Computational modelling in disorders of consciousness: Closing the gap towards personalised models for restoring consciousness

Andrea I Luppi
ShiNung Ching
Michael N Diringer
et al.

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.
Computational modelling in disorders of consciousness: Closing the gap towards personalised models for restoring consciousness

Andrea I. Luppi a, h, *, Joana Cabral g, l, Rodrigo Cofre d, e, l, Pedro A.M. Mediano f, g, l, Fernando E. Rosas h, l, j, k, l, Abid Y. Qureshi l, Amy Kuceyeski m, Enzo Tagliazucchi n, o, Federico Raimondo h, q, Gustavo Deco r, s, t, u, James M. Shine v, Morten L. Kringlebach k, w, x, Patricio Ori o y, ShiNung Ching z, Yonatan Sanz Perl r, a a, a b, Michael N. Diringer a c, Robert D. Stevens ad, Jacobo Diego Sitt a a, a e, e**

a Division of Anaesthesia and Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK
b Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
c Life and Health Sciences Research Institute, University of Minho, Portugal
d Centre National de la Recherche Scientifique (CNRS), Institute of Neuroscience (NeuroPSI), Paris-Saclay University, Gif-sur-Yvette, France
e Department of Computing, Imperial College London, London, UK
f Department of Psychology, University of Cambridge, Cambridge, UK
g Department of Informatics, University of Sussex, Brighton, UK
h Centre for Psychedelic Research, Department of Brain Sciences, Imperial College London, London, UK
i Centre for Complexity Science, Imperial College London, London, UK
j Centre for Eudaimonia and Human Flourishing, Linacre College, University of Oxford, Oxford, UK
k University of Kansas Medical Center, Kansas City, MO, USA
l Department of Radiology, Weill Cornell Medicine, New York, USA
m Department of Physics (UBA) e Instituto de Física de Buenos Aires (CONICET), Buenos Aires, Argentina
n Latin American Brain Health Institute (BrainLat), Universidad Adolfo Ibáñez, Santiago, Chile
o Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Germany
p Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
q Center for Brain and Cognition, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
r Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
s Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
t Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia
u Brain and Mind Center, The University of Sydney, Sydney, Australia
v Department of Psychiatry, University of Oxford, Oxford, UK
w Center for Music in the Brain, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
x Centro Interdisciplinario de Neurociencia de Valparaíso and Instituto de Neurociencia, Universidad de Valparaíso, Valparaíso, Chile
y Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA
z Institut du Cerveau et de la Moelle épinière - Paris Brain Institute, ICM, Paris, France
a National Scientific and Technical Research Council (CONICET), Goya de Cruz, CABA 2290, Argentina
b Department of Neurology and Neurosurgery, Washington University in St. Louis, St. Louis, MO, USA
c Departments of Anesthesiology and Critical Care Medicine, Neurology, and Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
d Sorbonne Université, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France

Keywords:
Disorders of consciousness
Computational models
Generative models
Statistical models
Biophysical models
Machine learning

* Corresponding author at: Division of Anaesthesia and Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK.
** Corresponding author at: Institut du Cerveau et de la Moelle épinière - Paris Brain Institute, ICM, Paris, France.
E-mail addresses: al857@cam.ac.uk (A.I. Luppi), jacobo.sitt@icm-institute.org (J.D. Sitt).
1 These authors contributed equally to this work.

Disorders of consciousness are complex conditions characterised by persistent loss of responsiveness due to brain injury. They present diagnostic challenges and limited options for treatment, and highlight the urgent need for a more thorough understanding of how human consciousness arises from coordinated neural activity. The increasing availability of multimodal neuroimaging data has given rise to a wide range of clinically- and scientifically-motivated modelling efforts, seeking to improve data-driven stratification of patients, to identify causal mechanisms for patient pathophysiology and loss of consciousness more broadly, and to develop simulations as a means

Received 15 January 2023; Received in revised form 16 April 2023; Accepted 9 May 2023
Available online 15 May 2023.
1053-8119/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
of testing in silico potential treatment avenues to restore consciousness. As a dedicated Working Group of clinicians and neuroscientists of the international Caring Coma Campaign, here we provide our framework and vision to understand the diverse statistical and generative computational modelling approaches that are being employed in this fast-growing field. We identify the gaps that exist between the current state-of-the-art in statistical and biophysical computational modelling in human neuroscience, and the aspirational goal of a mature field of modelling disorders of consciousness; which might drive improved treatments and outcomes in the clinic. Finally, we make several recommendations for how the field as a whole can work together to address these challenges.

1. Introduction

1.1. Disorders of consciousness

Advances in intensive care medicine have led to an increasing number of patients surviving severe brain injuries. Some of these patients regain consciousness, but others will remain in a Disorder of Consciousness (DOC) state. Post-comatose DOC patients are characterized by having their eyes open (awake) while remaining seemingly unable to respond or communicate (Claassen et al., 2021; Giacino, 2004). Chronic impairments of consciousness constitute a challenging situation, which requires better understanding of the underlying physiopathology to develop novel diagnostic, prognostic, and therapeutic tools to individually and optimally take care of each patient (Edlow et al., 2020). This clinical goal is well aligned (though of course not identical) with the scientific goal of a robust theory of consciousness (Luppi et al., 2021). Improving our ability to detect consciousness and promote its recovery will enhance our ability to study it; and in turn, obtaining a better understanding of consciousness and its neural bases is expected to increase our ability to identify its presence or the causes of its absence, and how to promote its recovery in the clinic.

Patients with disorders of consciousness occupy a behavioural spectrum currently divided into different conditions: coma, vegetative state or unresponsive wakefulness syndrome (VS/UWS) and minimally conscious state (MCS). Patients in a coma are characterised by lack of arousal and responsiveness. On the other hand, both VS/UWS and MCS patients exhibit similarly preserved arousal, but while VS/UWS patients remain largely unresponsive (with comatose patients also having their eyes closed), MCS patients show fluctuations in their behaviour, exhibiting temporal windows where volitional behaviour can be inferred (Bodien et al., 2022; Giacino, 2004, 2002). In the acute stage, prognosis is variable and highly dependent on the severity of the current state. Prognosis for chronic DOC patients is typically poor, with many patients remaining chronically unresponsive, and often dying without regaining consciousness (Estraneo et al., 2020). Treatment options remain limited, and rates of success are modest at best, although vigorous research is taking place to provide alternatives (Schnakers, 2017).

This challenging situation is perhaps not surprising, when one considers that even clinically distinguishing between VS/UWS and MCS patients is far from trivial: on the contrary, this diagnosis can be challenging for physicians without specialist training, and misdiagnosis rates reach up to 43% (Schnakers et al., 2009). Misdiagnosis can have important consequences, such as inadequate pain management, prognosis underestimation and even improper end-of-life decisions. Moreover, even when correctly performed, behaviour-based diagnosis is not fully accurate and might erroneously label some patients who in fact retain covert awareness. Functional brain imaging (fMRI) studies on VS/UWS patients provide evidence for covert intentional brain activity (Monti et al., 2010; Owen et al., 2006) in 10–15% of unresponsive patients and, in a few cases, similar methods have been used to enable simple functional communication with these patients (Claassen et al., 2019; Monti et al., 2010). Therefore, there is growing consensus for a need to go beyond behavioural observation alone, which has been matched by increasing availability of neuroimaging techniques which provide insights about patients’ brains: structural and diffusion MRI to identify the impact of lesions on anatomical structures and their connections, but also functional MRI, electroencephalography (EEG) (Naci et al., 2017) and functional near-infrared spectroscopy (fNIRS) (Abdalmalak et al., 2021) to observe brain activity, and positron emission tomography (PET) measures of metabolism (Bodart et al., 2017; Gołkowski et al., 2017; Hermann et al., 2021; Sala et al., 2021) and receptor density (Qin et al., 2015) to name just a few.

The question therefore arises: can we capitalise on this wealth of data and decades of research to inform diagnosis in the clinic, and predict prognosis and devise more suitable treatments?

