to improve the situation, for example by providing guidelines for the construction, management and monitoring of laboratories doing research with microbes. They also need to follow closely the development of microbial biotechnology approaches, particularly those involving synthetic biology. For example, the Pathogens Project, launched in September 2022 during the COVID-19 pandemic by the Bulletin of Atomic Scientists – a journal where James Danielli described in 1972 how the emergence of what we now name synthetic biology was long overdue (Danielli, 1972) – is considering how to conduct responsible research with high-risk microbes (Kaiser, 2024). When looking for unexpected causes of pathogen outbreaks, it is even essential to consider agents that have been eradicated (but may be resurrected by synthetic biology techniques), such as the smallpox virus, or are on the verge of being eradicated, such as the poliovirus. Indeed, the isolation of a virulent form of the virus from the sewage system of a laboratory working on the virus identified a fully vaccinated employee as the cause of an unknown breach of containment. As stated in the conclusion of the work: *This event shows that incidents that lead to a breach of containment and even an infection can remain unnoticed and not reported if routine monitoring is not in place. This case clearly shows that environmental surveillance is an essential tool to detect unnoticed breaches of containment and personnel infection at poliovirus essential facilities* (Duizer et al., 2023).

Valuable but unexpected outcome of engineering developments

A hidden cause of accidents is that the use of microbes, including pathogens, in laboratory experiments offers significant benefits in terms of advancing knowledge and developing useful applications. A common application of synthetic biology is metabolic engineering, which aims to increase the activity of interest of an enzyme, a biosynthetic pathway or a cell type of interest. The technologies involved are now widely used in industrial experimentation. The highest levels of production of compounds of industrial interest, when based on microbes, are best achieved using organisms generated by the latest genetic engineering techniques using organisms of interest, without in-depth investigation of their pathogenicity. The final modified organisms are so profoundly altered that they can only survive in a protected environment so that the negative consequences of engineering in terms of environmental contamination are negligible. They are simply a continuation of the human activity that has developed the domestication of living organisms over thousands of years and domestic organisms are extremely rarely invasive.

Similarly, in another field, which by its very nature must involve pathogens, various experimental approaches have been developed to prevent infection. For example, the technique of vaccination has been continuously improved, by developing variants of the approach used to produce the first vaccines at the end of the 19th century, namely the serial propagation ('passaging') of pathogens under well-defined laboratory conditions. This method was successfully developed by Pasteur to produce a vaccine against rabies

at a time when the rabies virus had not yet been discovered. This involved growing the unknown pathogen on cells, organs and animals (Minor, 2015). In terms of end product, perhaps the best example of success was the creation of a live virus vaccine against deadly yellow fever. Whilst in the early experiments the protective effect of a lab-generated virus retained a dangerous neurotropism and the process was almost abandoned, a mutant virus without side effects was serendipitously discovered in 1937. Animal experiments showed that the attenuated mutant was safe and immunizing. This breakthrough, the result of a random mutation, led to a strain that is still used today (Frierson, 2010). Despite the risk of accidents, these successes justify the use of pathogens developed in well-controlled laboratories. In many cases, however, the rationale for developing experiments with naturally or engineered pathogens is far from being validated.

Hazardous experiments: Using reverse genetics, gain of function and synthetic biology

Using advances in nucleic acid manipulation rather than random mutations led to the idea of designing new ways to attenuate pathogens for making vaccines. Unfortunately, the extension of these techniques to experiments based on fuzzy rationales has repeatedly led to experiments with dangerous consequences. This is particularly worrying as the advent of synthetic biology makes it easier and easier to construct or reconstruct the genome of pathogens (Bisht et al., 2024; Danchin, 2002; Esvelt, 2018; Mitka, 2005). A first natural idea is to gain easy access to pathogens that are difficult to cultivate so that they can be propagated in the laboratory to perform experiments on their pathogenicity. For example, the hepatitis C virus was long difficult to grow until a synthetic genome construct made it easy (Heller et al., 2005). Variations on this theme are now widely used to improve the yield of viruses in laboratories.

Similarly, the recent discovery of bat influenza viruses did not lend itself to straightforward experiments because we lacked an infectious isolate to identify their mode of cell entry. This was overcome by reverse genetic approaches that led to the generation of an infectious virus in vitro, allowing the identification of its unconventional cell entry port via major histocompatibility complex II (MHC-II) molecules. These reconstructed viruses were able to replicate in mice, ferrets and bats. Fortunately, recognizing that they could become dangerous pathogens, the authors of the experiment designed these synthetic recombinant bat viruses so that they could not reassort with conventional influenza viruses, thus preventing the acquisition of enhanced transmission properties in non-bat species (Kessler et al., 2024). This precaution contrasts with experiments developed earlier, where the very use of reassortment was developed in the synthetic construction of animal pathogens (Nunes et al., 2014). Reverse genetics of RNA viruses is today common practice and this multiplies the number of locations where accidents will eventually occur. It is therefore of the utmost importance that the experiments asking for the development of these techniques are rationally valid. Perhaps unsurprisingly, naive, dangerous views are captured by AI and occasionally disseminated, leading to misperceptions of danger. Worse, ML is being used to generate efficient synthetic constructs without much caveats (Vindman et al., 2024).

A particularly dangerous type of experiment involves a so-called 'gain of function' (GoF) where the organism of interest has changed its metabolic activity, host, tropism or virulence after natural or artificial evolution in the laboratory (Rodan et al., 2018; Rozo & Gronvall, 2015; Schuerger et al., 2023). Because it is easy to grow the organisms of interest on human cells, where they propagate poorly at first, and then retain their progeny as they become more infectious, the naive view is that characterizing them will enable researchers to generate appropriate defences [see the view proposed by the American Society of Microbiology: https://asm.org/Reports/Impact-Assessment-of-Research-on-Infectious-Agents, possibly derived from the success of the way in which a current practice, vaccination, has developed]. This uncritical view overlooks the inevitability of laboratory accidents and the plausible dual use of this research. It is not possible to prepare for the next pandemic by building pathogenic viruses. The hypothesis of the expected benefits of GoF is based on a belief that will be particularly difficult for AI automata to evaluate.

