Brain, Behavior, and Immunity 115 (2024) 470–479

Contents lists available at ScienceDirect

Brain Behavior and Immunity

journal homepage: www.elsevier.com/locate/ybrbi

Machine learning and artificial intelligence in neuroscience: A primer for researchers

Fakhirah Badrulhisham [a], Esther Pogatzki-Zahn [b], Daniel Segelcke [b], Tamas Spisak [c,d], Jan Vollert [e,f,*]

[a] Royal Devon and Exeter Hospital NHS Trust, Exeter, United Kingdom
[b] Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital Muenster, Muenster, Germany
[c] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Medicine Essen, Essen, Germany
[d] Center for Translational Neuro- and Behavioral Sciences, Department of Neurology, University Medicine Essen, Essen, Germany
[e] Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom
[f] Pain Research, Department of Surgery and Cancer, Imperial College London, London, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Artificial intelligence
Predictive modelling
Neuroscience
Pain
fMRI
Behavioural research
*omics

ABSTRACT

Artificial intelligence (AI) is often used to describe the automation of complex tasks that we would attribute intelligence to. Machine learning (ML) is commonly understood as a set of methods used to develop an AI. Both have seen a recent boom in usage, both in scientific and commercial fields.

For the scientific community, ML can solve bottle necks created by complex, multi-dimensional data generated, for example, by functional brain imaging or *omics approaches. ML can here identify patterns that could not have been found using traditional statistic approaches. However, ML comes with serious limitations that need to be kept in mind: their tendency to optimise solutions for the input data means it is of crucial importance to externally validate any findings before considering them more than a hypothesis. Their black-box nature implies that their decisions usually cannot be understood, which renders their use in medical decision making problematic and can lead to ethical issues.

Here, we present an introduction for the curious to the field of ML/AI. We explain the principles as commonly used methods as well as recent methodological advancements before we discuss risks and what we see as future directions of the field. Finally, we show practical examples of neuroscience to illustrate the use and limitations of ML.

1. Introduction

In multiple domains of healthcare and biology, we face problems for which there are no mono-causal solutions: either, because we do not know possible paths to solution, due to the numerous, multidimensional variables involved or because we can imagine a path to solution, but it turns out to be a puzzle so complex that we cannot solve it. An example for the former could be attempting to predict the development of tumours based on *omics-datasets: we have only very limited understanding and hypotheses of which of these data holds potential as predictive cancer biomarkers, but we assume that there is a reasonable chance that some of them do. An example for the latter would be the folding of a potential protein based on its amino-acid sequence: we generally know the principles, however, taking all interactions into account makes it computationally intractable to fully calculate the precise folding (Dill et al., 2008). Both examples have more things in common. For one, if we cannot find a *perfect* solution, we would still be excited about a very close estimate: for proteins, all we need to know is if it will modulate relevant pathways, and for the prediction of cancer, any improvement over current clinical algorithms of prediction would be valuable, even if they are not perfect. Second, even if we do not understand how the result is reached, in both cases, we could still put the result to a meaningful use. And lastly, for both cases, we could draw on large datasets of annotated previous cases.

Computationally, we have found that a group of algorithms widely known as *machine learning* (ML) can produce extraordinary results on

* Corresponding author at: Department of Clinical and Biomedical Sciences, University of Exeter Medical School, EMS Building G09, St Luke's Campus, Heavitree Rd., Exeter EX1 2LU, United Kingdom.
E-mail address: j.vollert@exeter.ac.uk (J. Vollert).

https://doi.org/10.1016/j.bbi.2023.11.005
Received 26 April 2023; Received in revised form 16 October 2023; Accepted 8 November 2023
Available online 14 November 2023

these kinds of problems, which are too complex to solve perfectly. This solution generally requires the use of enormous amounts of exemplary annotated data for which we can accept a good result, even if we cannot explain the solution path. To some degree, this can be thought to mimic human decisions: a "good" doctor will decide to take out a tumour they assume to be malignant based largely on experience (using clinical guidelines and algorithms as a handrail). One of the central promises of artificial intelligence in healthcare therefor is to replace the experience of *one* doctor with that of *thousands* of doctors.

While leading to breakthroughs in some of these questions like protein folding prediction (Jumper et al., 2021), ML algorithms face multiple challenges, which need to be taken into serious account before applying and when judging the results. In this review, we aim to introduce frequently used techniques, explain basic concepts, show opportunities and risks in the field, and provide some examples of neuroscientific research applications. We intend this review as a primer for the interested – we therefore will use simplifications that can be seen as oversimplifications at points. We do so to keep to a beginner's level, but please keep in mind that many, most, or all of the concepts explained here are more nuanced and complex than portrayed here when drilling down into them.

1.1. Terminology and prominent methods

While clear definitions are missing for most of the terms in the field, the term *artificial intelligence* (AI) is often used to describe the automation of complex tasks we would generally attribute intelligence to (Luger, 2004). *Machine Learning* (ML), then, is a loosely connected group of methods to achieve an AI approach.

Here, we will talk exclusively about *specific* AI: an AI used to solve a specific problem, which it has been trained for. An unfortunate extrapolation is often made from the advancement in *specific* AIs to the rise of a *general* AI (the AI you know from mo. vies and literature, an artificial intelligence being able to solve a wide range of problems and being able to find solutions to new problems on its own). It is possible, but currently not foreseeable that the advancements in specific AI will lead to the development of a general AI. Think of it in the same way that the discovery of a possible drug target in an animal model might or might not lead to safe and effective medication against dementia in humans. Of course, the discovery of the drug target increases the chance of future treatment, but it is only one of many steps in a long journey, each of which can fail, and none of which we can predict how long they will take.

Further subdivisions can go along with intent of application (diagnostic versus prediction), input needed (supervised versus unsupervised learning), or mathematical background of approaches used (linear versus non-linear, symmetric versus non-symmetric). The difference between diagnostic versus prediction is whether an AI scanning a picture of a mole can decide if this mole *is* cancerous, or whether it *will become* cancerous. There are specialised approaches for both, but for the purpose of this primer, it suffices to know that they exist, and the general principles and challenges remain the same. An exploratory approach aiming to define subtypes of moles could be unsupervised (without the need for prior grouping, the groups found might or might not overlap with being benign or malign), while an approach actively seeking to separate cancerous versus benign moles would be supervised: it would learn from a dataset of moles labelled as "cancerous" or "non-cancerous". Alas, there is no similarly simple explanation for the mathematical theories involved, but we can see them as an increasingly complex modelling, with linear more simplistic than non-linear, and symmetric more simplistic than non-symmetric. Simplicity is ideal: if a simple model can explain your data, this is the ideal case, and more complex models should only be used if simple approaches fail.

1.2. Curve fitting – The path to the optimal solution

AI and ML approaches solve problems by optimizing complex forms of curve fitting. For example, data points for two groups (distinguished by shape and colour) are plotted in Fig. 1 according to hypothetical Features A and B, and the goal is to determine the relationship between these features and the grouping. In the simplest of cases, this will be a straight line. Realistically, more complex models will provide better fits for real data. The simple straight line will then be *under-fitted*: it will perform its task poorly, due to an overly simplistic model.

On the other end, a more complex model will always be able to provide a better fit for the given data, to the point of perfection. This is due to the nature of the method: it will always optimise the fit based on the data provided. Every input dataset will, however, be flawed in some ways, and can never be a perfect representation of the real world – therefore, external validation (replication of findings in a data set separately from the original data, ideally independent in means of population and collection) of each finding is mandatory for findings based on ML. For instance, a ML-model that performs near perfectly on its training data set and badly on external (validation) data is *over-fitted*: it is optimised for specific data rather than the general case and hence useless for all practical purposes (Fig. 1B).

This shows why for the development of an AI, at least two independent datasets are needed: a *training* dataset, on which the "learning" (or model optimisation) takes place, and a *validation* dataset, on which the derived AI is validated to show it is not over-fitted and generalises to unseen data. Choosing ideal training and validation datasets defines success and failure of any AI approach, and it cannot be overstated how carefully these need to be chosen. Ideally, validation and training datasets should be completely independent – for example, data collected at a different hospital by separate examiners, on another continent, and so on, to prove maximal generalizability of the AI. Often, this will not realistically be possible, at the very least, however, if only a single dataset is available, it must be split, and a proportion kept separately as validation dataset. In this case, the results will not be immediately generalisable: any AI is only ever validated for the data is has been tested on.