1.2. Why modelling?

Neuroscience is witnessing vigorous modelling efforts, from the scale of single neurons to whole-brain models (Amunts et al., 2022; D’Angelo and Jirsa, 2022; Einevoll et al., 2019; Roland et al., 2019). One way to understand what we mean by “model” is “a quantitative specification of a theory about how some aspect of the world works”. That can be as simple as a linear model describing the correlation between two variables; or it can take a more complicated form, such as coupled differential equations describing the evolution of a system. It is clear that obtaining quantitative models of DOC would be valuable not only in terms of advancing our scientific understanding of consciousness and its neural bases, but also in the clinic: for making more accurate diagnoses based on quantitative evidence, and for having a more robust understanding of the pathology’s trajectory in order to try and steer it towards more favourable outcomes.

The first, crucial step towards modelling is therefore the identification and quantification of relevant properties. In the recent decades, with the advent of modern functional neuroimaging techniques, neuroscience has increasingly been able to identify links between loss of consciousness - whether pathological or pharmacological - and altered properties of brain activity and its dynamics (Afrasiabi et al., 2021; Barttfeld et al., 2015; Bonhomme et al., 2019; Campbell et al., 2020; Demertz et al., 2019; Gutierrez-Barragan et al., 2021; Huang et al., 2020; Hutchison et al., 2014; Luppi et al., 2019; Panda et al., 2022; Redinbaugh et al., 2020; Song et al., 2018). Combined with increasingly detailed information about the healthy brain’s macroscale structural and functional, microstructural, and molecular organization (Markello et al., 2022), this information can be leveraged by statistical models, such as machine learning (ML) algorithms, to characterise and categorise patient groups and sub-groups, but also to identify the features that best relate to DOC pathophysiology.

The ultimate end-goal of any DOC intervention is arguably to restore consciousness by re-establishing appropriate brain activity patterns (assuming a direct causal link from brain activity to consciousness). Descriptive statistical models that summarise the data along specific dimensions (whether in the form of Generalised Linear Models, or via machine learning methods), can reveal features of brain activity that are altered in DOCs. However, they are limited in their ability to predict how external manipulations might interact with neural circuits. For this purpose, it is necessary to develop generative models that allow in silico (i.e., through computational modelling rather than in vivo) exploration of hypothetical therapeutics aiming to rebalance brain activity in DOC patients. Generative models serve to describe how a system behaves under certain conditions or in response to perturbation. A major promise of generative models of brain activity is that they can be sys-
tematically and reversibly perturbed, making it possible to probe in silico interventions that are still beyond the capabilities of experimental research — whether in humans or animals (Cabral et al., 2017a; Cofré et al., 2020; Kringlebach and Deco, 2020; Shine et al., 2021). For these reasons, generative computational modelling of brain activity is gaining traction as a tool of choice for investigating the causal mechanisms that drive brain activity in both healthy and pathological conditions, complementing experimental research (Cabral et al., 2017a; Cofré et al., 2020; Kringlebach and Deco, 2020; Shine et al., 2021; Ramezanian-Panahi et al., 2022). However, current generative models of brain activity remain rather abstract, and their predictive power and fidelity to consciousness remain to be validated with experiments in vivo, before they can be translated into clinical applications (Kurtin et al., 2023). Once sufficient knowledge is reached regarding the key features of brain activity that reflect consciousness and how these can be modulated (first in silico then in vivo), novel avenues in addressing DOC can emerge.

Ultimately, the aspirational goal for a mature field of DOC modelling is one of personalised medicine, where models of the healthy brain can be obtained from comprehensive subject-specific multimodal data at the individual level, informed by aggregate data about relevant features from the broader population, and then perturbed to match each patient’s unique patterns of structural organization and neural signatures. Therefore, the disorder-mimicking model can be used as an in silico test-bed (i.e., a “digital twin” of the patient (Erol et al., 2020)) to predict the outcome of alternative interventions and define the optimal therapeutic strategy for each patient aimed at restoring consciousness, operationalised as displacing the modelled activity in the general direction of restored consciousness and cognitive function (Fig. 1). The prospect of personalised computational models enabling this kind of “Phase 0 clinical trials” is especially appealing for DOC, because patients can vary widely in terms of aetiology, lesion site and extent, and symptoms - in turn calling for different treatment avenues with no one-size-fits-all approaches.

However, we survey how different modelling approaches are being employed to address disorders of consciousness; we outline the gaps that exist between the current state-of-the-art, and the aspirational goal of a mature field of DOC modelling capable of finding application in the clinic; and we propose how some of these challenges could be addressed, to bring the field closer to this ambition.

2. Modelling approaches: lay of the land

To clarify how we hope for the field of DOC modelling to develop, it is of course necessary to outline not only what we wish to achieve, but also how close we currently are to achieving it. In turn, this requires an outline of current modelling approaches. In a rapidly developing field at the intersection of disciplines that are themselves rapidly evolving, such a taxonomy is a Sisyphean task, already outdated the moment it is written. However, our aim with presenting this taxonomy is not one of exhaustive enumeration; rather, below we provide a general map, omitting the details of the territory to more clearly convey a sense of the general landscape.

With this caveat out of the way, a first distinction to be made is between statistical models and biophysical computational models (Fig. 2). Statistical models can be subdivided based on whether they seek to characterise observed data in terms of summary statistics (descriptive models), or they simulate the data-generating processes, in order to create new instances (generative models). Descriptive models find widespread use for hypothesis testing in the form of Generalised Linear Models, to quantify correlations or group-level differences in terms of some dimension(s) of interest: “Is there a statistically significant difference between patients and controls in terms of measure X?”. However, descriptive statistical models can also be used in a hypothesis-free approach through machine learning techniques, to identify data-driven clusters or features that best discriminate between clinically relevant categories: “Based on this set of neuroimaging/clinical features, can we find previously undetected subgroups of patients? Along what dimension are patients most discriminable from healthy controls?”. Generative statistical models explicitly learn how the data are distributed, enabling such models to generate new instances that are consistent with relevant statistical properties of the observed data. Of note, here we contrast generative with “descriptive” models, rather than with “discriminative” models; we use this broader term because we deal with a correspondingly broader class of statistical models than just classifiers intended to discriminate between categories.

In contrast, biophysical[ly-inspired] models in computational neuroscience employ equations derived from the underlying biophysics of neural activity (the physical, chemical and/or biological processes governing the dynamics of the system, at some suitable level of simplification and abstraction) to simulate the time evolution of the system, typically in the form of a system of differential equations (Breakspear, 2017; Shine et al., 2021; Ramezanian-Panahi et al., 2022). Note that such models also allow the generation of new data instances: not by learning the underlying probability distributions, but rather via direct simulation of the process that generated the data. By modelling the data-generating process rather than only the resulting data, biophysical computational models provide an avenue to test and evaluate possible causal interventions, addressing counterfactual questions about what would happen if some aspects of the process were disrupted or altered (e.g., “What happens if this connection is lesioned?”, “What happens if inhibition is increased?”). Generative models of brain activity have been shown to approximate the nonlinear response to different types of perturbations,
such as electromagnetic stimulation, structural lesions, or psychoactive compounds (Burt et al., 2021; Deco et al., 2018a; Luppi et al., 2022c). With the caveat that experimental manipulation could be considered as the only true arbiter of causality, biophysical computational models are perhaps the closest in theory.

Below, we describe each of these model families in more detail.

2.1. Statistical models

2.1.1. Descriptive statistical models

Descriptive statistical models comprise a large family, ranging from simple General Linear Models, to ML approaches (e.g. deep neural networks) that can identify discriminative features in the data. General Linear Models are perhaps the simplest form of descriptive statistical modelling, where a statistical relationship is hypothesized and then tested between two (or more) features of interest - for instance, some aspect of brain activity or anatomy versus diagnosis. This approach is widely used to identify predictors/correlates of disease severity or recovery, or to assess differences between patients and controls, or between subgroups of patients. These models are mostly ways of testing specific hypotheses about the data (Demertzis et al., 2015; Luppi et al., 2021a; Lutkenhoff et al., 2020).