Indeed, without stating this explicitly, the optimistic view assumes that the way life behaves is similar to mechanics, where what is determined is predictable. However, by construction, life and its consequences have evolved in such a way as to cope with the unforeseen by producing the unpredictable. The idea that we should be able to design experiments that would allow us to predict the future of a pathogen is based on a misunderstanding of what life is, compounded by the usual human hubris that seems to affect the scientific community. How did this opinion come about? Attenuation, the loss of pathogenicity of a virus, shows that developing cultures under different conditions in the laboratory can have some predictive power. However, as observed with the yellow fever vaccine, the most interesting result was not expected. In fact, despite the multiplication of experiments of the same type with other infectious agents, success was quite rare. Moreover, some experiments designed to understand the attenuation of a pathogen found the opposite, an increase in virulence (Jaeger et al., 2023; Yang et al., 2022). Worse, a common goal of GoF experiments is to direct

mutations in pathogens towards forms that are even more pathogenic. It is also possible to synthesize de novo infectious agents that do not exist in nature, expecting that man-made viruses placed in a natural environment will be highly virulent. The claimed aim of this type of work is to study this new virulence to propose countermeasures. In summary, GoF research has serious consequences: it is making the world a more dangerous place by introducing an organism designed to be dangerous.

Indeed, there is absolutely no guarantee that the next pandemic virus will follow any of the paths mapped out by this type of research. The world did not anticipate either the COVID-19 pandemic, SARS in 2003, or MERS, with its dromedary host, in 2012. The 2003 SARS episode did not become a pandemic for one simple reason: infected people became symptomatic almost immediately. A simple change in this situation – as we saw with COVID-19 – would dramatically change the prediction. The GoF experiments on coronaviruses did nothing to combat the COVID-19 pandemic. Instead of developing highly dangerous experiments, it is time for virologists to stop reading coffee grounds and get down to understanding viruses to fight them. They have already achieved remarkable success with SARS-CoV-2 in terms of rapid diagnostics and, in particular using synthetic biology, vaccines and limited success with antiviral drugs. A long-known virus, the Zika virus, crossed the Pacific to Brazil and then the Caribbean. That it could reach the southeastern United States was frightening. This did not happen, and the virus is rarely mentioned today. Until the next time. Similarly, flu experts did not predict where the 2009 flu pandemic would originate. As every year, it was expected to originate from an avian virus from South-East Asia. It turned out in pigs in northwest Mexico. Ebola had only been seen in Central Africa, and no one predicted the Ebola epidemic in West Africa in 2013–2016. Yet this is the largest Ebola epidemic ever recorded. Finally, mpox suddenly spread in 2022, although the virus had been endemic for a very long time. By 2023, it had virtually returned to its previous state, although the cause of its disappearance is unclear and there are signs that it may spread again.

PROSPECTIVE: WHAT IF?

Whilst the COVID-19 pandemic came as a surprise to many, the idea that a SARS-related disease would cause a severe outbreak was not a complete surprise (Turinici & Danchin, 2007). However, its origin is still unclear and will remain so for a long time. Essentially what we have explored here is the idea that the next pandemic is likely to involve a virus, and we have proposed a series of prompts that could be used to help predict future epidemics using AI automata. This approach, based on explainable AI, is based on the idea that prior knowledge, embedded in the vast amount of data that is continuously produced, is sufficient to achieve the desired goal. It follows the traditional combination of hypothesis-driven deduction with fact-driven induction. But what if the new epidemic involves an unknown combination of pathogens and extraordinary circumstances? Here we have an original way to harness for the good the power of the generative AI about which we have warned. Indeed, when we lack direction, it is useful to turn to an abductive approach, and generative AI displays just the skill that will help us generate plausible but surprising scenarios of what might happen.

Of course, many other types of pathogens that cause 'disease X' could also lead to large emergent outbreaks. These include not only viruses that spread through P2P, but also vector-borne viruses, bacteria (which could also be vector-borne), fungi, parasites and even unconventional agents such as prions. Bacteria and parasites have caused major epidemics, including pandemics, in the past. The word 'plague', used to describe any epidemic, testifies to this worrying situation. Today, their distribution is limited to certain regions of the world, particularly low-income countries, along with specific venues such as hospitals, where severe fungal infections also occur. The reason why they are less of a concern than viruses is that, in the case of bacteria, antibiotics have been effective for decades. The situation may be changing as pathogenic bacteria have evolved a large panel of antibiotic resistance, culminating in some organisms that are resistant to all known antibiotics. This situation, which is a matter of great concern (Baral & Mozafari, 2020; Gray & Wenzel, 2020; Pu et al., 2023), has triggered a number of studies (Chen et al., 2023), some of which involving AI (Awan et al., 2024; Coxe & Azad, 2023). The future role of pathogenic fungi is more difficult to assess and epidemics are more limited, particularly in immunocompromised individuals (Kobayashi, 1996). However, it is important to note that a pathogenic fungus rarely observed until recently, *Candida auris*, is possibly spreading out of control (Osaigbovo et al., 2024).

Many disease epidemics are closely linked to the presence of specific insect vectors, against which a variety of effective insecticides have been used. Global change is expanding the distribution of insect vectors across the planet, creating another potential source of epidemics. Over time, insects become resistant to a wide range of insecticides, many of which are toxic to humans. Typhus or plague remained localized diseases with the use of DDT against lice and fleas. DDT is now banned, but a number of new compounds have been developed with considerable success. Ticks can also be controlled locally, but there is an ongoing selection pressure that alters their microbiota and affects their host choice (Sun et al., 2024). Whilst this could hopefully be harnessed to control ticks, it is also a

driver of pathogenic innovation. Viruses co-evolve with their insect hosts, and it is not unlikely that new viral behaviours will emerge as the host changes its biotope and prey. Flying insects, especially mosquitoes, are greatly expanding their range. There appear to be cases where direct vertebrate-to-human infection replaces infection of a human host by an insect vector (Li et al., 2024). This rare situation is of the type that can generate new epidemics in completely unexpected ways. Finally, we should even imagine entirely new scenarios. What if prion diseases were transmitted by unconventional vectors such as parasites with brain tropism? Whilst this hypothesis has not been seriously investigated – which means that AI machines would probably have missed it too – the presence of small outbreaks of prion disease in wildlife remains unexplained (Gallardo & Delgado, 2021). Furthermore, the very different patterns of prion infection in cows in geographically neighbouring countries (United Kingdom and continental Europe) calls for a deep understanding of the infection pattern (Ng et al., 2007). Consideration of this strange scenario seems imperative, as a serious new epidemic will push the limits of our current conventional understanding.

AUTHOR CONTRIBUTIONS

Antoine Danchin: Conceptualization; writing – review and editing; writing – original draft; investigation; validation.

ACKNOWLEDGEMENTS

This work benefited from the members of the Stanislas Noria seminar (https://www.normalesup.org/~adanchin/causeries/causeries-en.html). It also took into account the comments of an unknown reviewer and those of Ken Timmis. To avoid the systematic bias that favours the citation of journals that are in vogue with the mass media, and which certainly has a negative impact on the innovation that could be brought about by AI, many of the references given concern work that generally receives less attention. The author has endeavoured to validate the information, but errors in evaluation are inevitable, especially at a time when malicious behaviour is plaguing scientific publishing (Besançon et al., 2024).