1.3. Unsupervised learning

Unsupervised learning uses a set of methods that do not depend on knowledge of outcome, for example, modern evolutionary trees are based on clustering of genetic resemblance. The outcome (the tree) was unknown to begin with and is a sole result of the principles of the algorithm. Unsupervised learning methods can generally be said to search for symmetry, order, or structure in data (see Fig. 2 for a visualisation of various methods (Pedregosa et al., 2011)). This can be exemplified by using two-dimensional data: if you can visually observe patterns (like in the rows 1,2,4,5 of Fig. 2), the right unsupervised learning will also find these patterns – only that we usually use high-dimensional, not two-dimensional data, which cannot be plotted easily. An advantage is that a symmetric form is less prone to overfitting and results can often more easily be transferred to external data. The shapes constructed by unsupervised ML algorithms are also easier to interpret, providing the opportunity to shape our understanding of biology.

Some of the most-used methods include dimension-reduction approaches like principal component analysis (Pearson, 1901); hierarchical clustering, like Ward (Ward, 1963) or Neighbour-joining (Saitou and Nei, 1987); fast, heuristic methods ideally suited for very large datasets like k-means (MacQueen, 1967); and density-based approaches like DBSCAN (density-based spatial clustering of applications with noise) (Ester et al., 1996). As shown in Fig. 2, none of these are inherently better than others, how they perform will depend on the structure of your data.

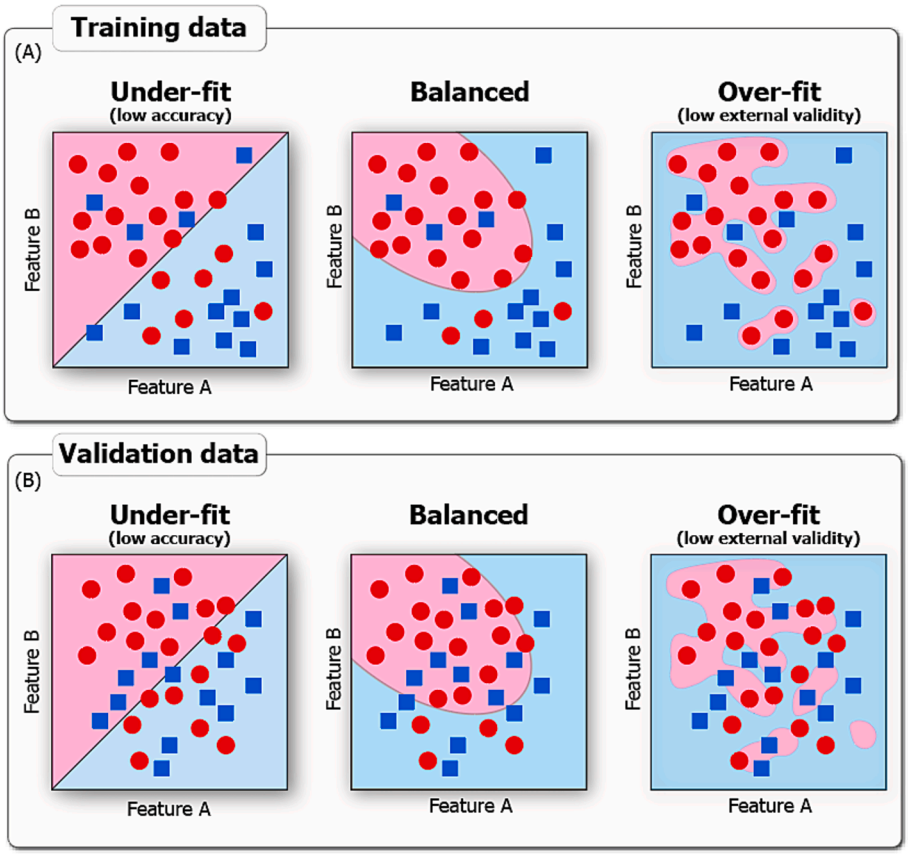

Fig. 1. Curve-fitting models. These are artificial data for illustration only. The linear underfitting, both in the training data set (A) and in the validation data set (B) cannot capture the nonlinear shape of the data, while the over-fitting can perfectly separate the training data set from the validation data set, but this model leads to a reduced performance on another data set. Accepting the error of the balance fit leads to better generalizability overall and thus to a more robust model.

1.4. Supervised learning

Supervised learning depends on a training dataset with known outcomes or classification, for example, pictures of moles and the information if these are malignant or not. The advantage is that they are able to uncover non-linear, non-symmetrical relationships, however, the concept of fully optimising a function on a given training dataset makes them prone to over-fitting or adjusting to bias in the dataset (Spisak, 2022), which is not always easily spotted.

In its simplest (though by no means simple) form, a supervised machine learning can be a multiple linear regression that can be replaced by regularised linear regression models that can control model complexity or non-linear regression models if warranted by the data.

Complex, non-linear, non-symmetrical methods have been responsible for most of the recent fast advances in the field like facial detection and other complex image processing tasks, protein folding and form the basis of *ChatGPT*. Support Vector Machines (Cortes and Vapnik, 1995), k-nearest neighbours (Cover and Hart, 1967), and Hidden Markov Models (Rabiner and Juang, 1986) and Markov Chain Monte Carlo methods (Hamra et al., 2013) are some of the most commonly used examples. They are based on highly distinct mathematical models and assumptions, and their use differs, as they have been demonstrated to excel in different fields. Here, as a general example for supervised machine learning, we will focus on neural networks (Hopfield, 1982), since they have gained most attention, which has been particularly the case through the development of Deep Learning (Hinton et al., 2006), a variant of neural networks.

1.5. Neural networks and deep learning

Neural networks are a method that has been developed mimicking biological process, similar to evolutionary algorithms. A neural network in its simplest form consists of an input layer (the features of your data), a hidden layer, and an output layer (the classification or prediction) (McCulloch and Pitts, 1943). The multiple nodes (perceptrons) of these layers are interconnected (see Fig. 3), and the number of hidden nodes will determine the complexity of the network (as well as its risk of overfitting). A node is modelled after a synapse, based on input signals, it provides forward a signal, or not (inspired by a post-synaptic potential either reaching the activation threshold at the axon hillock or not). In our (by now slightly overused) example of pictures of benign and malignant moles, the input layer would be those pictures, the output layer would be the label provided ("cancerous" or "non-cancerous"). The hidden layer is the "ghost in the machine": the black box classifying, with no explanation how the classification was reached.

Many recent advances in machine learning have been rooted in deep learning, using the principles of neural networks with multiple hidden layers, and hundreds, thousands, millions, and billions of nodes. Each layer can be interpreted as a level of abstraction, and the algorithms allow the network to choose and refine its own structure (LeCun et al., 2015). Deep learning has overtaken many classic AI techniques, and is behind many of the most impressive recent successes, including general use like chatGPT and its successor, GPT-4 (OpenAI, 2023) and scientific tasks like protein folding (Jumper et al., 2021). However, deep learning is not a "magic bullet", depending on dataset structure, classical machine learning models (like tree-based models) can outperform them in many real scenarios (Grinsztajn et al., 2022).

Supervised (deep) learning approaches are generally "black boxes".