More recently, ML efforts are focusing on developing algorithms that learn statistical regularities in the input training data (defined by a set of features) in a more hypothesis-free manner, in order to make predictions on unseen data (e.g., diagnostic or prognostic predictions) (Bareham et al., 2018; Campbell et al., 2020; Chennu et al., 2017; Engemann et al., 2018; Hermann et al., 2021; Riganello et al., 2018; Stefan et al., 2018; Wielek et al., 2018; Zheng et al., 2017). Such algorithms can be as simple as a logistic regression and as complex as deep learning techniques. It would be beyond the scope of this article to describe all possible types of ML approaches. Briefly, the most relevant distinction for current applications is arguably between supervised (including semi- and self-supervised flavours) and unsupervised ML models. Supervised models are given ground-truth information, e.g., clinical labels about the diagnosis or outcome or treatment response of each subject, and their task is to find features in the neuroimaging data that best tell the data-points apart: that is, they perform classification or regression tasks. Unsupervised models instead typically seek to cluster data-points based on their features, to reveal similarities and differences by learning the statistical structure of the input data in the absence of predefined labels (Khosla et al., 2019).

Descriptive statistical models can also be characterised by the various input features considered. First, in terms of estimating the predictive power of single features or the optimal combination of multiple features (i.e., univariate versus multivariate analysis). Second, by the type of features used (e.g., unimodal features - single neuroimaging modality - or multimodal features - multiple modalities). Third, descriptive statistical models can vary in terms of the “feature engineering” strategy adopted: namely, investigators may pre-select features based on theoretical predictions (possibly after complex data-processing steps), or they may choose a more data-driven approach to feature selection. Since descriptive modelling approaches produce classification or regression, they can be used to determine the relevance of a given feature, or identify the most important feature(s) out of many, or classify data-points into controls or patients, or into patient sub-groups.

Note that another kind of “statistical model” can be identified: namely, “null” models that are used to test whether a given observation is statistically unexpected, not against the healthy population or a different patient sub-group (as is typically done when using descriptive statistical models to study DOC), but rather against some hypothesised process. We will not address this kind of model here (nor the related class of generative null models), but we refer the interested reader to an excellent recent review by Váša and Mišić (2022).

2.1.2. Generative statistical models

Generative statistical models aim to reproduce the system under study or some of its (statistical) features, helping to understand what makes the recorded signal behave the way it does. These models are based on function approximation, utilising random processes that have been shown to – or are thought to – describe biological and brain data reasonably, to obtain “new” data that are consistent with statistical properties of the empirical data. Examples of these models are those based on Markov processes, which describe the evolution of a dynamical system by the probabilities of transitioning between separable “sub-states” of activity. From this perspective, each recording of brain activity can be interpreted as a series of transitions between such sub-states. Different approaches have been proposed to decompose brain activity detected from fMRI into a subset of states, including clustering algorithms, Hidden Markov Models (HMMs) or more advanced manifold learning algorithms (Busch et al., 2023). With HMMs, each sub-state is associated with a set of parameters of the observed data, with the most common choice being a multivariate autoregressive (MAR) model (Ou et al., 2015; Vidaurre et al., 2017, 2016). Alternatively, the activity or connectivity patterns observed over time can be clustered into a reduced set of clusters (or sub-states) and the dynamics can be similarly analysed as a Markov process by the transition probabilities between sub-states (Allen et al., 2014; Perti et al., 2017; Vohrýzek et al., 2020). Overall, although the optimal approach to define brain sub-states remains under debate, the approach to characterise brain dynamics from fMRI data as trajectories in a state space (i.e. as a Markov process) has revealed high sensitivity to differentiate across a wide range of conditions.
namely between sleep stages (Stevner et al., 2019), between controls and patients with psychiatric symptoms (Zarghami and Friston, 2020; Alonso Martínez et al., 2020; Farinha et al., 2022), between cognitive traits (Cabral et al., 2017b; Vidaurre et al., 2017; Uddin et al., 2021), and in psychedelic-induced altered states of consciousness (Lord et al., 2019; Olsen et al., 2022).

Note that these models are agnostic to the underlying (bio)physics; in other words, they do not seek to simulate the process that leads to the data, but only the statistical features of the data. Therefore, as with other kinds of generative statistical models, the mechanisms within these models may be biologically implausible, even though their outputs match the statistical features of measured data - an important distinction with biophysical computational models.

2.1.3. Statistical models of DOC: current approaches

The diagnosis of consciousness in patients with DOC poses important challenges, and relying solely on clinical assessments of behaviour has limitations. Recent findings indicate that complementing clinical behavioural assessments with statistical neuroimaging analysis can improve the diagnosis accuracy and the evaluation of intervention outcomes. Various statistical approaches have been proposed to identify and classify patients with DOC using data obtained from different neuroimaging techniques, which could complement systematic behavioural assessment and help reduce the misdiagnosis rate reported in these patients. These methods encompass extracting markers from electrophysiological and neuroimaging data, and using them for multivariate statistical models, aiming to obtain insights on diagnostic and prognostic measures.

Sitt and colleagues used feature engineering to quantify high density EEG (hdEEG) putative neuronal signatures of consciousness (such as interareal connectivity, complexity, spectral activity, as predicted by current theories of consciousness) and quantify their performance to predict the state of consciousness of DOC patients (Sitt et al., 2014; King et al., 2013). They also used ML (support vector machine classifiers) to optimally combine those features and demonstrate that they carry independent predictive information. Specific patterns of resting brain connectivity measured through hdEEG have been found to strongly correlate with the re-emergence of consciousness after brain injury (Bareham et al., 2018). Machine learning analysis of sleep patterns using EEG has also been shown to accurately predict the level of consciousness in patients with DOC (Wielek et al., 2018). Graph theory has been applied to spectral connectivity estimated from EEG, and key quantitative metrics of these networks have been found to correlate with the continuum of behavioural recovery in patients with DOC (Chennu et al., 2017).

A deeper evaluation of EEG-based diagnosis of DOC patients performed by Engemann et al. (2018) who depicted an automated procedure that was suitable for cross-site and cross-protocol diagnosis of DOC. Based on ensembles of decision trees, they concluded that fluctuations in the power of theta and alpha EEG frequency bands were the most consistent and relevant markers. In line with these results, Stefan et al. (2018) showed that the power in alpha frequency band was the most effective at distinguishing patients in minimally conscious state (MCS) from those in unresponsive wakefulness syndrome (UWS/VS), while the average clustering coefficient obtained from beta-band coherence networks was the best predictor of outcome.

Resting-state fMRI (rs-fMRI) has also been used to identify differences in local, regional, and network activity between DOC patients and healthy controls. Machine learning models trained to distinguish between conscious wakefulness and anaesthetic-induced unconsciousness were investigated for their ability to identify pathologically induced unconsciousness (Campbell et al., 2020). The models achieved reliable performance within and across datasets and demonstrated potential for discriminating between degrees of pathological unconsciousness in clinical patients. Analysing rs-fMRI from the perspective of a trajectory in a state space, where the states were defined by clustering instantaneous patterns of phase coherence between brain areas detected over time, revealed that UWS/VS patients show primarily a brain pattern of low interareal phase coherence, with reduced transition probabilities (meaning that this state is more stable) when compared with healthy individuals and minimally conscious patients (Demertzi et al., 2019). In turn, the latter exhibit higher predominance of patterns in which brain regions activate in anti-phase, and switch more often between states.

At the structural level, Annen et al. (2018) used T1-weighted MRI images to extract regional brain volumes of white and grey matter, which were later used in a ML model to diagnose DOC patients. Machine learning based on diffusion MRI tractography was used to identify regions along the tracks that were most informative in distinguishing amongst DOC patients in distinct groups: UWS/VS, and two sub-groups of minimally conscious state, termed MCS+ and MCS- (Zheng et al., 2017). These results indicated that thalamo-cortical connections play a role in patients’ behavioural profile and level of consciousness, and diffusion tensor imaging combined with ML algorithms could potentially facilitate diagnostic distinctions in DOC.

Moreover, multimodal approaches have been proposed to study cross-modal relations with respect to diagnosis and prognosis of DOC. In Hermann et al. (2021), the authors combined FDG-PET and EEG-based classification used based on a support vector machine to optimise diagnostic performance and predict 6-month command-following recovery in DOC patients (Hermann et al., 2021). A more recent work provided a systematic comparison of EEG-extracted features, visual interpretation as well as functional connectivity from rs-fMRI in models to diagnose DOC in the intensive care unit (Amiri et al., 2023).