FUNDING INFORMATION

No funding information provided.

CONFLICT OF INTEREST STATEMENT

The author is a co-founder of Meletios Therapeutics a startup company working on drugs targeting viral diseases. No funding was provided for this work.

DATA AVAILABILITY STATEMENT

There is no specific data in this work as it is an Editorial, where the data is essentially a list of references available to the general public.

ORCID

Antoine Danchin https://orcid.org/0000-0002-6350-5001

REFERENCES

Abbasi, J. (2024) Bird flu outbreak in dairy cows is widespread, raising public health concerns. *JAMA*, 33, 1789–1791.

Académie Nationale de Médecine. (2022) Généraliser la détection du SARS-CoV-2 dans les eaux usées : une mesure urgente en période de reflux épidémique. *Bulletin de l'Académie Nationale de Médecine*, 206, 1–2.

Ahmed, F., Shafer, L., Malla, P., Hopkins, R., Moreland, S., Zviedrite, N. et al. (2024) Systematic review of empiric studies on lockdowns, workplace closures, and other non-pharmaceutical interventions in non-healthcare workplaces during the initial year of the COVID-19 pandemic: benefits and selected unintended consequences. *BMC Public Health*, 24, 884.

Alivernini, S., Cañete, J.D., Bacardit, J. & Kurowska-Stolarska, M. (2024) Using explainable artificial intelligence to predict and forestall flare in rheumatoid arthritis. *Nature Medicine*, 30, 925–926.

Allen, B. (2024) The promise of explainable AI in digital health for precision medicine: a systematic review. *Journal of Personalized Medicine*, 14, 277.

Amann, R.I., Baichoo, S., Blencowe, B.J., Bork, P., Borodovsky, M., Brooksbank, C. et al. (2019) Toward unrestricted use of public genomic data. *Science*, 363, 350–352.

Anthony, S.J., St Leger, J.A., Pugliares, K., Ip, H.S., Chan, J.M., Carpenter, Z.W. et al. (2012) Emergence of fatal avian influenza in New England harbor seals. *mBio*, 3, e00166-12.

Armengaud, J., Delaunay-Moisan, A., Thuret, J.-Y., van Anken, E., Acosta-Alvear, D., Aragón, T. et al. (2020) The importance of naturally attenuated SARS-CoV-2in the fight against COVID-19. *Environmental Microbiology*, 22, 1997–2000.

Awan, R.E., Zainab, S., Yousuf, F.J. & Mughal, S. (2024) AI-driven drug discovery: exploring Abaucin as a promising treatment against multidrug-resistant *Acinetobacter baumannii*. *Health Science Reports*, 7, e2150.

Baral, B. & Mozafari, M.R. (2020) Strategic moves of "superbugs" against available chemical scaffolds: signaling, regulation, and challenges. *ACS Pharmacology & Translational Science*, 3, 373–400.

Barman, S., Turner, J.C.M., Kamrul Hasan, M., Akhtar, S., Jeevan, T., Franks, J. et al. (2023) Emergence of a new genotype of clade 2.3.4.4b H5N1 highly pathogenic avian influenza A viruses in Bangladesh. *Emerging Microbes & Infections*, 12, e2252510.

Bentley, R.A. & Ormerod, P. (2009) Social versus independent interest in "bird flu" and "swine flu". *PLoS Currents*, 1, RRN1036.

Berger, Y.G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953–964.

Berlivet, L. & Löwy, I. (2020) Hydroxychloroquine controversies: clinical trials, epistemology, and the democratization of science. *Medical Anthropology Quarterly*, 34, 525–541.

Besançon, L., Cabanac, G., Labbé, C. & Magazinov, A. (2024) Sneaked references: fabricated reference metadata distort citation counts. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24896

Biever, C. (2023) ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*, 619, 686–689.

Birhane, A. (2022) The unseen Black faces of AI algorithms. *Nature*, 610, 451–452.

Bisht, D., Salave, S., Desai, N., Gogoi, P., Rana, D., Biswal, P. et al. (2024) Genome editing and its role in vaccine, diagnosis, and therapeutic advancement. *International Journal of Biological Macromolecules*, 269, 131802.

Blacksell, S.D., Dhawan, S., Kusumoto, M., Le, K.K., Summermatter, K., O'Keefe, J. et al. (2023) Laboratory-acquired infections and pathogen escapes worldwide between 2000 and 2021: a scoping review. *The Lancet Microbe*, 5, e194–e202.

Boulos, L., Curran, J.A., Gallant, A., Wong, H., Johnson, C., Delahunty-Pike, A. et al. (2023) Effectiveness of face masks for reducing transmission of SARS-CoV-2: a rapid systematic review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381, 20230133.

Bowman, S.R. (2023) Eight things to know about large language models. *arXiv* 2304.00612.

Brown, R.B. (2022) Biases in COVID-19 case and death definitions: potential causes and consequences. *Disaster Medicine and Public Health Preparedness*, 17, e313.

Brüssow, H. (2024a) Avian influenza virus cross-infections as test case for pandemic preparedness: from epidemiological hazard models to sequence-based early viral warning systems. *Microbial Biotechnology*, 17, e14389.

Brüssow, H. (2024b) Pandemic preparedness: on the efficacy of non-pharmaceutical interventions in COVID-19 and about approaches to predict future pandemic viruses. *Microbial Biotechnology*, 17, e14431.

Cakmak, A. & Ozsoyoglu, G. (2007) Annotating genes using textual patterns. *Pacific Symposium on Biocomputing*, 2007, 221–232.

Carneiro, H.A. & Mylonakis, E. (2009) Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49, 1557–1564.

Centers for Disease Control and Prevention (CDC). (1997) Isolation of avian influenza A (H5N1) viruses from humans – Hong Kong, May-December 1997. *Morbidity and Mortality Weekly Report*, 46, 1204–1207.

Cerchia, C. & Lavecchia, A. (2023) New avenues in artificial-intelligence-assisted drug discovery. *Drug Discovery Today*, 28, 103516.

Chaber, A.-L. (2018) The era of human-induced diseases. *EcoHealth*, 15, 8–11.

Changeux, J.P., Courrège, P. & Danchin, A. (1973) A theory of the epigenesis of neuronal networks by selective stabilization of synapses. *Proceedings of the National Academy of Sciences of the United States of America*, 70, 2974–2978.

Charlier, J., Barkema, H.W., Becher, P., De Benedictis, P., Hansson, I., Hennig-Pauka, I. et al. (2022) Disease control tools to secure animal and public health in a densely populated world. *Lancet Planetary Health*, 6, e812–e824.