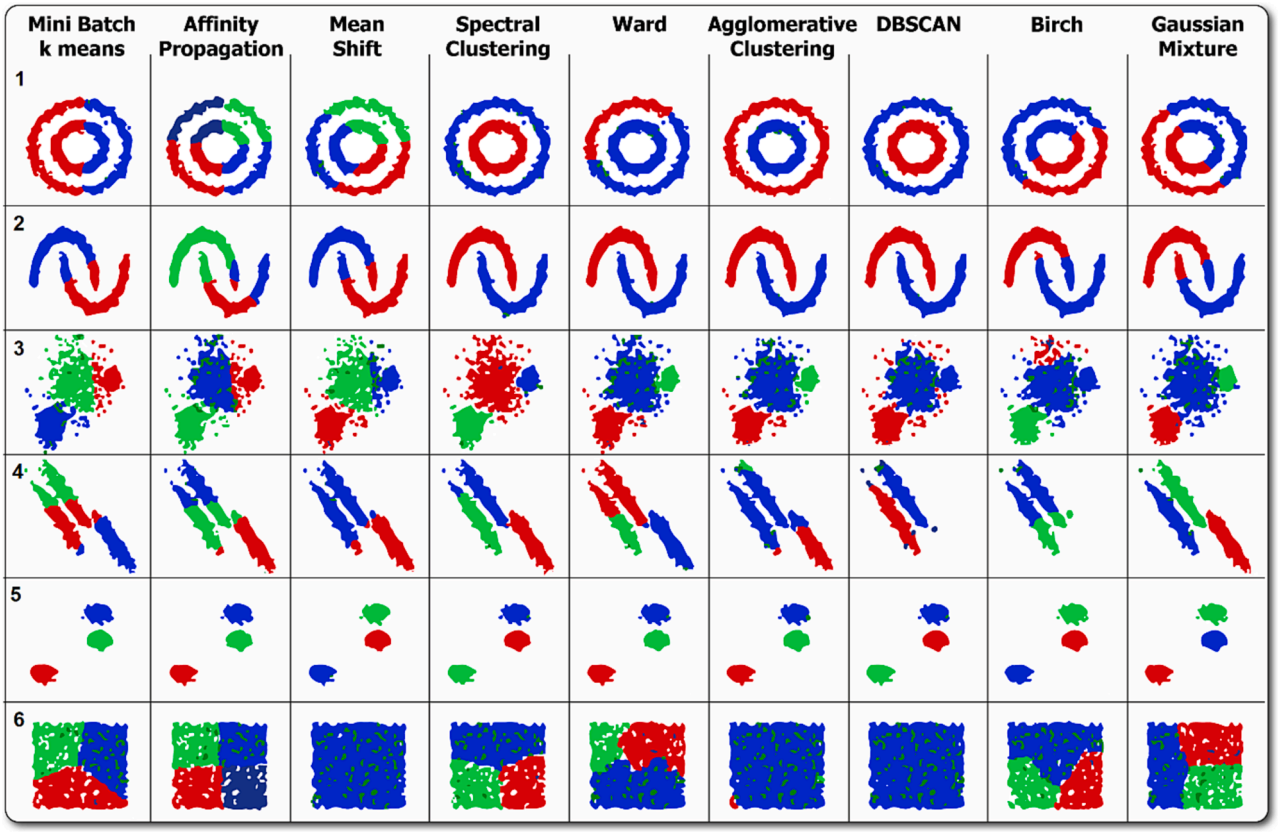

Fig. 2. The ability of various cluster methods to separate non-linear datasets, simulated using scikit-learn (Pedregosa et al., 2011). These are artificial data for illustration only. The ability of the clustering method to separate your data will depend on the structure of your data, which, in multiple dimensions, cannot be plotted and therefore remains usually unknown. The lowest panel shows that, even for uniform data distribution, the algorithm will always present a "best" solution.

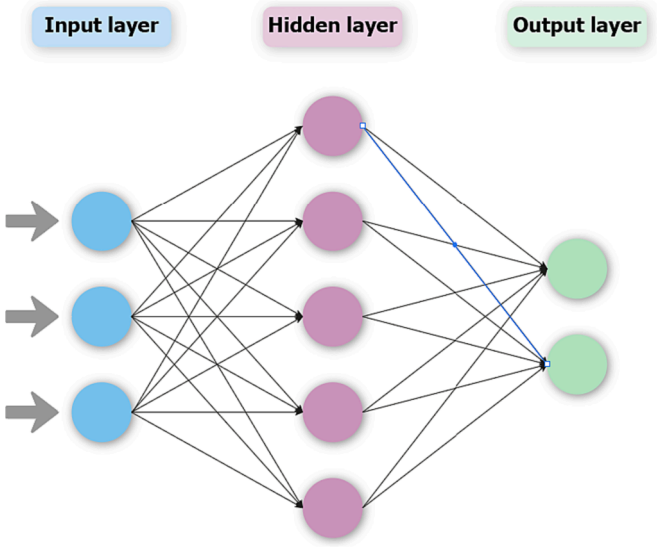

Fig. 3. Simple example of a three-layer perceptron network. The input layer provides the knowledge we have, the output layer the result. The hidden layer is the "ghost in the machine", the black box-part of the decision-making process. Modern deep learning AIs can use multiple hidden layers with almost endless nodes, GPT-4 uses billions of nodes, and trillions of parameters (OpenAI, 2023).

They do not follow clear and identifiable rules but greedily learn from examples. This is similar to how humans learn intuitively compared to how we learn at school: you are often more likely to know grammatic rules of a second language you learned at school than your native language. At the same time, you might well be better at grammar in your native language, as you do not base your decision on the (simplified) grammatical rules but on practical usage experience of the common case and the many exemptions.

2. Risks

With growing use, one of AI approaches inherent features – its black box approach – is becoming a growing concern. When used for exploratory research, AI can find novel targets that are subsequently investigated. In this case, a black box is of little concern, as the meaningfulness of biological pathways and medical sciences will be described in subsequent experiments. However, AI is now used to assist human decision making (not only) in medicine, where it can be of high importance to know why and how a decision was formed.

Both the optimisation to training data and the subsequent application of AI by a wide range of users can lead to unintended consequences. An AI trained to detect malignant skin lesions has been shown to have learned that the presence of rulers in pictures shows malignance, as they are used more often for scaling in pictures of malignant lesions, biasing the training data set (Narla et al., 2018). On the user end, dermatologists will often use markers to highlight the lesion on a picture, which can also increase the chance of the AI classifying the lesion as malignant (Winkler et al., 2019). An AI predicting complications of pneumonia wrongly suggested patients with asthma as low risk of pneumonia (while the opposite is the case) (Caruana et al., 2015). While these were easy-to-uncover teething problems, they illustrate the unforeseen consequences of invisible bias in data, and the fatal consequences blind trust in AI-assisted decision making can have.

AIs are trained on large datasets, and results can therefore only be as

good as the input data ("garbage in, garbage out" rule of databases). This means that during data collection, careful consideration must be given to avoid errors in labelling or similar aspects that can lead to distorted results. As AI methods become more widely available, they will often be used on datasets that are fundamentally of a too-low quality. Results should be discarded, as only high-quality input data can lead to results that we can have enough trust in to warrant further use or research.

Even the best and well-collated existing large datasets are biased. This is partly based in human history. For example, image processing AIs often learn on ImageNet, but this database is heavily biased towards data from the US and Western Europe, while China and India, comprising more than a third of the world's population, only constituted just 3 % of ImageNet data in 2017 (Shankar et al., 2017). Since AIs aim to optimise their accuracy, ignoring minorities is a frequent feature (Zou and Schiebinger, 2018).

In addition, by learning from human annotated data, the AI will learn our biases. In the beginning of this piece, we used the term "replace the experience of *one* doctor with that of *thousands* of doctors". This will also include replacing the *bias* of one doctor with that of thousands of doctors. While we often believe any computer decision is impartial and unbiased, it has been shown multiple times that this is not the case: internet search algorithms propagate gender stereotypes (Vlasceanu and Amodio, 2022), and AI-based decision making in US hospitals has been shown to disadvantage Black patients, negating them necessary care and treatment a similar White patient would have received (Obermeyer et al., 2019).

2.1. Explainable AI

One approach to reduce the impact of the risks and limitations is opening up the black box by using explainable AI (XAI) methods (Vilone and Longo, 2021). In contrast to the black box of classic AI, they aim to create a white box, where decision making can be traced and understood (Castelvecchi, 2016). XAIs aim to be transparent, interpretable and explainable, features which would make them ideal for assisted decision making in healthcare (Rieg et al., 2020). However, with growing complexity of AIs, explainability becomes near impossible to achieve, which can be understood when considering how for example neural networks aim to mimic, rather than understand human decision making – they do so as we cannot sufficiently explain how we come to decisions and cannot easily put our experience into simple rules. By training AIs to do similar, we end up with similarly complex decision making processes, that inherently cannot be explained in simple ways (Bhatt et al., 2020).