Finally, statistical models have detected relevant markers for DOC in other physiological signals extending beyond the brain, such as in heart rate. Indeed, it has been shown that electrocardiography can also be used to diagnose DOC (Raimondo et al., 2017), providing partially independent information from the EEG signals. In a later publication, Riganello et al. (2018) showed that heart rate variability (HRV) entropy analysis, specifically the “complexity index”, can serve as a feature for differentiating between VS/UWS and MCS patients. Similarly, Candida Rivera and colleagues demonstrated that features extracted from the heart-evoked potential can be optimally combined to predict the state of consciousness of patients in resting state and during a task (Candida Rivera et al., 2021, 2023). These studies suggest that heart rate monitoring can provide an easy, inexpensive, and non-invasive diagnostic tool for disorders of consciousness, with the aid of statistical modelling techniques.

However, it is also important to clarify that these models have certain limitations. First, several of the previous works do not fully validate the models on new datasets and lack performance estimates beyond the initial cross-validated performances, which are often overestimated (Varoquaux, 2018). Second, their utility as part of clinical decision systems remains unknown. While the models provide novel insights into the DOC population, it is not yet clear what is the cost versus benefit of using such models. Further holistic analysis considering the human, technical and economic cost are required. Finally, and most importantly, statistical models of DOC are well suited to identify relevant features out of many possible candidates, but they do not explicitly model the process by which such features come to be relevant, i.e., they do not, on their own, provide insight on how to act on them, so they provide little insight about clinical interventions.

2.2. Biophysical and biologically-inspired computational models

2.2.1. A vast space of biophysical models

Complementing purely statistical modelling techniques, in silico simulations of brain activity represent a powerful set of tools to study macroscale mechanistic questions in neuroscience (Cabral et al., 2017a; Cofré et al., 2020; Deco and Kringelbach, 2014; Kringelbach and Deco, 2020; Ramezanian-Panahi et al., 2022). Biophysical and biologically-inspired computational models incorporate some aspect of biology (e.g., anatomical connectivity, excitatory and inhibitory pop-
ulations etc.), and exist on a continuum, with varying biological plausibility and varying complexity and detail, which depends in part on the scale of the system being modelled (synapse, cell, region, whole brain network). As the terminology varies, we include in this admittedly broad category models that produce simulations of brain activity over time; such models also typically incorporate empirical data in the simulation process, such as the empirical connectivity between brain regions.

To date, there is no single model that can reproduce all the myriad aspects of human brain activity that span multiple spatial and temporal scales. The pursuit of such a universal model poses overwhelming conceptual and computational challenges. For this reason, researchers typically employ different computational models that are shaped according to their specific research question. Such models can vary widely in terms of their complexity, neurobiological realism of inputs and outputs, and even the target of the modelling; some summarise brain activity via a single global parameter (for example, fitting the model to the point where the mean firing rate becomes unstable, or a marker of criticality), whereas others seek to reproduce aspects of regional activity, or inter-regional connectivity (e.g., the pattern of functional correlations between regions). Likewise, models can vary in the spatial scale of interest - from small neuronal populations to the entire brain - and the temporal scale, from millisecond-resolution electrophysiology to the in-fraslow fluctuations of the BOLD signal (Cabral et al., 2017a; Cofré et al., 2020; Kringlebach and Deco, 2020). There is also a great deal of heterogeneity regarding the level of biophysical plausibility of computational models - abstract (sometimes called “phenomenological”); Ramezanian-Panahi et al., 2022; Pathak et al., 2022; Kurtin et al., 2023) models trade-off biological specificity for a clarity of insight, however at the potential cost of solutions that betray the true processes occurring in the brain.

In turn, both hypothesis-driven and data-driven approaches to modelling can be employed: some researchers seek to find model parameters that best fit the observed data, whereas others seek to assess the effects of perturbing the models in specific ways. An example of the former kind is Dynamic Causal Modelling: this approach employs biophysical equations to model fMRI or EEG activity, but its principal use is not for generating new data, but rather for selecting between competing accounts of a given phenomenon (i.e., “model selection”) (Casey et al., 2022; Friston et al., 2019, 2003; Preller et al., 2019; Stolker et al., 2022). An example of the latter kind is network control theory (Gu et al., 2015), which models activity through an autoregressive process based on the simplifying assumption of macroscopically linear dynamics. As a result of this simplification, network control theory can characterise the propensity of brain networks to steer brain dynamics in a desired direction, or support the spreading of perturbations (e.g., transcranial magnetic stimulation pulses or deep brain stimulation), and identify stimulation regimes capable of producing a desired activity state (Betzel et al., 2016; Cornblath et al., 2018; Gu et al., 2015; Kim et al., 2018; Lynn and Bassett, 2019; Medaglia et al., 2017; Singleton et al., 2022; Tang et al., 2020, 2017; Zarkali et al., 2020) (but see also Pasqualetti et al. (2019), Suweis et al. (2019), Tu et al. (2018)) for a discussion of this approach and its limitations).

2.2.2. Biophysical models in action

This state of affairs generates a vast space of phenomena to be explained, model types, and investigative approaches. However, we believe that each of these aspects should be considered in light of one overarching question: what unique insights does a particular combination provide us?

For instance, models with greater biophysical realism (e.g., dynamic mean field, Jansen-Rit) are especially well suited to investigate the effects of neuromodulatory influences at the macroscale, since they take into account the presence of distinct excitatory and inhibitory populations (Deco et al., 2014, 2013). These approaches have become increasingly prominent thanks to the availability of empirical measurements of the cortical distributions of neurotransmitter receptors and transporters from in vivo PET (Hansen et al., 2022) and post-mortem autoradiography (Goulas et al., 2021; Zilles and Palermo-Gallagher, 2017), as well as the regional expression of associated genes from transcriptomics (Arnatekvićute et al., 2019; Hawrylcz et al., 2012; Markello et al., 2021). Incorporating such biological information has led not only to more realistic models (Deco et al., 2021; Demirtaş et al., 2019; Luppi et al., 2022a; Müller et al., 2020), but also to models capable of simulating pharmacological interventions with a variety of different drugs, covering the range from psychopharmacology to anaesthetics (Burt et al., 2021; Coronel-Oliveros et al., 2023, 2021; Deco et al., 2018a; Kringlebach et al., 2020; Luppi et al., 2022c). However, we note that a high degree of biological realism is not mandatory for a modelling approach to be able to capture the effects of pharmacological interventions, as recently demonstrated e.g. with extensions of network control theory that incorporate receptor expression to simulate the effects of psychedelics (Singleton et al., 2022), and previous work simulating the effects of anaesthesia using generalised Ising models from statistical mechanics (Kandeepan et al., 2020; Straamgla et al., 2017).

On the other hand, Hopf/Stuart-Landau and Kuramoto models are suitable for modelling the oscillatory character of brain activity, and study aspects such as synchrony and metastability of neuronal oscillations (Cabral et al., 2017a; Deco et al., 2017; Deco and Kringlebach, 2016; Viša et al., 2015). Some approaches (e.g. Jansen-Rit) allow for easier translation across different simulations of functional brain activity (fMRI and EEG) (Coronel-Oliveros et al., 2021) and others can provide more latitude for perturbation. For example, regional oscillations in a Hopf model can be made subcritical (i.e., their amplitude naturally decay over time) or supercritical (i.e., the oscillations are sustained with constant amplitude) (Deco et al., 2019; Hahn et al., 2020, 2021; Ipíña et al., 2020; Jobst et al., 2017; López-González et al., 2021). Overall, although it may be tempting to categorise modelling efforts in terms of their choice of model (model family), we believe that the emphasis should be on what the model does and does not reproduce about brain function, and what opportunities and insights it offers.

2.2.3. Biophysical computational models of DOC: current approaches

Over the last decade, whole-brain models of network dynamics have shown promising potential to investigate the causes of altered brain activity detected across DOC. Considering the dynamics of brain areas interacting in the neuroanatomical network (with different degrees of biophysical realism), the simulated activity is found to reveal features qualitatively similar to experimentally-recorded brain signals.