Chen, L., Kumar, S. & Wu, H. (2023) A review of current antibiotic resistance and promising antibiotics with novel modes of action to combat antibiotic resistance. *Archives of Microbiology*, 205, 356.

Chen, X., Kalyar, F., Chughtai, A.A. & MacIntyre, C.R. (2024) Use of a risk assessment tool to determine the origin of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Risk Analysis*, 44, 1896–1906.

Cheng, V.C.C., Lau, S.K.P., Woo, P.C.Y. & Yuen, K.Y. (2007) Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clinical Microbiology Reviews*, 20, 660–694.

Chomsky, N. (2015) *Aspects of the theory of syntax: 50th anniversary edition, fiftieth anniversary edition*. Cambridge, MA: The MIT Press.

Chomsky, N. (2017) Language architecture and its import for evolution. *Neuroscience & Biobehavioral Reviews*, 81, 295–300.

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R. et al. (2022) Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8, 59.

Cluzel, N., Lambert, A., Maday, Y., Turinici, G. & Danchin, A. (2020) Biochemical and statistical lessons from the evolution of the SARS-CoV-2 virus: paths for novel antiviral warfare. *Comptes Rendus Biologies*, 343, 177–209.

Cottam, E.M., Wadsworth, J., Shaw, A.E., Rowlands, R.J., Goatley, L., Maan, S. et al. (2008) Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS Pathogens*, 4, e1000050.

Coxe, T. & Azad, R.K. (2023) Silicon versus superbug: assessing machine learning's role in the fight against antimicrobial resistance. *Antibiotics*, 12, 1604.

Danchin, A. (1996) On genomes and cosmologies. In: Collado-Vides, J., Magasanik, B. & Smith, T.F. (Eds.) *Integrative approaches to molecular biology*. Cambridge, MA: The MIT Press, pp. 91–112.

Danchin, A. (2002) Not every truth is good. The dangers of publishing knowledge about potential bioweapons. *EMBO Reports*, 3, 102–104.

Danchin, A. (2003) Infection of society. As diseases have evolved to exploit the holes in our defences, including weaknesses in society, we have to reconsider our way of life, otherwise they will continue to haunt us. *EMBO Reports*, 4, 333–335.

Danchin, A., Ouzounis, C., Tokuyasu, T. & Zucker, J.-D. (2018) No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. *Microbial Biotechnology*, 11, 588–605.

Danielli, J.F. (1972) IV. Artificial synthesis of new life forms. *Bulletin of the Atomic Scientists*, 28, 20–24.

Dick, S. (2019) *Artificial intelligence*. Cambridge, MA: MIT Press.

Douven, I. (2022) What is abduction? In: *The art of abduction*. Cambridge, MA: The MIT Press, pp. 29–68.

Dresher, B.E. & Kaye, J.D. (1990) A computational learning model for metrical phonology. *Cognition*, 34, 137–195.

Duda, R.O. & Shortliffe, E.H. (1983) Expert systems research. *Science*, 220, 261–268.

Duizer, E., Ruijs, W.L., Putri Hintaran, A., Hafkamp, M.C., Van Der Veer, M. & Te Wierik, M.J. (2023) Wild poliovirus type 3 (WPV3)-shedding event following detection in environmental surveillance of poliovirus essential facilities, the Netherlands, November 2022 to January 2023. *Eurosurveillance*, 28, 2300049.

Ebright, R.H., MacIntyre, R., Dudley, J.P., Butler, C.D., Goffinet, A., Hammond, E. et al. (2024) Implementing governmental oversight of enhanced potential pandemic pathogen research. *Journal of Virology*, 98, e0023724.

Ekambaram, A., Sørensen, A.Ø., Bull-Berg, H. & Olsson, N.O.E. (2018) The role of big data and knowledge management in improving projects and project-based organizations. *Procedia Computer Science*, 138, 851–858.

Emmert-Streib, F., Yli-Harja, O. & Dehmer, M. (2020) Artificial intelligence: a clarification of misconceptions, myths and desired status. *Frontiers in Artificial Intelligence*, 3, 524339.

Esvelt, K.M. (2018) Inoculating science against potential pandemics and information hazards. *PLoS Pathogens*, 14, e1007286.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A. et al. (2024) Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30, 958–968.

Frank, F., Florens, N., Meyerowitz-katz, G., Barriere, J., Billy, É., Saada, V. et al. (2023) Raising concerns on questionable ethics approvals – a case study of 456 trials from the Institut Hospitalo-Universitaire Méditerranée infection. *Research Integrity and Peer Review*, 8, 9.

Frierson, J.G. (2010) The yellow fever vaccine: a history. *The Yale Journal of Biology and Medicine*, 83, 77–85.

Gallardo, M.J. & Delgado, F.O. (2021) Animal prion diseases: a review of intraspecies transmission. *Open Veterinary Journal*, 11, 707–723.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S. & Ouzounis, C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18, 1641–1649.