2.2. Reporting, transparency, reproducibility, and replicability

As, unfortunately, in all areas of research, there are concerns about transparent reporting and "spin" practices in machine learning, underlined by a recent systematic review (Andaur Navarro et al., 2023). There are no current gold standards on conduct and reporting of ML and AI medical research, partly of course, as they are just a method for multiple means. Generally, reporting guidelines on AI/ML will include elements like mentioning the use of ML, as well as describing the specific methods and purpose, methods to control for errors, and reporting of common model-metrics.

Specific reporting guidelines can be found in the REWARD EQUATOR network for clinical trial protocols (SPIRIT-AI (Rivera et al., 2020)) and clinical trials (CONSORT-AI (Liu et al., 2020)). When developing tools used for assistance in decision making, DECIDE-AI should be followed (Vasey et al., 2022).

A framework for reporting applicable to all medical AI research can be found in the MI-CLAIM checklist (Norgeot et al., 2020). Here, the authors also touch on the difficult topic of reproducibility. While, in some situations, ML/AI approaches can lead to findings that are easier to replicate as the effect sizes are higher (Spisak et al., 2023), making it more likely to achieve similar qualitative findings in a separate approach, that is not always the case (Marek et al., 2022). The black box nature and often commercial use of ML products can lead to reproducibility far lower compared to using traditional statistics (and it is not high there to begin with), as step-by-step replication cannot necessarily be done. This is compounded by the wide range of available methods (Hoffmann et al., 2021), making it difficult for external groups to independently replicate findings. The hypothesis-free, pattern-searching approach of AI is, as mentioned above, prone to over-fitting, which can be abused to present findings that are likely random as real – similar to p-hacking or HARKing techniques in traditional statistics. However, black box findings that are unexplainable, unreproducible, and unreplicable, can only be detrimental to the scientific progress, and oppose the principles of scientific research. We therefore note that conduct guidelines are needed, and ethical frameworks need to be developed which also take transparency of reporting data and code sharing and reproducibility into account.

2.3. A note on anthropomorphisms and hype

The field of AI/ML is, unfortunately, rich in anthropomorphising terms that install unrealistic expectations for readers and listeners. "We used a machine-learning approach to train a neural network with evolutionary improvement to achieve an artificial intelligence" is a sentence not unthinkable to read, or "brain-inspired deep neural networks with attention mechanism that learned complex hidden representations to achieve an artificial intelligence system for diagnosis of cancer". While there are historic reasons for each of these terms (often, they were developed by mimicking human or biological behaviours and strategies), they evoke science-fiction associations they do not warrant. This can go so far as to computer programmers (Metz, 2022) or journalists (Roose, 2023) believing the chatbot they converse with is sentient (or close to it). When looking at a robotic production arm in an industrial setting, we do not automatically think of the *Terminator*, in the same way, we should not think of *Blade Runner* when reading about *ChatGPT*. The literature and film versions of robots and AIs are general: they have fully mastered to mimic humans and are able to make decisions and alter their behaviour almost freely. The robotic production arm and *ChatGPT* only mimic one very specific task, and, without explicit instructions, will not do that, either.

3. Application in neuroscience

3.1. Machine learning in preclinical neuroscience

Identifying the neural elements that influence naturalistic behavioral motifs in freely moving animals stands as one of the paramount challenges in contemporary neuroscience. To unravel these mechanisms, a variety of sophisticated ML/AI approaches are being formulated and utilized to discern behavioral patterns in rodents. (Fig. 4A). The detailed objective annotation of behavioural trajectories in real time without known influencing variables such as day and night phase or experimenters are crucial characteristics of modern methods (Sorge et al., 2014; Sadler et al., 2022). The capability to automatically extract behaviors in rodents is a developmental leap that can fulfil this requirement. In the age of machine and deep learning, it is possible to extract and quantify an almost infinite number of behavioural variables, to decompose behaviours into categories, subcategories and into minute behavioural sequences. However, the booming field of behavioural neuroethology still has limitations because the community has not yet consolidated, developed and applied methods, which translates to an insufficient transfer of models from lab to lab. This arises from inadequately established benchmarking and the scarce availability of extensive, thoroughly annotated data sets. In addition, the extraction of numerous variables correlates with an increasing amount of data, which requires data organisation, transfer and storage options. This is associated with a lack of platforms that enable the sharing of large data sets,

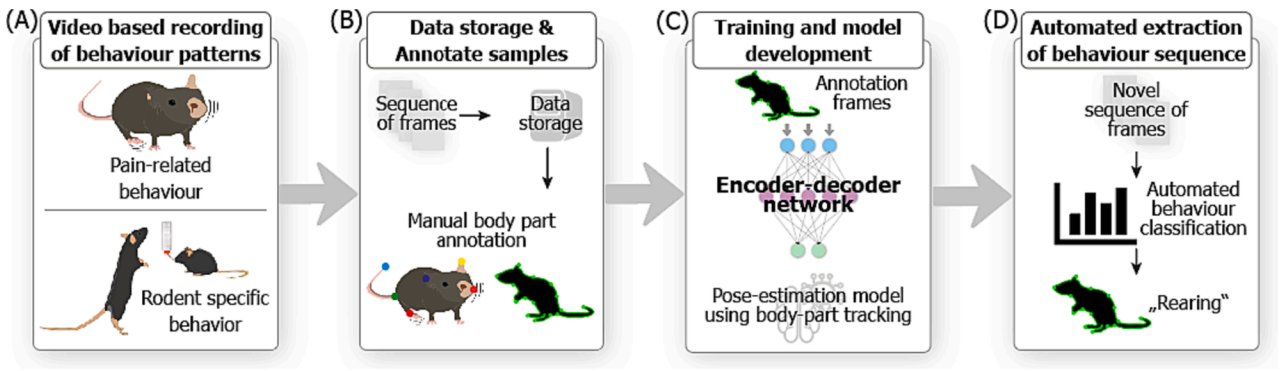

Fig. 4. Applications of machine learning in animal behavioural analyses. (A) Using AI/ML approaches, rodent-specific behaviours can be identified, isolated, and characterised. (B) Due to the enormous amount of data collected, data storage and the manual annotation of data sets are challenges in this method. By manually annotating body parts of rodents, such as tails, paws, ears or even parts of the face, (C) estimate body posture models over time can be generated via an encoder-decoder network and in the next step, (D) behavioural components can be automatically detected and named.

similar to sequencing databases in Omics (e.g., https://www.omicsdi.org and (Conesa and Beck, 2019)). In addition, most behavioural research labs have limited access to the latest tools for extracting and analysing behaviour, as their implementation requires advanced computer skills.

The automated identification of behavioral motifs (stereotypical, sub-second) in most protocols unfolds through a graduated process and can be categorized into supervised and unsupervised approaches (as previously mentioned). In supervised approaches (e.g., JAABA (Kabra et al., 2013)), a user trains algorithms to recognise behavioural motives. In contrast, the unsupervised method (e.g., MoSeq (Wiltschko et al., 2015)) separates individual vi. deo sequences into behavioural syllables without bias. The current development of applications that use indirect methods for behavioural extraction has emerged. Unsupervised or supervised ML approaches are also used in this indirect approach, the latter being the more common. In this method, virtual body landmarks of the animal (e.g., ears, paw, tail, etc, see Fig. 4B) are used to estimate body posture over time. Various open-source programmes can follow this approach. Examples of these tools include DeepLabCut (Mathis et al., 2018; Lauer et al., 2022), SLEAP (Pereira et al., 2022), DANNCE (Dunn et al., 2021) or Anipose (Karashchuk et al., 2021).