In addition to models explicitly aimed at recapitulating DOC (Abeyasinghe et al., 2020; Luppi et al., 2022c; Sanz Perl et al., 2021), biophysical models have also been employed to study other relevant and related conditions, such as the effects of brain lesions in conscious patients (e.g., stroke, brain injury; Rocha et al., 2022; Favaretto et al., 2022), or the effects of loss of consciousness in the healthy brain (e.g., sleep, anaesthesia) (Deco et al., 2017; Stramaglia et al., 2017; Kandeepan et al., 2020; Ipíña et al., 2020; Hahn et al., 2020). Indeed, recent efforts have also been undertaken to explain the changes in brain function observed in DOC patients - typically capitalising on the comparison with anaesthetic-induced unconsciousness to distinguish lesion-specific and consciousness-specific effects. This has leveraged recent empirical work that demonstrated important similarities between the brain dynamics of DOC patients and those of anaesthetised individuals (Barttfeld et al., 2015; Bonhomme et al., 2019; Campbell et al., 2020; Cao et al., 2019; Demertzi et al., 2019; Golkowski et al., 2021, 2017; Gutierrez-Barragan et al., 2021; Huang et al., 2020; Luppi et al., 2019; Panda et al., 2022; Song et al., 2018) with subsequent extenssions identifying common underlying neuromodulatory mechanisms (Spindler et al., 2021).

Such models have to date capitalised on the similarities and differences between DOCs and anaesthesia, incorporating empirical evidence about patients’ disrupted patterns of structural connectivity between brain regions (Abeyasinghe et al., 2020; Luppi et al., 2022a, 2023;
Sanz Perl et al., 2021). For instance, Luppi and colleagues (Luppi et al., 2022c) showed that the dynamics of a dynamic mean-field model can be altered in comparable ways by a pharmacological perturbation (increase of model regional inhibition in proportion to the empirical distribution of GABA-A receptors, to simulate the effects of the GABA-ergic agent propofol) or by perturbing the structural connections to be analogous to the connectome of DOC patients. This work sought to explain how different changes to the brain’s normal functioning (transient pharmacological intervention versus chronic structural lesion) can lead to similar patterns of brain dynamics. Conversely, Sanz Perl and colleagues (Sanz Perl et al., 2021) demonstrated that the states of anaesthesia and DOC can be distinguished in terms of how responsive they are to external perturbations, with DOCs being more resistant to change - consistent with their persistent nature versus the transient nature of anaesthesia.

The focus on responsiveness to perturbations is no coincidence: one of the best-performing empirical methods to estimate an individual’s residual consciousness, the Perturbational Complexity Index, relies on evaluating the EEG response to brief perturbations induced by pulses of transcranial magnetic stimulation (Lee et al., 2022; Casali et al., 2013; Casarotto et al., 2016; Rosanova et al., 2018; Sarasso et al., 2021). Therefore, several efforts have been underway that seek to reproduce this phenomenon in silico (Bensaid et al., 2019), including incorporation into the popular modelling framework of The Virtual Brain (Goldman et al., 2021). For instance, a recent study (Luppi et al., 2023) used a dynamic mean-field model to demonstrate that the structural network alterations of DOC patients are sufficient to induce less hierarchical propagation of spontaneous events (“intrinsic ignition” (Deco et al., 2017; Deco and Kringelbach, 2017)), a phenomenon that is also observed empirically in patients’ brains, and that correlates with compromised measures of network controllability. Although these models offer some mechanistic insight to explain the features that differentiate between conditions, they still lack precise predictive value and more efforts are needed to fully realise the potential of computational models of brain stimulation, in particular for the design of personalised stimulation strategies with increased effectiveness (Kurtin et al., 2023; Vohryzek et al., 2022).

Other modelling efforts have also sought to capture aspects of the recovery process, by focusing on how brain activity spreads on the connectome and how this changes as a result of perturbations that alter the network’s topology (Vasa et al., 2015; Cabral et al., 2012; Deco et al., 2018b). For instance, it has been shown that recovery from DOCs induced by severe injury may depend on re-routing of functional connections whose structural connections are impaired, without necessarily changing the SC itself (Kuceyksi et al., 2016).

2.3. Mixed modelling strategies

2.3.1. Combining models to overcome their specific limitations

At this point, it is worth acknowledging that in describing the different model types, we have inevitably highlighted their differences - but similarities of course abound. All models are in some sense data-driven: whether because they look for patterns in the data, or because they use data to determine model fit and model parameters. Moreover, the choice of model type and features (if applicable) determines what the model can provide, such that none of these models is completely data-driven.

More broadly, a modelling workflow may involve multiple model types, by combining descriptive and generative statistical modelling (e.g., by performing statistical comparisons based on features identified by ML), or by combining statistical and biophysical modelling. For instance, unsupervised k-means clustering (a type of descriptive statistical modelling that clusters data based on specific features) may be used to identify distinct brain-states that can then be tested for differences in terms of relevant features (e.g., state occupancy) through statistical General Linear Models (Bartfeld et al., 2015; Lord et al., 2019) or used to identify target features to fit a biophysical computational model (Deco et al., 2019; Kringelbach et al., 2020). In addition to ML, biophysical computational models can also be used to extract additional features for discriminative models, such as identifying the global coupling parameter G that enables a biophysical model to best fit each subject’s empirical data, and then comparing groups based on this (Coronel-Oliveros et al., 2023; Deco et al., 2017).

As we described above, the application of ML algorithms to neuroimaging data shows great promise for classifying physiological and pathological brain states. However, classifiers trained on high-dimensional data are prone to overfitting, especially for a low number of training samples. To overcome this roadblock, over the last years strategies were developed that combine whole-brain computational models with statistical models (Vohryzek et al., 2022). The main rationale behind these strategies is to take advantage of the generative capabilities of the whole-brain model to meet the requirements that the statistical models have, in terms of the amount of data required to achieve sufficient statistical power and make generalisable predictions. In this vein, Arbábyazd and colleagues developed whole-brain models to create surrogate data to train random forest and Boost algorithm ML models to classify Alzheimer’s disease patients and healthy participants (Arbábyazd et al., 2021). The authors demonstrated that the performance of both classifiers is comparable with that obtained when the models are trained with empirical data. Another strategy that combines both whole-brain and statistical models is postulated by Gilson and colleagues (Gilson et al., 2019). The whole-brain model fitting procedure generates model parameters, and the resulting effective connectivity between brain regions (quantification of the influence of one region’s activity over the activity of another) can be used to train statistical discriminative models to classify different brain states. This shows that the modelling types can be combined into a “virtuous cycle”, progressively refining the features and identifying which ones could be intervened upon. Generative models can also be used for data augmentation (i.e., a mathematical process to generate synthetic data with the purpose of augmenting the training dataset, thus enhancing the model’s learning capacity) to help improve the discriminative models, so the synergy between the two approaches can go both ways.

2.3.2. Mixed models of DOC: current approaches

Combining machine learning and descriptive statistical modelling, Demertz et al. (2019) and Luppi et al. (2019) both used k-means clustering to identify multiple dynamic states of brain connectivity from functional MRI of DOC patients and healthy controls. Demertz and colleagues then extracted the prevalence of each state, and through descriptive statistical modelling demonstrated that UWS patients spend a greater proportion of time in a pattern characterised by high coupling with the underlying structural connectivity, with smaller chances to transition between patterns. This study also shows the value of combining not only multiple modelling strategies, but also multiple imaging modalities. Luppi and colleagues instead focused on network properties of the different dynamic functional connectivity states identified by k-means clustering, demonstrating through descriptive statistical modelling that a brain state of high network integration is especially affected by loss of consciousness, both in DOC patients and anaesthetised individuals.

In recent work, Sanz Perl et al. (2020) proposed implementing whole-brain models as a dynamical model informed data augmentation procedure to create meaningful surrogate data keeping the spatiotemporal structure of the original data. In that work, a random forest classifier was trained to discriminate between sleep stages and wakefulness with surrogate data generated with whole-brain models fitted to individual and group average empirical data for different stages of the wake-sleep human cycle. In both cases, the classifiers showed good performance when evaluated against empirical data, demonstrating that statistical models can be trained with whole-brain model individual and group average synthetic data.