Grassly, N.C. & Fraser, C. (2008) Mathematical models of infectious disease transmission. *Nature Reviews. Microbiology*, 6, 477–487.
Gray, D.A. & Wenzel, M. (2020) Multitarget approaches against multiresistant superbugs. *ACS Infectious Diseases*, 6, 1346–1365.
Groves, H.E., Piché-Renaud, P.-P., Peci, A., Farrar, D.S., Buckrell, S., Bancej, C. et al. (2021) The impact of the COVID-19 pandemic on influenza, respiratory syncytial virus, and other seasonal respiratory virus circulation in Canada: a population-based study. *The Lancet Regional Health – Americas*, 1, 100015.
Guan, L. (2024) Reaching carbon neutrality requires energy-efficient training of AI. *Nature*, 626, 33.
Guo, C.-Y., Zhang, W.-X., Zhou, Y.-G., Zhang, S.-S., Xi, L., Zheng, R.-R. et al. (2024) Dynamics of respiratory infectious diseases under rapid urbanization and COVID-19 pandemic in the subcenter of Beijing during 2014-2022. *Heliyon*, 10, e29987.
Gurney, K. (2018) *An introduction to neural networks*. Boca Raton, Florida, USA: CRC Press.
Haiech, J. & Sallantin, J. (1985) Computer search of calcium binding sites in a gene data bank: use of learning techniques to build an expert system. *Biochimie*, 67, 555–560.
Hanson, A.D., Pribat, A., Waller, J.C. & de Crécy-Lagard, V. (2009) "Unknown" proteins and "orphan" enzymes: the missing half of the engineering parts list – and how to find it. *The Biochemical Journal*, 425, 1–11.
Haque, S., Mengersen, K., Barr, I., Wang, L., Yang, W., Vardoulakis, S. et al. (2024) Towards development of functional climate-driven early warning systems for climate-sensitive infectious diseases: statistical models and recommendations. *Environmental Research*, 249, 118568.
Hawkins, J. & Blakeslee, S. (2005) *On intelligence. How a new understanding of the brain will Lead to the creation of truly intelligent machines*, 1st edition. New York, NY: Owl Books.
Hayes-Roth, F., Waterman, D.A. & Lenat, D.B. (Eds.). (1983) *Building expert systems*. Reading, MA: Addison-Wesley Pub. Co.
Hegde, S.T., Lee, K.H., Styczynski, A., Jones, F.K., Gomes, I., Das, P. et al. (2024) Potential for person-to-person transmission of henipaviruses: a systematic review of the literature. *The Journal of Infectious Diseases*, 229, 733–742.
Heller, T., Saito, S., Auerbach, J., Williams, T., Moreen, T.R., Jazwinski, A. et al. (2005) An in vitro model of hepatitis C virion production. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2579–2583.
Henaut, A. & Danchin, A. (1996) Analysis and predictions from *Escherichia coli* sequences. In: *Escherichia coli and Salmonella typhimurium cellular and molecular biology*. Washington, DC: ASM Press, pp. 2047–2066.
Heymann, D.L., Aylward, R.B. & Wolff, C. (2004) Dangerous pathogens in the laboratory: from smallpox to today's SARS setbacks and tomorrow's polio-free world. *Lancet*, 363, 1566–1568.
Hill. (1956) Proceedings of the meeting. *Journal of the Royal Statistical Society Series A (General)*, 119, 146–149.
Hofstadter, D.R. (1979) *Gödel, Escher, Bach: an eternal golden braid*. New York: Basic Books.
Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79, 2554–2558.
Horton, P.B. & Kanehisa, M. (1992) An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucleic Acids Research*, 20, 4331–4338.
Hung, K.K.C., Mark, C.K.M., Yeung, M.P.S., Chan, E.Y.Y. & Graham, C.A. (2018) The role of the hotel industry in the response to emerging epidemics: a case study of SARS in 2003 and H1N1 swine flu in 2009 in Hong Kong. *Globalization and Health*, 14, 117.
Islam, A., Munro, S., Hassan, M.M., Epstein, J.H. & Klaassen, M. (2023) The role of vaccination and environmental factors on outbreaks of high pathogenicity avian influenza H5N1 in Bangladesh. *One Health*, 17, 100655.
Jaeger, A.S., Marano, J., Riemersma, K.K., Castaneda, D., Pritchard, E.M., Pritchard, J.C. et al. (2023) Gain without pain: adaptation and increased virulence of Zika virus in vertebrate host without fitness cost in mosquito vector. *Journal of Virology*, 97, e01162-23.
Jang, W.D., Kim, G.B., Kim, Y. & Lee, S.Y. (2022) Applications of artificial intelligence to enzyme and pathway design for metabolic engineering. *Current Opinion in Biotechnology*, 73, 101–107.
Jefferson, T., Dooley, L., Ferroni, E., Al-Ansary, L.A., van Driel, M.L., Bawazeer, G.A. et al. (2023) Physical interventions to interrupt or reduce the spread of respiratory viruses. *Cochrane Database of Systematic Reviews*, 1, CD006207.
Kaiser, J. (2024) International panel calls for tighter oversight of risky pathogen studies. *Science*Insider, 28 February 2024.
Kalimeri, K., Delfino, M., Cattuto, C., Perrotta, D., Colizza, V., Guerrisi, C. et al. (2019) Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms. *PLoS Computational Biology*, 15, e1006173.
Kendal, A.P., Noble, G.R., Skehel, J.J. & Dowdle, W.R. (1978) Antigenic similarity of influenza A(H1N1) viruses from epidemics in 1977–1978 to "Scandinavian" strains isolated in epidemics of 1950–1951. *Virology*, 89, 632–636.
Kermack, W.O. & McKendrick, A.G. (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115, 700–721.
Kessler, S., García-Sastre, A., Schwemmle, M. & Ciminski, K. (2024) Reverse genetics of bat influenza A viruses. *Methods in Molecular Biology*, 2733, 75–86.
Kibenge, F.S.B. (2023) A One Health approach to mitigate the impact of influenza a virus (IAV) reverse zoonosis is by vaccinating humans and susceptible farmed and pet animals. *American Journal of Veterinary Research*, 84, ajvr.23.03.0053.
Kilbourne, E.D. (2006) Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, 12, 9–14.
Kobayashi, G.S. (1996) Disease mechanisms of fungi. In: *Medical microbiology*. Galveston, TX: University of Texas Medical Branch at Galveston.
Kong, W., Hu, Y., Zhang, J. & Tan, Q. (2022) Application of SMILES-based molecular generative model in new drug design. *Frontiers in Pharmacology*, 13, 1046524.
Kortemme, T. (2024) De novo protein design-from new structures to programmable functions. *Cell*, 187, 526–544.
Kress, A., Poch, O., Lecompte, O. & Thompson, J.D. (2023) Real or fake? Measuring the impact of protein annotation errors on estimates of domain gain and loss events. *Frontiers in Bioinformatics*, 3, 1178926.
Kung, H.C., Jen, K.F., Yuan, W.C., Tien, S.F. & Chu, C.M. (1978) Influenza in China in 1977: recurrence of influenzavirus A subtype H1N1. *Bulletin of the World Health Organization*, 56, 913–918.
Larrouquere, L., Gabin, M., Poingt, E., Mouffak, A., Hlavaty, A., Lepelley, M. et al. (2020) Genesis of an emergency public drug information website by the French Society of Pharmacology and Therapeutics during the COVID-19 pandemic. *Fundamental & Clinical Pharmacology*, 34, 389–396.
Lauber, C., Zhang, X., Vaas, J., Klingler, F., Mutz, P., Dubin, A. et al. (2024) Deep mining of the sequence read archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. *PLoS Pathogens*, 20, e1012163.
Laude, H., Van Reeth, K. & Pensaert, M. (1993) Porcine respiratory coronavirus: molecular features and virus-host interactions. *Veterinary Research*, 24, 125–150.
LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521, 436–444.
Lehto, K.-M., Länsivaara, A., Hyder, R., Luomala, O., Lipponen, A., Hokajärvi, A.-M. et al. (2024) Wastewater-based surveillance is