Initiating the identification of behavioral patterns begins with vi. deo recording. In recent years, the enhancement in camera image quality, coupled with a substantial reduction in initial costs, has rendered this a feasible option for a wide-ranging community. This is especially true for the vi. deo sampling rate (frames per second), which is necessary to extract behavioural sequences that occur in the sub-second range (e.g., hind paw withdrawal response to a stimulus of different modality) and recording spectra (e.g., infrared). But also, the depth of field is a crucial characteristic, which, especially during observation in the home cage setting, is needed to track individual animals in three-dimensional space and to isolate complex behavioural patterns. Here, 3D approaches can also be helpful, using multiple cameras with different viewing angles to refine behaviour estimation in complex environments (e.g. laboratory cages) with multiple animals (Nath et al., 2019). Point estimation models allow not only individual observation but also behavioural extraction of multiple animals within the same vi. deo (same cage) sequence and the associated social interaction (Lauer et al., 2022; Pereira et al., 2022). The construction of these markerless point estimation models employs neural networks with an encoder-decoder architecture to generate probability density diagrams. These diagrams are derived from features that the network is required to learn. Output diagram shows the probability of the existence of the learned feature. The resulting probability densities will be used for localisation, which is the basis for whole-body or whole-limb point skeletons for subsequent pose estimation and behaviour classification. In order to extract more nuanced behavioural signatures, such as grooming, straightening, rearing, social interaction from the point probabilities, advanced tracking algorithms are used in combination with other algorithms (e.g. UMAP (McInnes et al., 2018), random forest classifiers), or deep-learning models (e.g., recurrent convolutional neural networks). Tools used in this field to extract and quantify information on attitude and behaviour include SiMBA (Nilsson et al., 2020), B-SOiD (Hsu and Yttri, 2021), MoSeq (Wiltschko et al., 2015) and uBAM (Brattoli et al., 2021), among others. The possibility of longitudinal 24/7 extraction of naturalistic behaviour, hypothesis-driven modulation of cage environment (e.g. day/night cycle, enrichment, selective socialisation of cage mates with different health status (Segelcke et al., 2023)) and integrating state-of-the-art optogenetic tools for the targeted modulation of neuronal structures makes such approaches even more valuable (Hao et al., 2021).

Some of these tools and analysis pipelines have already been validated for the extraction of highly complex pain-related behaviour. Pain-related behaviours can be expressed in a variety of behavioural ways, but most assays have focused on experimental stimulation of the hind paw with noxious or non-noxious stimuli of different modalities (e.g., mechanical, thermal) and the resulting paw withdrawal response has established itself as the most common method for detecting pain-related behaviours (Deuis et al., 2017). Present ML approaches concentrate on enhancing the binary assessment of the withdrawal response by automating the reflection and affective components, aiming to isolate signatures differentiating between noxious and non-noxious stimuli. In recent work, paw and body movement features can be automatically extracted from behavioural trajectories using e.g., PAWS (Pain Assessment at Withdrawal Speeds, based on SLEAP) to then identify information from paw kinematics (Abdus-Saboor et al., 2019; Jones et al., 2020). Based on these data, a univariate pain score was developed using ordinal logistic regression for different harmless and noxious stimuli at the posterior paw and validated with basolateral amygdala activation (Jones et al., 2020).

Taken together, ML/AI pipelines for automated behavioural analysis have proven to be extremely powerful in different research directions, with only a subset of current pain research in mice described here as an example. Increased implementation of these automated behavioural approaches (preferably using comparable systems) can, consequently, increase the efficiency and translational potential of preclinical investigations and improve their reproducibility. For scientists working in neuro-behavioural research, there are unprecedented opportunities in implementing automated behavioural analysis tools. Increased tool implementation can consequently enhance the efficiency and translational potential of preclinical investigations. However, a real improvement in the replicability and reproducibility of data from such innovative approaches can only be achieved in the long term based on comparable or uniform standards: a challenge that must be met in the

future.

3.2. Machine learning in neuro-gastroenterology

Neuro-gastroenterology is the study of functional gut disorders such as irritable bowel syndrome (IBS) and functional dyspepsia, which are prevalent worldwide and can be challenging for clinicians to diagnose and treat and are critical for public health as they result in disability, impaired quality of life, and economic burden (Black and Ford, 2020).

At its core is the understanding of the gut-brain axis, which describes the complex, bi-directional communication, and interaction between the central and enteric nervous system (Carabotti et al., 2015). Neuro-gastroenterology is a field with challenges that make it particularly suited to attempt machine learning approaches: for better understanding of the underlying mechanisms, using microbiomic and metabolomic datasets results in high-dimensional data, which needs to be integrated with multiple levels of patient-reported outcomes or imaging data (fMRI). Such complex, multi-layered data can be mined using ML (Kaur et al., 2021).

There is increasing evidence that gut microbiota plays a key role in the regulation of the gut-brain axis. In addition to their local interactions with intestinal cells and the enteric nervous system, microbes in the gut have also been shown to modulate the central nervous system through neuroendocrine and metabolic pathways (Martin et al., 2018). It is becoming clear that the composition of the gut microbiome can therefore influence a wide range of disorders – proximal somatoform gastrointestinal disorders such as functional dyspepsia and IBS, but also mental health disorders such as depression (Morais et al., 2021). Using ML, microbial signatures have been shown to potentially play a role in psychological distress in IBS (Peter et al., 2018), depression phenotype (Stevens et al., 2021), amnestic mild cognitive impairment (Liu et al., 2021), autism-spectrum disorders (Wu et al., 2020), and others. A pilot clinical study demonstrated how using ML to personalise nutritional strategy based on individual gut microbiome features could lead the way towards a personalised treatment for IBS (Ghaffari et al., 2022). In this study, individual microbiome modulation through diet significantly improves IBS-related symptoms in patients with IBS-mixed over regular, non-individualised IBS diet (Karakan et al., 2022). While all these findings should be seen as serendipitous, exploratory findings for now, they might advance our understanding of the generation, and potential treatment, of these diseases in previously unexpected ways.

AI-assisted decision-making might in the future add to clinical algorithms (Kordi et al., 2022), although current findings need to be independently validated and replicated. However, we are starting to learn crucial lessons along the way: for example, IBS-constipation and functional constipation have been treated as distinct conditions, thought to have distinct pathophysiology. Using an ML approach to compare the accuracy of diagnostic models for IBS-constipation and functional constipation based on 'uni-symptomatic' versus 'syndromic' models, Ruffle et al (Ruffle et al., 2021) have shown that syndromic models do not significantly improve diagnostic accuracy, which suggests that they are not separate conditions but a single syndrome within one clinical spectrum.

Integration of structural and functional brain imaging into neuro-gastroenterology will lead to a deeper understanding of disease mechanisms, and a better understanding of the microbiome-gut-brain axis (Mayer et al., 2019). Using a support vector machine ML approach, Mao et al (Mao et al., 2020) showed an altered resting-state functional connectivity and effective connectivity of the habenula for IBS patients compared to healthy controls, advancing our understanding of the brain regions involved in IBS.

Taken together, neuro-gastroenterology is a field that can certainly profit from the application of ML approaches. However, current studies often suffer from low reporting quality, and the complex nature of data involved calls for the creation of larger, multi-site consortia to generate reliable, high-quality, multi-dimensional data of high external validity (Mayer et al., 2019). We look forward to the findings of prospectively designed and registered studies, which will provide the first confirmatory results in the field (Berentsen et al., 2020).

3.3. AI in the intersection of cognitive, computational, and clinical neurosciences

AI and cognitive neuroscience live in a symbiotic relationship. While the former continuously draws inspiration from our knowledge of biological neural systems to develop artificial neural networks, the latter harnesses the power of AI to expand our understanding of these biological systems (Kriegeskorte and Douglas, 2018). Examples range from the use of computational models based on reinforcement learning algorithms or recurrent neural networks to model human adaptive behaviour and decision making (Gläscher et al., 2010; Ito et al., 2022), to deep neural networks that help us better understand and decode how brain activity represents images viewed (Seeliger et al., 2018), and words heard or spoken by (Anumanchipalli et al., 2019; Goldstein et al., 2022) human participants.

In addition to enhancing our understanding of micro- and macro-scale neurocomputational processes, AI/ML have the potential to open up new avenues for translational and clinical research. ML-based predictive models, commonly referred to as "neural signatures" or "neuromarkers", integrate information from complex neural measures (fMRI, EEG, MEG, etc.) to decode and predict various clinical and behavioural traits or states.