Dynamical model data augmentation procedure was also used to train unsupervised statistical models. For instance, in Perl et al. (2020) the authors trained a variational autoencoder (VAE) with syn-
thetic functional connectivity corresponding to wakefulness and deep sleep to obtain a low-dimensional representation of brain states. This strategy allowed the authors to find an orderly trajectory from wakefulness to brain injured patients in a latent space whose coordinates represent metrics related to functional modularity and structure-function coupling, both increasing alongside loss of consciousness. These results suggest that other brain states (e.g., DOC) could be captured and understood in terms of trajectories within a low-dimensional latent space, with potential applications in diagnosis and prognosis.

3. Interim summary

Overall, while both generative statistical and biophysical computational models aim to simulate the outcome, only the biologically-inspired models aim to simulate the entire data-generating process. This causal flavour is derived from their biological inspiration - with the understanding that insight into causality should be rooted on models that are biologically realistic. This means that biophysical computational models are also well suited to providing insight about counterfactuals, by directly simulating the effects of an intervention, (Fig. 3) and therefore about the likely effects of a given treatment option. This stands in contrast with statistical models, which are most suited to diagnosis and prognosis. Of course, since generative statistical models learn the joint probability distribution of multiple features, it is possible to use such models to explore how variation in one feature would affect another. For instance, to test the effects of a certain drug, one could first assess how this drug modifies the features of interest in the data, and then explore the region of modelled data-space that corresponds to these changes. However, such out-of-sample generalisations are notoriously difficult. In contrast, having a causal model of how one variable influences the whole system’s behaviour is precisely the raison d’etre of biophysical computational models, as such relationships can be explicitly encoded.

In other words, descriptive models are well suited to identify relevant features out of many possible candidates, but because they do not explicitly model the process by which such features come to be relevant, they do not on their own provide insight on how to act on them: they do not tell the clinician how to intervene. Generative statistical models and biophysical models can help to test possible interventions for the promising features identified by discriminative models: the foundations towards “digital twins”. The downside is that finding the appropriate level of biological complexity is not straightforward: a mismatch with reality will always be present, to some extent, and it can be challenging to determine whether it is innocuous abstraction, inevitable noise, or misleading mis-specification. However, as reviewed in the section on Mixed modelling strategies, research efforts have been ongoing to combine the strengths of both modelling approaches, and to use one to mitigate the shortcomings of the other, representing the source of numerous recent insights.

4. The road towards a mature field of DOC modelling

Having provided an overview of the different kinds of modelling approaches that are being employed to investigate disorders of consciousness, in this second part of our article we provide recommendations for how the field can move forward: the challenges that lie ahead, and how we believe that they can be overcome.

4.1. Open challenges in simulating DOC

Despite these favourable characteristics, applications of whole-brain modelling to DOCs are relatively recent, arguably because this endeavour involves substantial challenges. In effect, the complexity and heterogeneity of DOCs in terms of aetiology — with each patient typically

Fig. 3. Overview of using biophysical computational models to evaluate causal interventions. Empirical structural connectivity between regions can be reconstructed from diffusion MRI tractography, which provides the global connectivity between regions. Perturbations can be applied to this connectivity, to simulate lesions or plasticity. Additionally, perturbations to the local dynamics of each region can simulate neuromodulatory influences and pharmacology. The resulting simulated activity and functional connectivity can then be compared against empirical data, to evaluate the model's goodness-of-fit.
presenting unique patterns of cerebral lesions and deficits — greatly complicates extrapolating modelling paradigms based on healthy brains.

One way to approach the challenge of modelling brain activity of patients with DOC is to decompose the problem into two more tractable questions: (i) modelling unconsciousness, and (ii) modelling brain damage. As testsbeds to model unconsciousness, at least three candidates present themselves: the endogenous, transient state of unconsciousness comprising dreamless (deep) sleep; the endogenous but pathological unconsciousness of epileptic seizures; and exogenously induced unconsciousness resulting from the administration of general anaesthetics - which unlike sleep, is a perturbation of the brain’s spontaneous state, in this sense resembling DOC more than sleep does. These conditions can also be studied in animal models, where there is enormous capacity for experimental access and manipulation, leading to substantial advances in our understanding of their neurobiology. In turn, this neurobiological knowledge makes it possible to formulate specific mechanistic hypotheses based on experimental results, and develop suitable computational models spanning from very biologically detailed ones (Ching et al., 2010, 2012) to more abstract ones that nevertheless capture well-known features of the anaesthetised or sleeping brain (Deco et al., 2017; Kandeepan et al., 2020; Stramaglia et al., 2017).

On the other hand, anaesthesia and sleep are typically studied in the healthy brain in humans, whereas DOCs typically involve lesions of varied extent and location, such that both grey and white matter can be affected. However, not all brain lesions result in chronic disorders of consciousness, and neuroscientists have developed computational models to understand the effects of brain damage on patients’ cognition and brain function. For example, previous studies have shown that computational modelling can provide specific predictions of cognitive impairment linking structural lesions and their effect on neurodynamics (Cabral et al., 2017a; Vasa et al., 2015; Rocha et al., 2022; Favaretto et al., 2022) — in particular highlighting the role of the integrity of “hub” nodes for the assessment of cognitive loss. For both models of unconsciousness and models of brain injury, animal models have played a fundamental role in shaping our understanding, thanks to their greater experimental accessibility (Beppi et al., 2023; Redinbaugh et al., 2020; Bastos et al., 2021; Tasserie et al., 2022; Bartfefd et al., 2015; Gutierrez-Barragán et al., 2021). Studies in anaesthetised macaques provided the first evidence that brain structural and functional connectivity become more similar when consciousness is lost (Bartfeld et al., 2015) - which was later confirmed both in mice (Gutierrez-Barragán et al., 2021) and humans (Demertz et al., 2019). More recently, several studies demonstrated that electrical stimulation of the central thalamus can restore neural and behavioural signatures of consciousness in anaesthetised macaques (Redinbaugh et al., 2020; Bastos et al., 2021; Tasserie et al., 2022), a feat which is possible thanks to animals’ greater experimental accessibility. Developing computational models of animal models (as done e.g. by Hahn et al. 2020) is therefore going to be a key stepping stone towards the human counterpart.

We note that in order to characterise the current gaps that need to be bridged, it is not sufficient to identify the current state of the art: one must also clearly define the goals to be achieved. Upon doing so, it becomes apparent that for the present endeavour, two different kinds of goals exist, which are closely aligned but nevertheless not identical. On one hand, there is the clinicians’ goal to improve diagnosis and prognosis and ultimately treat (or at least alleviate the suffering of) patients with DOC. On the other hand, there is the scientific goal of understanding how the brain can become stuck in a state of chronic unconsciousness, and how this illuminates the mechanisms that in the healthy brain enable consciousness to emerge from matter. The history of medicine is replete with examples of treatments that were discovered serendipitously or through trial and error, and hence employed before being fully understood. Arguably, the mechanisms of anaesthesia, one of the most important tools in medical history, remain incompletely understood - which fortunately does not prevent anaesthetists from using it to spare patients the intolerable suffering of surgery. Nevertheless, a greater scientific understanding can only benefit the clinical approach, both in terms of diagnosis and treatment. However, a scientist may be satisfied with more abstract understanding at the group level, whereas clinicians inevitably do not treat groups, but individuals - and they have a stronger incentive to provide timely answers, since they cannot indefinitely put off decisions about treatment avenues. Therefore, in this aspect the scientific and clinical goals diverge, and this divergence needs to be considered when outlining our desiderata as a community and how we intend to attain them.

In particular, the clinical and scientific perspectives may hold different views on the main trade-off involved in biophysical/generative modelling of DOC: biological realism versus complexity. If we want to have models that help us to predict outcomes or evaluate in silico the potential for different treatments, then a mis-specified model that fails to take into account relevant aspects of neurobiology (the definition of which is itself part of the problem, of course) may constitute a costly mistake. Likewise for the case where statistical/discriminative models are being used for diagnostic/prognostic purposes, or to identify patients for more in-depth investigation. This issue is intertwined with the modeller’s perennial question: what counts as a well-fitting model? There are two parts to this question. On one hand, the model must be able to match the desired aspect of the data. But perhaps even more importantly, such an aspect (the loss function in terms of which the model is evaluated) must be properly identified. A model that faithfully reproduces the wrong function is at best useless, and at worst actively misleading. Finding a suitable objective function - in this case, a neuroimaging marker that a given brain is capable of entertaining consciousness - is a task for empirical investigations of consciousness, to be aided by discriminative statistical models.