an efficient monitoring tool for tracking influenza a in the community. *Water Research*, 257, 121650.
Lenharo, M. (2023) GISAID in crisis: can the controversial COVID genome database survive? *Nature*, 617, 455–457.
Leung, N.H.L. (2021) Transmissibility and transmission of respiratory viruses. *Nature Reviews. Microbiology*, 19, 528–545.
Leuzzi, A., Di Martino, M.L., Campilongo, R., Falconi, M., Barbagallo, M., Marcocci, L. et al. (2015) Multifactor regulation of the MdtJI polyamine transporter in Shigella. *PLoS One*, 10, e0136744.
Li, H., Huang, J., Lian, X., Zhao, Y., Yan, W., Zhang, L. et al. (2023) Impact of human mobility on the epidemic spread during holidays. *Infectious Disease Modelling*, 8, 1108–1116.
Li, J., Ma, Y., Xu, X., Pei, J. & He, Y. (2022) A study on epidemic information screening, prevention and control of public opinion based on health and medical big data: a case study of COVID-19. *International Journal of Environmental Research and Public Health*, 19, 9819.
Li, J., Wang, C., Li, X., Zhang, G., Sun, S., Wang, Z. et al. (2024) Direct transmission of severe fever with thrombocytopenia syndrome virus from farm-raised fur animals to workers in Weihai, China. *Virology Journal*, 21, 113.
Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. (2022) A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33, 6999–7019.
Lim, P.L., Kurup, A., Gopalakrishna, G., Chan, K.P., Wong, C.W., Ng, L.C. et al. (2004) Laboratory-acquired severe acute respiratory syndrome. *The New England Journal of Medicine*, 350, 1740–1745.
Liu, Z., Ding, M., Hu, C., Rong, R., Lin, C., Yao, G. et al. (2024) Susceptibility and exposure risk to airborne aerosols in intra-urban microclimate: evidence from subway system of megacities. *Science of the Total Environment*, 917, 170514.
Liu, Z., Zubatiuk, T., Roitberg, A. & Isayev, O. (2022) Auto3D: automatic generation of the low-energy 3D structures with ANI neural network potentials. *Journal of Chemical Information and Modeling*, 62, 5373–5382.
Lou, J., Liang, W., Cao, L., Hu, I., Zhao, S., Chen, Z. et al. (2024) Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations. *Nature Communications*, 15, 2546.
MacIntyre, C.R. (2023) Illegal biolabs in the community – is Reedley a one-off? *Global Biosecurity*, 5, gbio.237. https://doi.org/10.31646/gbio.237
Manheim, D. & Lewis, G. (2021) High-risk human-caused pathogen exposure events from 1975-2016. *F1000Research*, 10, 752.
McCorduck, P. (2019) *This could be important: my life and times with the artificial intelligentsia*. Pittsburgh, PA: Carnegie Mellon University, ETC Press, Signature.
McCulloch, W.S. & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
McQuiston, J.H., Luce, R., Kazadi, D.M., Bwangandu, C.N., Mbala-Kingebeni, P., Anderson, M. et al. (2024) U.S. preparedness and response to increasing clade I Mpox cases in the Democratic Republic of the Congo — United States, 2024. *Morbidity and Mortality Weekly Report*, 73, 435–440.
Médigue, C., Calteau, A., Cruveiller, S., Gachet, M., Gautreau, G., Josso, A. et al. (2019) MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. *Briefings in Bioinformatics*, 20, 1071–1084.
Meijers, M., Ruchnewitz, D., Eberhardt, J., Łuksza, M. & Lässig, M. (2023) Population immunity predicts evolutionary trajectories of SARS-CoV-2. *Cell*, 186, 5151–5164.e13.
Merz, J.J., Barnard, P., Rees, W.E., Smith, D., Maroni, M., Rhodes, C.J. et al. (2023) World scientists' warning: the behavioural crisis driving ecological overshoot. *Science Progress*, 106, 368504231201372.
Millett, P. (2023) Safeguard the world's worst pathogens. *Science*, 382, 243.
Millett, P. (2024) EU regulations for maximum containment labs-response. *Science*, 383, 1188–1189.
Minor, P.D. (2015) Live attenuated vaccines: historical successes and current challenges. *Virology*, 479–480, 379–392.
Mitka, M. (2005) 1918 killer flu virus reconstructed, may help prevent future outbreaks. *JAMA*, 294, 2416.
Monto, A.S., Black, S., Plotkin, S.A. & Orenstein, W.A. (2011) Response to the 2009 pandemic: effect on influenza control in wealthy and poor countries. *Vaccine*, 29, 6427–6431.
Muraille, E., Naccache, P. & Pillot, J. (2022) The tragedy of liberal democratic governance in the face of global threats. *Frontiers in Public Health*, 10, 902724.
Musen, M.A., O'Connor, M.J., Schultes, E., Martínez-Romero, M., Hardi, J. & Graybeal, J. (2022) Modeling community standards for metadata as templates makes data FAIR. *Scientific Data*, 9, 696.
Nakajima, K., Desselberger, U. & Palese, P. (1978) Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. *Nature*, 274, 334–339.
Nelson, M.I., Viboud, C., Simonsen, L., Bennett, R.T., Griesemer, S.B., St. George, K. et al. (2008) Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathogens*, 4, e1000012.
Ng, T.-W., Turinici, G., Ching, W.-K., Chung, S.-K. & Danchin, A. (2007) A parasite vector-host epidemic model for TSE propagation. *Medical Science Monitor*, 13, BR59–BR66.
Ng, T.W., Turinici, G. & Danchin, A. (2003) A double epidemic model for the SARS propagation. *BMC Infectious Diseases*, 3, 19.
Nguyen, C.T. (2024) The limits of data. *Issues*, 40, 94–101.
Nunes, S.F., Hamers, C., Ratinier, M., Shaw, A., Brunet, S., Hudelet, P. et al. (2014) A synthetic biology approach for a vaccine platform against known and newly emerging serotypes of bluetongue virus. *Journal of Virology*, 88, 12222–12232.
Oakley, R., Hedrich, N., Walker, A., Dinkita, H.M., Tschopp, R., Abongomera, C. et al. (2024) Status of zoonotic disease research in refugees, asylum seekers and internally displaced people, globally: a scoping review of forty clinically important zoonotic pathogens. *PLoS Neglected Tropical Diseases*, 18, e0012164.
Oliveira, M., Wilming, R., Clark, B., Budding, C., Eitel, F., Ritter, K. et al. (2024) Benchmarking the influence of pre-training on explanation performance in MR image classification. *Frontiers in Artificial Intelligence*, 7, 1330919.
Olson, D.R., Konty, K.J., Paladini, M., Viboud, C. & Simonsen, L. (2013) Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9, e1003256.
Omar, M., Brin, D., Glicksberg, B. & Klang, E. (2024) Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: a systematic review. *American Journal of Infection Control*, 52, 992–1001.
Osaigbovo, I.I., Ekeng, B.E., Davies, A.A., Ebeigbe, E., Bongomin, F., Kanyua, A. et al. (2024) *Candida auris*: a systematic review of a globally emerging fungal pathogen in Africa. *Open Forum Infectious Diseases*, 11, ofad681.
Pepin, K.M., Carlisle, K., Anderson, D., Baker, M.G., Chipman, R.B., Benschop, J. et al. (2024) Steps towards operationalizing One Health approaches. *One Health*, 18, 100740.
Perrow, C. (1999) *Normal accidents: living with high-risk technologies*. Princeton, NJ: Princeton University Press.
Perrow, C. (2011) *The next catastrophe: reducing our vulnerabilities to natural, industrial, and terrorist disasters*. Princeton, NJ: Princeton University Press.
Peters, U. (2022) Algorithmic political bias in artificial intelligence systems. *Philosophy and Technology*, 35, 25.