Several studies aim to construct neuromarkers that can directly diagnose or characterise various clinical conditions (de Vos et al., 2018; Horien et al., 2022; Jiang et al., 2023). In such studies, however, it often becomes hard to disentangle what is being modelled because of the multidimensional and heterogeneous nature of clinical conditions and co-occurring health conditions (e.g., co-morbidities, medication use). Neuromarker research has thus turned towards the so-called "component process approach" (Woo et al., 2017), which aims to first develop neural signatures for basic "component processes", i.e. basic traits or states that can be examined in standardised circumstances and even experimentally manipulated in some cases. The resulting neural signatures can serve as robust and explainable intermediate features for the modelling of multiple clinical conditions. One of the pioneering examples is the Neurologic Pain Signature (NPS (Wager et al., 2013)), a machine learning model that derives an objective readout of ongoing pain experience from brain activity, as measured by fMRI. The NPS has been extensively validated by a series of studies and was found to display high reliability, broad external validity, and strong effect sizes in large independent samples (Zunhammer et al., 2018; Han et al., 2022). Task-elicited brain activity has also been found to be predictive for vicarious pain (Zhou et al., 2020), fear (Zhou et al., 2021), negative affect (Chang et al., 2015), craving (Garrison et al., 2023), reward (Speer et al., 2023) and many other states and traits. Multivariate ML models can also capitalise on brain activity measured in lack of any explicit stimulation (resting state), or even on brain morphology, to predict individual traits like pain sensitivity (Spisak et al., 2020; Kotikalapudi et al., 2023), learning (Kincses et al., 2023), cognition (Sripada et al., 2020), intellectual capacity (Tong et al., 2022), and others.

While ML-based brain signatures can reach unprecedented effect sizes (Hedges $g = 2.3$ in case of the NPS), predictive modelling itself is not a magic bullet. The lack of external validation and bad methodological practice lead, in many cases, to overly optimistic performance estimates and unrealistic expectations regarding the usefulness of such models (Sui et al., 2020; Varoquaux and Cheplygina, 2022). Just like traditional univariate analyses, such low-performing models still suffer from limited power, replicability, and predictive utility even with sample sizes in the thousands (Marek et al., 2022; Spisak et al., 2023). Another problem is that such brain-based models are susceptible to capture spurious or out-of-interest associations that can be detrimental to the model's clinical validity and generalizability and lead to

sensitivity to artefacts – in practice, this can mean minority-disadvantaging or racially biased models (Spisak, 2022).

In summary, AI holds immense potential not only for expanding our understanding of how the brain works but also for making this knowledge applicable in clinical contexts and to complement existing clinical approaches. However, to realise this potential, neuromarkers of the future must overcome significant challenges, such as ensuring broad generalizability across diverse contexts, promoting equity across subpopulations, and developing models with high neuroscientific validity and interpretability.

4. Resources and further reading

There are multiple starting points to experimenting with ML, for a user experienced in using Python, we highly recommend *scikit-learn* (Pedregosa et al., 2011). Its rich online resources and as well as its focus on essential methods make it a great place for beginners that still goes a long way. Fig. 2 has been created using sample data from scikit-learn. For a (pun intended) deeper dive, *tensorflow* (Ramsundar, 2018) and *PyTorch* (Ketkar, 2017) are open-source platforms for deep learning by Google and OpenAI.

A developing quasi-standard in biological data science is R statistical computing (R Core Team, 2022). R is an incredibly rich open-source project, with endless resources for biomedical sciences. A dedicated online community has created packages (collections of functions) for almost every possible task. For machine learning specifically, this includes for example *caret* (for regression models), *e1071* (for k-nearest neighbours and support vector machines), *neuralnet* (for neural networks), and *keras* (for deep learning). The advantage of using R is that is does not stop at ML: there are excellent tools for any aspect of biomedical data, many of them have been collated within the Bioconductor project (Gentleman et al., 2004). Visualisation of any plot can be achieved using *ggplot2*, and *shinyapps* allow construction of web-based user interfaces. There are also commercial packages for ML, *Matlab* for example has a valuable ML toolbox and is used for both commercial and research applications.

We have aimed to provide a short introduction to machine learning in general, and its application in neurosciences, but of course, this has remained somewhat superficial. Others have taken similar, but complimentary approaches. Connor (Connor, 2019) provides a more methods-focussed introduction, while others focus on practical aspects for application in, e.g., pain research (Lötsch et al., 2022). For deeper reads, there are many. For using R in biomedical research, the University of California, Riverside, has collated excellent learning materials (GEN242, 2022). Finally, Zou and Schiebinger (Zou and Schiebinger, 2018) summarise bias inherent in human data and what it means for AI in a plastic way.

5. Future directions and closing remarks

ML and AI have multiple inherent risks and fallacies; however, their success is undeniable, and for better or worse, their use in biomedical sciences is unstoppable at this point. The recent advancements in deep learning, as exemplified in the changes between GPT-3 and GPT-4 have been at an unforeseen pace, and teething problems aside, AIs will soon outperform humans in countless tasks.

We would argue that as most AIs remain a black box, with decision making that can be influenced by human bias or unexpected elements of training data, their best use is for hypothesis generation, exploratory, and discovery research. Their use in medical decision making depends on the context and, in many circumstances can be problematic, while ethical issues are not resolved, and explainable AI has not moved forward significantly. Since currently, AIs remain black boxes, these should at most be one of multiple indicators for human-centred decision making.

CRediT authorship contribution statement

Fakhirah Badrulhisham: Writing – review & editing, Writing – original draft. **Esther Pogatzki-Zahn:** Writing – review & editing. **Daniel Segelcke:** Visualization, Writing – review & editing, Writing – original draft. **Tamas Spisak:** Writing – review & editing, Writing – original draft. **Jan Vollert:** Visualization, Writing – review & editing, Conceptualization, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Abdus-Saboor, I., Fried, N.T., Lay, M., Burdge, J., Swanson, K., Fischer, R., Jones, J., Dong, P., Cai, W., Guo, X., Tao, Y.-X., Bethea, J., Ma, M., Dong, X., Ding, L., Luo, W., 2019. Development of a Mouse Pain Scale Using Sub-second Behavioral Mapping and Statistical Modeling. Cell Reports 28, 1623–1634.e4.

Andaur Navarro, C.L., Damen, J.A., Takada, T., Nijman, S.W.J., Dhiman, P., Ma, J., Collins, G.S., Bajpai, R., Riley, R.D., Moons, K.G., Hooft, L., 2023. Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning-based prediction models. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2023.03.024.

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498.

Berentsen, B., Nagaraja, B.H., Teige, E.P., Lied, G.A., Lundervold, A.J., Lundervold, K., Steinsvik, E.K., Hillestad, E.R., Valeur, J., Brønstad, I., Gilja, O.H., Osnes, B., Hatlebakk, J.G., Haász, J., Labus, J., Gupta, A., Mayer, E.A., Benitez-Páez, A., Sanz, Y., Lundervold, A., Hausken, T., 2020. Study protocol of the Bergen brain-gut-microbiota-axis study: A prospective case-report characterization and dietary intervention study to evaluate the effects of microbiota alterations on cognition and anatomical and functional brain connectivity in patients with irritable bowel syndrome. Medicine 99, e21950.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M.F., Eckersley, P., 2020. Explainable machine learning in deployment, in: FAT* '20. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency : January 27-30, 2020, Barcelona, Spain. FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain. 27 01 2020 30 01 2020. The Association for Computing Machinery, New York, New York, pp. 648–657.

Black, C.J., Ford, A.C., 2020. Global burden of irritable bowel syndrome: trends, predictions and risk factors. Nature Reviews. Gastroenterology & Hepatology 17, 473–486.

Brattoli, B., Büchler, U., Dorkenwald, M., Reiser, P., Filli, L., Helmchen, F., Wahl, A.-S., Ommer, B., 2021. Unsupervised behaviour analysis and magnification (uBAM) using deep learning. Nat Mach Intell 3, 495–506.