With this in mind, we lay out four main aspirations for the modelling of DOC, for each desideratum, we also highlight what we see as the most pressing challenges, and we outline some potential approaches that we believe could contribute to address such challenges.

4.2. Desiderata for the future of DOC modelling

4.2.1. Desideratum 1: greater generalisability

Models need to be able to generalise across individuals and across different aetiologies (e.g., traumatic versus anoxic/hypoxic injury). More broadly, models should generalise across different cohorts. Within the same individuals, models ought to be able to provide a match for a particular individual’s multimodal data, reflecting their capacity to truly provide insight about the patients’ brains. Generalising to the broader population is the implicit goal of the statistical tests we employ, and ability to do so is a marker that we possess true scientific insight into a given phenomenon.

Building models that capture multimodal data can help to address the first challenge that we face when trying to improve generalisability: the need to identify relevant spatial and temporal scales that our models should aim to capture. For instance, fMRI and EEG differ in terms of both spatial and temporal resolution, so a model that can simulate both would imply the ability to capture both slow and fast dynamics. Additionally, models that can provide a good fit to multiple modalities (e.g., both fMRI and EEG data) provide intrinsic validation for their biological plausibility. This avenue of addressing the first challenge to generalisation can capitalise on the existence of rich datasets about patients, spanning not only neuroimaging but also neurotransmitter expression, transcriptomics and proteomics, which are now becoming increasingly available. Of course, the success of this endeavour depends on the data being compatible and of high enough quality, to make sure that the additional data are not merely adding noise. In particular, multi-modal models may best enhance generalisability when cross-modal data at different scales are effectively integrated.

A second challenge to generalisation is that DOC patients vary widely not only in terms of diagnosis, but also in terms of lesion extent and location, as well as severity and aetiology. This means that the boundaries
for generalisation remain ill-defined: which subset of patients constitute an intended target for generalisation (such that failing to generalise to such patients indicates a shortcoming of the model), and which patients are instead simply beyond the scope of the model? To address this challenge, a useful starting point may be to identify homogeneous sub-groups of DOC patients with similar diagnosis and similar aetiology, and use this sub-group-level as a stepping stone between the individual-level and the broader group-level. In particular, it may be especially fruitful to focus on patients displaying similar constellations of symptoms, and building models that can be differentially perturbed to obtain individual-specific symptomatology.

An intriguing potential avenue for generalisation is the recent proposal that late stages of dementia may be behaviourally equivalent to DOC (Huntley et al., 2021). If so, this would provide an invaluable additional source of information, since the progression is gradual and observable, unlike most of the injuries that lead to current DOC patients’ admission. However, this avenue is itself not without challenges. Gradual atrophy of a region due to neurodegeneration need not produce the same cognitive and behavioural deficits caused by sudden ischaemic stroke or traumatic lesion of the same region. Therefore, the rate of change, and resulting role of plasticity, may point to distinct processes. Nevertheless, the greater similarity between gradual progression stages of brain degeneration in dementia may provide a suitable testing ground for models to generalise.

4.2.2. Desideratum 2: patient-specificity

As anticipated above, the clinical need is to treat individual patients, and one of the greatest promises of modelling is the possibility to develop personalised models for each patient, in the vein of the “Digital Twins” paradigm (Fig. 4). This can broadly take two forms. Under the first approach, a “digital twin” of patient X is a model obtained from aggregating large amounts of data from other patients who resemble patient X in some relevant aspect, with the aim of triangulating on this specific patient. Under the second approach, a “digital twin” is built from the patient’s own multimodal data. Of course, the two approaches are not mutually exclusive: group data can be used to “fill in” missing information about a patient in question.

This is especially valuable given the wide variability of DOC patients and their trajectories. Unfortunately, the risk is that models may end up fitting idiosyncrasies rather than capturing what is still common across DOC patients: their impaired consciousness. In other words, single-patient modelling risks overfitting to features unrelated to DOC (especially since each patient often presents a unique and complex aetiology, with a potential host of deficits and impairments that are not directly related to their unconscious status). Under such conditions, clinicians with an interest in simulating treatment options find themselves between the Scylla of using a group-level model that may not apply to their specific patient and his/her needs; and the Charybdis of using a personalised model, but with little confidence that it reflects generalizable information. Therefore, the main challenge to patient-specific models is to establish trust in single-patient predictions, mitigating the impact of overfitting and measurement noise.

Personalised models could be created, for example, using a transfer learning-like approach wherein the model begins by representing a population and is fine-tuned using an individual’s observed data (Gu et al., 2022). These personalised models would then have to be tested in how well they reflect empirical data. This could be accomplished by successfully predicting a patient’s trajectory, matching longitudinal data. A second, more demanding bar to clear would be to show that a model can successfully predict the effects of a given intervention on a patient. This could be done for patients who are already due to receive a given treatment (e.g., a specific medication). If replicable across patients, this would provide the kind of evidence that the model may also be suitable to inform the choice of treatment.

4.2.3. Desideratum 3: models of recovery

The current state of DOC modelling focuses primarily on distinguishing between groups, with the goal being either diagnosis or obtaining a mechanistic understanding of loss of consciousness (or a combination of both). However, the prime clinical goal is to achieve recovery, and therefore, there is a need for models that can reflect the recovery pro-
cess: both its spontaneous course (which would also aid prognosis), and how it can be promoted or accelerated through interventions. The main challenge to this goal is that few patients fully recover, and many instead decline; and for those who do recover, this process is typically very protracted and gradual, rather than an abrupt change (Eddow et al., 2020). If we do not understand how the brain recovers, we remain limited in our ability to help it to recover.

Conceptually, it cannot be expected that the process of recovery from DOC will mirror the process of loss of consciousness, as one might expect to happen for sleep or anaesthesia, because it is not possible to simply “un-do” the damage that a patient has suffered, in the way that anaesthesia wanes as the drug concentration is diminished. And in fact, even for sleep and anaesthesia there is growing evidence that loss and recovery of consciousness are actually asymmetric, with an “inertia” or hysteresis effect being observed for both, suggesting more complex transitions (Friedman et al., 2010; Kuizenga et al., 2018; Luppi et al., 2021; Proekt and Hudson, 2018). Additionally, even if the process were symmetric, we do not have imaging data about the moment of the injury, in the same way that we can image anaesthetic-induced loss of responsiveness.

Nevertheless, some ways to mitigate these limitations can be envisioned. First, emergence from sleep and anaesthesia, though imperfect models, can still provide valuable insights about the general process of recovering consciousness. Second, in some cases it may be possible to scan patients before and after treatment: especially if such treatments were to show a degree of success (such as the effects that zolpidem on a small subset of DOC patients (Noormandi et al., 2017)), then the pre- vs post-treatment comparison could provide insights about emergence to inform modelling efforts. Third, longitudinal imaging can provide data about both gradual recovery and gradual decline, which can then be used as benchmarks for models intended to capture disease progression rather than just static snapshots.

Imaging acute patients can also provide clearer data closer to the time of loss of consciousness, while also providing the occasion to make model-based predictions that can then be evaluated against the actual progression of the patient (although the acute phase during intensive care is inevitably complicated by practical difficulties and confounds such as sedation). In this context, studies of recovery from stroke or TBI without loss of consciousness could provide an avenue to decouple brain injury and repair from their effects on consciousness, for instance in terms of spatial distribution of damage (Maas et al., 2022, 2014; Olafson et al., 2022, 2021), which is undoubtedly one of the key challenges. This approach can then be complemented by studies of loss and recovery of consciousness in animal models, where causal intervention is more feasible and there can be greater experimental access; recent work has been identifying promising stimulation targets to awaken non-human primates from anaesthesia, and modelling this scenario offers a clear avenue to make progress on the field’s ability to model recovery of consciousness (Bastos et al., 2021; Donoghue et al., 2019; Tasserie et al., 2022). Together, these approaches may offer a way to devise models of recovery from DOC, by modelling different aspects of the recovery process, across multiple timescales. To this end, incorporating aspects of temporal evolution over the long term (e.g., plasticity of connections (Hellyer et al., 2016)) will be a key direction for the field.