Pitt, S.J. & Gunn, A. (2024) The One Health concept. *British Journal of Biomedical Science*, 81, 12366.

Plaza, P.I., Gamarra-Toledo, V., Euguí, J.R. & Lambertucci, S.A. (2024) Recent changes in patterns of mammal infection with highly pathogenic avian influenza A(H5N1) virus worldwide. *Emerging Infectious Diseases*, 30, 444–452.

Poptsova, M.S. & Gogarten, J.P. (2010) Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*, 156, 1909–1917.

Prelec, D., Seung, H.S. & McCoy, J. (2017) A solution to the single-question crowd wisdom problem. *Nature*, 541, 532–535.

Prifti, E., Chevaleyre, Y., Hanczar, B., Belda, E., Danchin, A., Clément, K. et al. (2020) Interpretable and accurate prediction models for metagenomics data. *GigaScience*, 9, giaa010.

Pu, D., Zhao, J., Chang, K., Zhuo, X. & Cao, B. (2023) "Superbugs" with hypervirulence and carbapenem resistance in *Klebsiella pneumoniae*: the rise of such emerging nosocomial pathogens in China. *Science Bulletin*, 68, 2658–2670.

Quinqueton, J. (1985) OURCIN: a tool to build expert systems. *Biochimie*, 67, 485–491.

Rapaport, W. (2006) Turing test. In: *Encyclopedia of language & linguistics*. Amsterdam: Elsevier, pp. 151–159.

Rewar, S., Mirdha, D. & Rewar, P. (2015) Treatment and prevention of pandemic H1N1 influenza. *Annals of Global Health*, 81, 645–653.

Rice, L., Wong, E. & Kolter, Z. (2020) Overfitting in adversarially robust deep learning. PMLR (pp. 8093–8104).

Rodan, L.H., Anyane-Yeboa, K., Chong, K., Klein Wassink-Ruiter, J.S., Wilson, A., Smith, L. et al. (2018) Gain-of-function variants in the ODC1 gene cause a syndromic neurodevelopmental disorder associated with macrocephaly, alopecia, dysmorphic features, and neuroimaging abnormalities. *American Journal of Medical Genetics. Part A*, 176, 2554–2560.

Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408.

Ross, E. & Harper, D.R. (2023) *Laboratory accidents and biocontainment breaches: policy options for improved safety and security.* Chatham House, London: Royal Institute of International Affairs.

Rozo, M. & Gronvall, G.K. (2015) The reemergent 1977 H1N1 strain and the gain-of-function debate. *mBio*, 6, e01013-15.

Sætra, H.S. (2023) Generative AI: here to stay, but for good? *Technology in Society*, 75, 102372.

Sah, R., Srivastava, S., Kumar, S., Mehta, R., Donovan, S., Sierra-Carrero, L. et al. (2024) Concerns on H5N1 avian influenza given the outbreak in U.S. dairy cattle. *The Lancet Regional Health – Americas*, 35, 100785.

Sahoh, B. & Choksuriwong, A. (2023) The role of explainable artificial intelligence in high-stakes decision-making systems: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 14, 7827–7843.

Salmanton-García, J., Wipfler, P., Leckler, J., Nauclér, P., Mallon, P.W., Bruijning-Verhagen, P.C.J.L. et al. (2024) Predicting the next pandemic: VACCELERATE ranking of the World Health Organization's Blueprint for Action to Prevent Epidemics. *Travel Medicine and Infectious Disease*, 57, 102676.

Satpathy, P., Kumar, S. & Prasad, P. (2021) Suitability of Google trends™ for digital surveillance during ongoing COVID-19 epidemic: a case study from India. *Disaster Medicine and Public Health Preparedness*, 17, e28.

Schriml, L.M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E.A., Finn, R.D., Hugenholtz, P. et al. (2020) COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data*, 7, 188.

Schuerger, C., Batalis, S., Quinn, K., Kinoshita, R., Daniels, O. & Puglisi, A. (2023) *Understanding the global gain-of-function research landscape*. Washington, DC, USA: Center for Security and Emerging Technology.

Seel, N.M. (Ed.). (2012) *Encyclopedia of the sciences of learning*. Boston, MA: Springer US.

Seo, D.-W. & Shin, S.-Y. (2017) Methods using social media and search queries to predict infectious disease outbreaks. *Healthcare Informatics Research*, 23, 343–348.

Shi, Z., Zhai, Y., Zhang, Y. & Wei, H. (2022) SNAL: sensitive non-associative learning network configuration for the automatic driving strategy. *Scientific Reports*, 12, 20045.

Shih, D.-H., Wu, Y.-H., Wu, T.-W., Chang, S.-C. & Shih, M.-H. (2024) Infodemiology of influenza-like illness: utilizing Google Trends' big data for epidemic surveillance. *Journal of Clinical Medicine*, 13, 1946.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. (2024) AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

Silva, H.M. (2024) Challenges and reflections on pandemic disinformation: the case of hydroxychloroquine and the implications for global public health. *Value in Health Regional Issues*, 43, 101005.

Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 13, 238–241.

Srihith, I.D., Donald, A.D., Srinivas, T.A.S., Anjali, D. & Varaprasad, R. (2023) The backbone of computing: an exploration of data structures. *International Journal of Advanced Research in Science, Communication and Technology*, 3, 155–163.

Su, W., Sun, L., Zhao, W., Song, S., Yang, Y., He, Y. et al. (2024) The prediction of influenza-like illness using national influenza surveillance data and Baidu query data. *BMC Public Health*, 24, 513.

Sun, H., Li, H., Tong, Q., Han, Q., Liu, J., Yu, H. et al. (2023) Airborne transmission of human-isolated avian H3N8 influenza virus between ferrets. *Cell*, 186, 4074–4084.e11.