Carabotti, M., Scirocco, A., Maselli, M.A., Severi, C., 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. Annals of Gastroenterology : Quarterly Publication of the Hellenic Society of Gastroenterology 28, 203–209.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible Models for HealthCare, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15: The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney NSW Australia. 10 08 2015 13 08 2015. ACM, New York, NY, pp. 1721–1730.

Castelvecchi, D., 2016. Can we open the black box of AI? Nature 538, 20–23.

Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., Wager, T.D., 2015. A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. PLoS Biology 13, e1002180.

Conesa, A., Beck, S., 2019. Making multi-omics data accessible to researchers. Sci Data 6, 251.

Connor, C.W., 2019. Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology 131, 1346–1359.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13, 21–27.

de Vos, F., Koini, M., Schouten, T.M., Seiler, S., van der Grond, J., Lechner, A., Schmidt, R., de Rooij, M., Rombouts, S.A.R.B., 2018. A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. NeuroImage 167, 62–72.

Deuis, J.R., Dvorakova, L.S., Vetter, I., 2017. Methods Used to Evaluate Pain Behaviors in Rodents. Frontiers in Molecular Neuroscience 10, 284.

Dill, K.A., Ozkan, S.B., Shell, M.S., Weikl, T.R., 2008. The protein folding problem. Annual Review of Biophysics 37, 289–316.

Dunn, T.W., Marshall, J.D., Severson, K.S., Aldarondo, D.E., Hildebrand, D.G.C., Chettih, S.N., Wang, W.L., Gellis, A.J., Carlson, D.E., Aronov, D., Freiwald, W.A., Wang, F., Ölveczky, B.P., 2021. Geometric deep learning enables 3D kinematic profiling across species and environments. Nature Methods 18, 564–573.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226–231.

Garrison, K.A., Sinha, R., Potenza, M.N., Gao, S., Liang, Q., Lacadie, C., Scheinost, D., 2023. Transdiagnostic Connectome-Based Prediction of Craving. The American Journal of Psychiatry appiajp21121207.

GEN242, 2022. Introduction. https://girke.bioinformatics.ucr.edu/GEN242/about/introduction/. Accessed 17 March 2023.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5, R80.

Ghaffari, P., Shoaie, S., Nielsen, L.K., 2022. Irritable bowel syndrome and microbiome; Switching from conventional diagnosis and therapies to personalized interventions. J Transl Med 20, 173.

Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K.A., Devinsky, O., Hasson, U., 2022. Shared computational principles for language processing in humans and deep language models. Nature Neuroscience 25, 369–380.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on tabular data? https://arxiv.org/pdf/2207.08815.

Hamra, G., MacLehose, R., Richardson, D., 2013. Markov chain Monte Carlo: an introduction for epidemiologists. International Journal of Epidemiology 42, 627–634.

Han, X., Ashar, Y.K., Kragel, P., Petre, B., Schelkun, V., Atlas, L.Y., Chang, L.J., Jepma, M., Koban, L., Losin, E.A.R., Roy, M., Woo, C.-W., Wager, T.D., 2022. Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. NeuroImage 247, 118844.

Hao, Y., Thomas, A.M., Li, N., 2021. Fully autonomous mouse behavioral and optogenetic experiments in home-cage. eLife 10.

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554.

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., Boulesteix, A.-L., 2021. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. Royal Society Open Science 8, 201925.

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.

Horien, C., Floris, D.L., Greene, A.S., Noble, S., Rolison, M., Tejavibulya, L., O'Connor, D., McPartland, J.C., Scheinost, D., Chawarska, K., Lake, E.M.R., Constable, R.T., 2022. Functional Connectome-Based Predictive Modeling in Autism. Biological Psychiatry 92, 626–642.

Hsu, A.I., Yttri, E.A., 2021. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. Nature Communications 12, 5188.

Ito, T., Yang, G.R., Laurent, P., Schultz, D.H., Cole, M.W., 2022. Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. Nat Commun 13, 673.

Jiang, Y., Wang, J., Zhou, E., Palaniyappan, L., Luo, C., Ji, G., Yang, J., Wang, Y., Zhang, Y., Huang, C.-C., Tsai, S.-J., Chang, X., Xie, C., Zhang, W., Lv, J., Chen, D.i., Shen, C., Wu, X., Zhang, B., Kuang, N., Sun, Y.-J., Kang, J., Zhang, J., Huang, H., He, H., Duan, M., Tang, Y., Zhang, T., Li, C., Yu, X., Si, T., Yue, W., Liu, Z., Cui, L.-B., Wang, K., Cheng, J., Lin, C.-P., Yao, D., Cheng, W., Feng, J., 2023. Neuroimaging biomarkers define neurophysiological subtypes with distinct trajectories in schizophrenia. Nat. Mental Health 1, 186–199.

Jones, J.M., Foster, W., Twomey, C.R., Burdge, J., Ahmed, O.M., Pereira, T.D., Wojick, J. A., Corder, G., Plotkin, J.B., Abdus-Saboor, I., 2020. A machine-vision approach for automated pain measurement at millisecond timescales. eLife 9.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., Branson, K., 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. Nature Methods 10, 64–67.

Karakan, T., Gundogdu, A., Alagözlü, H., Ekmen, N., Ozgul, S., Tunali, V., Hora, M., Beyazgul, D., Nalbantoglu, O.U., 2022. Artificial intelligence-based personalized diet: A pilot clinical study for irritable bowel syndrome. Gut Microbes 14, 2138672.

Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., Tuthill, J.C., 2021. Anipose: A toolkit for robust markerless 3D pose estimation. Cell Reports 36, 109730.

Kaur, H., Singh, Y., Singh, S., Singh, R.B., 2021. Gut microbiome-mediated epigenetic regulation of brain disorder and application of machine learning for multi-omics data analysis. Genome 64, 355–371.

Ketkar, N., 2017. Deep Learning with Python: A Hands-on Introduction, 1st ed. Apress, Berkeley, CA, Online-Ressourcen.

Kincses, B., Forkmann, K., Schlitt, F., Pawlik, R., Schmidt, K., Timmann, D., Elsenbruch, S., Wiech, K., Bingel, U., Spisak, T., 2023. RCPL preprint: An externally validated resting-state brain connectivity signature of pain-related learning.

Kordi, M., Dehghan, M.J., Shayesteh, A.A., Azizi, A., 2022. The impact of artificial intelligence algorithms on management of patients with irritable bowel syndrome: A systematic review. Informatics in Medicine Unlocked 29, 100891.

Kotikalapudi, R., Kincses, B., Zunhammer, M., Schlitt, F., Asan, L., Schmidt-Wilcke, T., Kincses, Z., Bingel, U., Spisak, T., 2023. Brain morphology predicts individual sensitivity to pain: a multi-center machine learning approach. Pain.

Kriegeskorte, N., Douglas, P.K., 2018. Cognitive computational neuroscience. Nature Neuroscience 21, 1148–1160.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V.N., Lauder, G., Dulac, C., Mathis, M.W., Mathis, A., 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. Nature Methods 19, 496–504.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Liu, P., Jia, X.-Z., Chen, Y., Yu, Y., Zhang, K., Lin, Y.-J., Wang, B.-H., Peng, G.-P., 2021. Gut microbiota interacts with intrinsic brain activity of patients with amnestic mild cognitive impairment. CNS Neuroscience & Therapeutics 27, 163–173.

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ (clinical Research Ed.) 370, m3164.

Lötsch, J., Ultsch, A., Mayer, B., Kringel, D., 2022. Artificial intelligence and machine learning in pain research: a data scientometric analysis. Pain Reports 7, e1044.

Luger, G.F., 2004. Artificial intelligence: Structures and Strategies for Complex Problem Solving, 5th ed. Addison-Wesley, Harlow, p. 936.

MacQueen, J.B., 1967. Some Methods for Classification and Analysis of Multivariate Observations, in: Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.

Mao, C.P., Chen, F.R., Huo, J.H., Zhang, L., Zhang, G.R., Zhang, B., Zhou, X.Q., 2020. Altered resting-state functional connectivity and effective connectivity of the habenula in irritable bowel syndrome: A cross-sectional and machine learning study. Human Brain Mapping 41, 3655–3666.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N. A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660.