4.2.4. Desideratum 4: improved confidence

This last desideratum is not unique to modelling DOCs, but rather it is a broader one that is shared with modern applications of modelling and machine learning, especially in science and healthcare. Namely, models need explainability, accountability, reliability, in order to be used with confidence. Can the model provide an understandable explanation of the decision? Does it explain how and what it learnt from the data? Some work has already been carried out in this direction, by adopting interpretable deep learning approaches to the classification of sleep, anaesthesia, and pathological unconsciousness (DOC) based on EEG features (Lee et al., 2022). These authors showed that the model can disambiguate between awareness and arousal, correlating with intervention-based approaches.

This desideratum can be summarised by saying that “models must not be black-boxes” (Castelvecchi, 2016; Heinrichs and Eickhoff, 2020). To address these goals, some considerations need to be first taken into account. Firstly, model fitting: many of the smaller-scale biophysical computational models are ill-parameterized and may result in overfitting; and the same may also apply to statistical approaches: an issue that is also related to our desideratum about generalisability. Multiple avenues exist to address this challenge. On one hand, repeated studies on the same patients could help to boost confidence and increase the signal-to-noise ratio within individual patients. This could be combined with increased data-sharing to pool samples across different cohorts, which would not only contribute to generalisability, but also provide better ways to avoid overfitting via cross-validation. This approach would also increase the representation of each sub-group of patients - as well as potentially helping to identify relevant sub-groups for stratification. With data-sharing comes the need to harmonise data acquisition paradigms from different sites (Orlach et al., 2022; Pomponio et al., 2020) as well as using homogeneous preprocessing pipelines and consistent fitting procedures across studies, to make research more comparable and facilitate identifying improvements over the state-of-the-art - which is typically not possible in current research, when studies differ in terms of model used but also cohort, data processing, and fitting criterion.

5. General recommendations

In addition to the Desiderata outlined above, there are also a number of more general recommendations that we believe would help both the study of DOCs as a whole, and the more specific endeavours of modelling DOCs.

First, adding to the call for data-sharing and harmonisation, we believe that DOC research should take inspiration from the non-human primate neuroscience. Like most DOC datasets, non-human primate neuroimaging datasets tend to comprise a relatively small number of individuals, due to the difficulties of data collection. In non-human primate neuroscience, one way to overcome this limitation has been by repeated data acquisition from the same individual, to boost the signal-to-noise ratio. To the extent that this is feasible for DOC patients, we believe that availability of multiple scanning sessions, both with the same and with different neuroimaging modalities, would be of great help to the modelling effort. Additionally, the field of non-human primate neuroscience has recently come together to organise large-scale data-sharing initiatives to accelerate research, and we hope for a similar initiative for DOC (Michael Milham et al., 2018).

A second avenue for progress in DOC modelling research will be to combine not only different flavours of models, but also combine and iterate between theory- and data-driven approaches (Luppi et al., 2021). We envision that the development of in silico models that reproduce the ML-derived insights will be an essential component of a mature field of DOC modelling.

Third, it will be important to obtain a more comprehensive (and ideally, more formal) understanding of how the space of possible models matches onto the space of modelling goals, in the context of DOC. In other words: when does each model stop being applicable? As modelling and model types evolve, so will this landscape change.

Finally, with new technologies inevitably come new considerations pertaining to their ethical use. In particular, we have advocated for the prospect of using models to assess the expected outcome of a given treatment option, for a given patient, in the paradigm of “digital twins” (Vohryzek et al., 2022). The question arises, however: if modelling suggests that a treatment is likely to be ineffective, to what extent should this be sufficient grounds for not attempting the treatment? We expect that there is no one-size-fits-all answer to this question: rather, as models become more personalised and their predictive power increases, they may become a greater component of the broader cost-benefit evalu-
tion pertaining to each patient. We do not expect that such models will replace clinicians’ assessment, nor do we think that this would be desirable: rather, they will provide clinicians with additional information when forming their expert judgments.

6. Concluding remarks

Overall, modelling approaches are perhaps one of the most promising avenues of progress in our understanding of disorders of consciousness, owing to the combination of increasing computational resources and increasingly detailed and powerful models. Although a number of gaps exist in the current state-of-the-art, we have outlined how the field could overcome these gaps to realise the full scientific and clinical power of computational modelling. The possibility of building “digital twins” and using them for “Phase 0 clinical trials” may substantially advance our ability to identify suitable treatments for individual patients, capitalising on the increasing amounts of multimodal neuroimaging data available.

In this context, we emphasise that there is a distinction to be drawn between modelling the difference between a conscious and an unconscious brain, and developing a full theory of consciousness (in terms of what consciousness “is”). We acknowledge that understanding consciousness is a challenging question to address, and a full understanding of human consciousness will require a concerted, multi-disciplinary approach well beyond the one outlined here: our goal is focused towards clinical relevance. Nevertheless, throughout the history of medicine, treatment has often preceded the detailed understanding of the ailment. We believe that a mature state of computational modelling of DOC will provide invaluable tools for personalised diagnosis, prognosis, and treatment, but we also believe that in the process of achieving this status, computational models will continue to provide important insights about consciousness itself.

We reiterate that our modelling taxonomy is not meant to be exhaustive or definitive. Indeed, our group of expert scientists and clinicians disagreed at times about the most suitable way to characterise the different model types, and which ones should be grouped together, and under what title. Are biophysical and statistical models “generative” in the same sense? Would it be more helpful to describe the main distinction between models as “top-down” (testing a given feature of interest) versus “bottom-up” (putting together the pieces to produce a representation of interest)? In some sense, the word “model” itself may simply hold different meanings for different researchers. We hope that this work will contribute to establishing a common ground and a common reference for discussing modelling approaches in the context of disorders of consciousness, so that clinicians and computationalists will achieve greater integration of different approaches, to overcome the challenges that we have outlined here, and achieve greater progress towards our shared goal: curing coma.

Ethics statement

This work did not involve collection or analysis of data.

Declaration of Competing Interest

The authors have no conflicts of interest to declare.

Credit authorship contribution statement

Andrea I. Luppi: Conceptualization, Writing – original draft, Writing – review & editing. Joana Cabral: Writing – original draft, Writing – review & editing, Visualization. Rodrigo Cofre: Writing – original draft, Writing – review & editing, Visualization. Pedro A.M. Mediano: Writing – review & editing. Fernando E. Rosas: Writing – review & editing. Abid Y. Qureshi: Writing – review & editing. Amy Kuceyeski: Writing – review & editing. Enzo Tagliazucchi: Writing – review & editing. Federico Raimondo: Writing – review & editing. Gustavo Deco: Writing – review & editing. James M. Shine: Writing – review & editing. Morten L. Kringelbach: Writing – review & editing. Patricio Orio: Writing – review & editing. ShiNung Ching: Writing – review & editing. Yonatan Sanz Perl: Writing – review & editing. Michael N. Diringer: Conceptualization, Writing – review & editing, Supervision. Robert D. Stevens: Conceptualization, Writing – review & editing, Supervision. Jacobo Diego Sitt: Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

Data availability

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

Acknowledgements

This Working Group was brought together under the auspices of the international Curing Coma Campaign.

Ail was supported by the Gates Cambridge Trust (OPP 1144). RC acknowledge the support the Human Brain Project, H2020-945539.

AK was supported by US National Institutes of Health grants R01NS102646 and RF1MH123232.

JC was supported by La Caixa Foundation (LCF/BQ/PR22/11920014).

JMS was supported by the Australian National Health and Medical Research Council (1193857).

PO is supported by grants Fondsdcyt 1211750 and Centro Basal AC3E FB0008 (ANID, Chile).

YSP is supported by European Union’s Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant 896354.

MLK is supported by the Centre for Eudaimonia and Human Flourishing (funded by the Pettit and Carlsberg Foundations) and Center for Music in the Brain (funded by the Danish National Research Foundation, DNRF117).

SC acknowledges support from the US National Institutes of Health (R01NS130693).

ET acknowledges support from FONDECyT (1220995).

JDS and GD acknowledge support from FONDECyT project ModelDX-Conssciousness

Part of this work was funded by a grant from ECOS Sud [project A2OM02] [JDS, ET] and STIC-AmSud [project SILIDOC - 21-STIC-11] [JDS, ET, PO].

References