Sun, J., Yuan, K., Chen, C., Xu, H., Wang, H., Zhi, Y. et al. (2023) Causality network of infectious disease revealed with causal decomposition. *IEEE Journal of Biomedical and Health Informatics*, 27, 3657–3665.

Sun, Y., Chen, C., Zeng, C., Xia, Q., Yuan, C. & Pei, H. (2024) Severe fever with thrombocytopenia syndrome virus infection shapes gut microbiome of the tick vector *Haemaphysalis longicornis*. *Parasites and Vectors*, 17, 107.

Takashita, E., Morita, H., Nagata, S., Shirakura, M., Fujisaki, S., Miura, H. et al. (2022) Antiviral susceptibilities of avian influenza A(H5), A(H7), and A(H9) viruses isolated in Japan. *Japanese Journal of Infectious Diseases*, 75, 398–402.

Taleb, N.N. & Taleb, N.N. (2016) *The black swan: the impact of the highly improbable*. New York: Random House.

Tan, R.K.J., Perera, D., Arasaratnam, S. & Kularathne, Y. (2024) Adapting an artificial intelligence sexually transmitted diseases symptom checker tool for Mpox detection: the HeHealth experience. *Sexual Health*, 21, SH23197.

Tang, L., Rhoads, W.J., Eichelberg, A., Hamilton, K.A. & Julian, T.R. (2024) Applications of quantitative microbial risk assessment to respiratory pathogens and implications for uptake in policy: a state-of-the-science review. *Environmental Health Perspectives*, 132, 56001.

Taylor, L. (2024) WHO pandemic treaty: "Torrent of fake news" has put negotiations at risk, says WHO chief. *BMJ*, 384, q243.

Taylor, M.R., Stevens, G.J., Agho, K.E., Kable, S.A. & Raphael, B. (2012) Crying wolf? Impact of the H1N1 2009 influenza pandemic on anticipated public response to a future pandemic. *The Medical Journal of Australia*, 197, 561–564.

Timme, R.E., Karsch-Mizrachi, I., Waheed, Z., Arita, M., MacCannell, D., Maguire, F. et al. (2023) Putting everything in its place: using the INSDC compliant pathogen data object model to

better structure genomic data submitted for public health applications. *Microbial Genomics*, 9, 001145.

Tumelty, L., Fa, J.E., Coad, L., Friant, S., Mbane, J., Kamogne, C.T. et al. (2023) A systematic mapping review of links between handling wild meat and zoonotic diseases. *One Health*, 17, 100637.

Turinici, G. & Danchin, A. (2007) The SARS case study: An alarm clock? In: Tibayrenc, M. (Ed.) *Encyclopedia of infectious diseases*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 151–162.

Undheim, T.A. (2024) The whack-a-mole governance challenge for AI-enabled synthetic biology: literature review and emerging frameworks. *Frontiers in Bioengineering and Biotechnology*, 12, 1359768.

Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. (2022) Dual use of artificial intelligence-powered drug discovery. *Nature Machine Intelligence*, 4, 189–191.

Vellani, V., Zheng, S., Ercelik, D. & Sharot, T. (2023) The illusory truth effect leads to the spread of misinformation. *Cognition*, 236, 105421.

Vindman, C., Trump, B., Cummings, C., Smith, M., Titus, A.J., Oye, K. et al. (2024) The convergence of AI and synthetic biology: the looming deluge, *arXiv* 2024. 18973.

Wang, X., Qiu, W., Hu, G., Diao, X., Li, Y., Li, Y. et al. (2024) NS7a of SADS-CoV promotes viral infection via inducing apoptosis to suppress type III interferon production. *Journal of Virology*, 98, e0031724.

Wardle, J., Bhatia, S., Kraemer, M.U.G., Nouvellet, P. & Cori, A. (2023) Gaps in mobility data and implications for modelling epidemic spread: a scoping review and simulation study. *Epidemics*, 42, 100666.

Wasserman, E.A., Kain, A.G. & O'Donoghue, E.M. (2023) Resolving the associative learning paradox by category learning in pigeons. *Current Biology*, 33, 1112–1116.e2.

Wells, H.L., Bonavita, C.M., Navarrete-Macias, I., Vilchez, B., Rasmussen, A.L. & Anthony, S.J. (2023) The coronavirus recombination pathway. *Cell Host & Microbe*, 31, 874–889.

Wilder-Smith, A. (2006) The severe acute respiratory syndrome: impact on travel and tourism. *Travel Medicine and Infectious Disease*, 4, 53–60.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A. et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

Wu, F. (2023) Updated analysis to reject the laboratory-engineering hypothesis of SARS-CoV-2. *Environmental Research*, 224, 115481.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S. et al. (2021) Artificial intelligence: a powerful paradigm for scientific research. *The Innovation*, 2, 100179.

Yang, F., Zhang, X., Liu, F., Yao, H., Wu, N. & Wu, H. (2022) Increased virulence of a novel reassortant H1N3 avian influenza virus in mice as a result of adaptive amino acid substitutions. *Virus Genes*, 58, 473–477.

Yang, L., Zhang, T., Han, X., Yang, J., Sun, Y., Ma, L. et al. (2023) Influenza epidemic trend surveillance and prediction based on search engine data: deep learning model study. *Journal of Medical Internet Research*, 25, e45085.

Yang, W., Schountz, T. & Ma, W. (2021) Bat influenza viruses: current status and perspective. *Viruses*, 13, 547.

Yoshikai, Y., Mizuno, T., Nemoto, S. & Kusuhara, H. (2024) Difficulty in chirality recognition for transformer architectures learning chemical structures from string representations. *Nature Communications*, 15, 1197.

Yu, S., Gu, C., Huang, K. & Li, P. (2024) Predicting the next sentence (not word) in large language models: what model-brain alignment tells us about discourse comprehension. *Science Advances*, 10, eadn7744.

Zhang, X.W., Yap, Y.L. & Danchin, A. (2005) Testing the hypothesis of a recombinant origin of the SARS-associated coronavirus. *Archives of Virology*, 150, 1–20.

Zheng, Y., Song, K., Xie, Z.-X., Han, M.-Z., Guo, F. & Yuan, Y.-J. (2023) Machine learning-aided scoring of synthesis difficulties for designer chromosomes. *Science China. Life Sciences*, 66, 1615–1625.

How to cite this article: Danchin, A. (2024) Artificial intelligence-based prediction of pathogen emergence and evolution in the world of synthetic biology. *Microbial Biotechnology*, 17, e70014. Available from: https://doi.org/10.1111/1751-7915.70014