Martin, C.R., Osadchiy, V., Kalani, A., Mayer, E.A., 2018. The Brain-Gut-Microbiome Axis. Cellular and Molecular Gastroenterology and Hepatology 6, 133–148.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience 21, 1281–1289.

Mayer, E.A., Labus, J., Aziz, Q., Tracey, I., Kilpatrick, L., Elsenbruch, S., Schweinhardt, P., van Oudenhove, L., Borsook, D., 2019. Role of brain imaging in disorders of brain-gut interaction: a Rome Working Team Report. Gut 68, 1701–1715.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/pdf/1802.03426.

Metz, R., 2022. No, Google's AI is not sentient. CNN.

Morais, L.H., Schreiber, H.L., Mazmanian, S.K., 2021. The gut microbiota-brain axis in behaviour and brain disorders. Nat Rev Microbiol 19, 241–255.

Narla, A., Kuprel, B., Sarin, K., Novoa, R., Ko, J., 2018. Automated Classification of Skin Lesions: From Pixels to Practice. The Journal of Investigative Dermatology 138, 2108–2110.

Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W., 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nature Protocols 14, 2152–2176.

Nilsson, S.R.O., Goodwin, N.L., Choong, J.J., Hwang, S., Wright, H.R., Norville, Z.C., Tong, X., Lin, D., Bentzley, B.S., Eshel, N., McLaughlin, R.J., Golden, S.A., 2020. Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals.

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., Butte, A.J., 2020. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nature Medicine 26, 1320–1324.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science 366, 447–453.

OpenAI, 2023. GPT-4 Technical Report, 99 pp. https://arxiv.org/pdf/2303.08774.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of 2, 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., McKenzie-Smith, G.C., Mitelut, C.C., Castro, M.D., D'Uva, J., Kislin, M., Sanes, D.H., Kocher, S.D., Wang, S.-S.-H., Falkner, A.L., Shaevitz, J.W., Murthy, M., 2022. SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods 19, 486–495.

Peter, J., Fournier, C., Durdevic, M., Knoblich, L., Keip, B., Dejaco, C., Trauner, M., Moser, G., 2018. A Microbial Signature of Psychological Distress in Irritable Bowel Syndrome. Psychosomatic Medicine 80, 698–709.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. Austria, Vienna https://www.R-project.org/.

Rabiner, L., Juang, B., 1986. An introduction to hidden Markov models. IEEE ASSP Mag. 3, 4–16.

Ramsundar, R.Z.B.B., 2018. TensorFlow for Deep Learning. O'Reilly Media Inc [Place of publication not identified], 1 online resource.

Rieg, T., Frick, J., Baumgartl, H., Buettner, R., 2020. Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. PloS One 15, e0243615.

Rivera, S.C., Liu, X., Chan, A.-W., Denniston, A.K., Calvert, M.J., 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ (clinical Research Ed.) 370, m3210.

Roose, K., 2023. Why a Conversation With Bing's Chatbot Left Me Deeply Unsettled. The. New York times.

Ruffle, J.K., Tinkler, L., Emmett, C., Ford, A.C., Nachev, P., Aziz, Q., Farmer, A.D., Yiannakou, Y., 2021. Constipation Predominant Irritable Bowel Syndrome and Functional Constipation Are Not Discrete Disorders: A Machine Learning Approach. The American Journal of Gastroenterology 116, 142–151.

Sadler, K.E., Mogil, J.S., Stucky, C.L., 2022. Innovations and advances in modelling and measuring pain in animals. Nat Rev Neurosci 23, 70–85.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., van Gerven, M.A.J., 2018. Generative adversarial networks for reconstructing natural images from brain activity. NeuroImage 181, 775–785.

Segelcke, D., Linnemann, J., Pradier, B., Kronenberg, D., Stange, R., Richter, S.H., Görlich, D., Baldini, N., Di Pompo, G., Verri, W.A., Avnet, S., Pogatzki-Zahn, E.M., 2023. Behavioral Voluntary and Social Bioassays Enabling Identification of Complex and Sex-Dependent Pain-(-Related) Phenotypes in Rats with Bone Cancer. Cancers 15.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D., 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. https://arxiv.org/pdf/1711.08536.

Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J.C.S., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A.P., Quinn, T., Frasnelli, J., Svensson, C.I., Sternberg, W.F., Mogil, J.S., 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nature Methods 11, 629–632.

Speer, S.P.H., Keysers, C., Barrios, J.C., Teurlings, C.J.S., Smidts, A., Boksem, M.A.S., Wager, T.D., Gazzola, V., 2023. A multivariate brain signature for reward. NeuroImage 271, 119990.

Spisak, T., 2022. Statistical quantification of confounding bias in machine learning models. GigaScience 11.

Spisak, T., Kincses, B., Schlitt, F., Zunhammer, M., Schmidt-Wilcke, T., Kincses, Z.T., Bingel, U., 2020. Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. Nat Commun 11, 187.

Spisak, T., Bingel, U., Wager, T.D., 2023. Multivariate BWAS can be replicable with moderate sample sizes. Nature 615, E4–E7.

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W.K., Luciana, M., Weigard, A., Hyde, L.H., Heitzeg, M., 2020. Prediction of neurocognition in youth from resting state fMRI. Molecular Psychiatry 25, 3413–3421.

Stevens, B.R., Roesch, L., Thiago, P., Russell, J.T., Pepine, C.J., Holbert, R.C., Raizada, M. K., Triplett, E.W., 2021. Depression phenotype identified by using single nucleotide exact amplicon sequence variants of the human gut microbiome. Molecular Psychiatry 26, 4277–4287.

Sui, J., Jiang, R., Bustillo, J., Calhoun, V., 2020. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. Biological Psychiatry 88, 818–828.

Tong, X., Xie, H., Carlisle, N., Fonzo, G.A., Oathes, D.J., Jiang, J., Zhang, Y., 2022. Transdiagnostic connectome signatures from resting-state fMRI predict individual-level intellectual capacity. Translational Psychiatry 12, 367.

Varoquaux, G., Cheplygina, V., 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digital Medicine 5, 48.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B.A., Mathur, P., McCradden, M.D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D.S.W., Watkinson, P., Weber, W., Wheatstone, P., McCulloch, P., 2022. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine 28, 924–933.

Vilone, G., Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76, 89–106.

Vlasceanu, M., Amodio, D.M., 2022. Propagation of societal gender inequality by internet search algorithms. Proceedings of the National Academy of Sciences of the United States of America 119, e2204529119.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. The New England Journal of Medicine 368, 1388–1397.

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244.

Wiltschko, A.B., Johnson, M.J., Iurilli, G., Peterson, R.E., Katon, J.M., Pashkovski, S.L., Abraira, V.E., Adams, R.P., Datta, S.R., 2015. Mapping Sub-Second Structure in Mouse Behavior. Neuron 88, 1121–1135.

Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Haenssle, H.A., 2019. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatology 155, 1135–1141.

Woo, C.-W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience 20, 365–377.

Wu, T., Wang, H., Lu, W., Zhai, Q., Zhang, Q., Yuan, W., Gu, Z., Zhao, J., Zhang, H., Chen, W., 2020. Potential of gut microbiome for detection of autism spectrum disorder. Microbial Pathogenesis 149, 104568.

Zhou, F., Li, J., Zhao, W., Xu, L., Zheng, X., Fu, M., Yao, S., Kendrick, K.M., Wager, T.D., Becker, B., 2020. Empathic pain evoked by sensory and emotional-communicative cues share common and process-specific neural representations. eLife 9.

Zhou, F., Zhao, W., Qi, Z., Geng, Y., Yao, S., Kendrick, K.M., Wager, T.D., Becker, B., 2021. A distributed fMRI-based signature for the subjective experience of fear. Nat Commun 12, 6643.

Zou, J., Schiebinger, L., 2018. AI can be sexist and racist - it's time to make it fair. Nature 559, 324–326.

Zunhammer, M., Bingel, U., Wager, T.D., 2018. Placebo Effects on the Neurologic Pain Signature: A Meta-analysis of Individual Participant Functional Magnetic Resonance Imaging Data. JAMA Neurology 75, 1321–1330.