Refining the Allostatic Self-Efficacy Theory of Fatigue and Depression Using Causal Inference

Alexander J. Hess [1,*] **Dina von Werder** [1,2,3] **Olivia K. Harrison** [1,4] **Jakob Heinzle** [1] **Klaas Enno Stephan** [1,5]
[1] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich
[2] Institute of Medical Technology, Brandenburgische Technische Universität Cottbus-Senftenberg
[2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians Universität München
[2] Department of Psychology, University of Otago, Dunedin
[2] Max Planck Institute for Metabolism Research, Cologne
[*] Correspondence: hess@biomed.ee.ethz.ch

ABSTRACT

Allostatic self-efficacy (ASE) represents a computational theory of fatigue and depression. In brief, it postulates that (i) fatigue is a feeling state triggered by a metacognitive diagnosis of loss of control over bodily states (persistently elevated interoceptive surprise); and that (ii) generalisation of low self-efficacy beliefs beyond bodily control induces depression.

Here, we convert ASE theory into a structural causal model (SCM). This allows for identification of empirically testable hypotheses regarding casual relationships between variables of interest. We use conditional independence tests on questionnaire data from healthy volunteers (N=60) to identify contradictions to the proposed SCM. Moreover, we estimate two causal effects proposed by ASE theory using three different methods.

Our analyses suggest that, in healthy volunteers, the data are not fully compatible with the proposed SCM. We therefore refine the SCM and present an updated version for future research. Second, we confirm the predicted negative average causal effect from metacognition of allostatic control to fatigue across all three different methods of estimation.

Our study represents an initial attempt to refine and formalise ASE theory using methods from causal inference. Our results confirm key predictions from the ASE theory but also suggest revisions which require empirical verification in future studies.

INTRODUCTION

Fatigue is a prominent symptom of major clinical significance in numerous disorders across medical disciplines [7,44]. It is fundamentally disabling for patients and profoundly affects their quality of life [10]. Fatigue is a common feature across a wide range of immunological and endocrine disorders, cancer, and neuropsychiatric diseases. In particular, it constitutes one of the core diagnostic criteria of major depression in standard psychiatric classification schemes (ICD-10 and DSM-5; [2,22]).

The clinical concept of fatigue is a heterogeneous construct, and fatiguability of cognitive and motor processes needs to be distinguished from the subjective perception of fatigue [18]. This study focuses on the latter. The pathophysiological mechanisms leading to fatigue are likely diverse [18]. Previous theories have focused on a variety of neurophysiological, immunological and inflammatory processes. Unfortunately, there are no mechanistically interpretable clinical tests available for fatigue that could be used to guide individual treatment [18].

More recently, a novel perspective on fatigue has been proposed – the allostatic self-efficacy theory (ASE; [18,27,40]). The ASE theory is based on computational concepts of brainbody interactions [27,40] which, in turn, are conceptually related to and inspired by Bayesian theories of perception (predictive coding; [12]) and action (active inference; [13]). The ASE theory emphasises the role of two cognitive factors for fatigue: interoception and metacognition.

Interoception corresponds to the perception of bodily states and is of major importance for understanding determinants of mental health [15,20]. Many contemporary concepts of interoception are grounded in Bayesian theories of perception and conceptualise interoception as an inference process based on the brain's generative model of sensory inputs from the body [1,15,27,28,36,37]. More specifically, interoception can be conceptualised as "inferences about bodily (physiological and biochemical) states that are coupled to regulatory processes which serve to control these states" [41]. Metacognition can be summarised as cognition about cognition [11], comprising a variety of evaluation processes by which the brain monitors its own performance. Building on a generic mathematical model of brainbody interactions, the ASE theory describes how the brain attempts to control bodily states via monitoring interoceptive surprise (as an index of the degree of dyshomeostasis; [40]).

In brief, the ASE theory proposes that the subjective experience of fatigue arises when, in a situation of persistent dyshomeostasis (and thus enduringly elevated interoceptive surprise), the brain arrives at the metacognitive diagnosis that its control over bodily states is failing; a condition also referred to as low allostatic self-efficacy. (Put differently, fatigue is a feeling state signalling the imperative need to rest because regulatory actions fail to resolve dyshomeostasis.) Once a generalisation of low self-efficacy beliefs beyond the body has taken place, leading to a general sense of helplessness and perceived lack of control, this is postulated to trigger the onset of depression [33,40].

At present, the ASE theory is arguably the only concept of fatigue that explains its ubiquitous occurrence across chronic disorders. It offers testable predictions based on either (i) computational quantities (prediction error or surprise) which can be estimated from behavioural and/or neurophysiological data or on (ii) self-report data about perceived control over bodily states (metacognition of allostatic control). In this study, we focus on the latter option.

Empirically, there is initial evidence that metacognition of allostatic control – as measured by a self-report questionnaire – is inversely associated with fatigue, as predicted by ASE theory [33]. However, a comprehensive investigation of the predictions made by the ASE theory is still lacking to date. Furthermore, as almost all disease concepts in psychiatry, ASE theory has been formulated verbally, but not as a precise causal model.

Here, we present an initial attempt to tackle the latter issue. To this end, we identify variables of central interest in the ASE theory, namely metacognition of allostatic control (M; specifically, the feeling of being in control over one's own bodily states), fatigue (F), general self-efficacy (S), and depression (D). We then formalize the causal structure implied by the ASE theory in the language of causal inference, more precisely, in the form of a structural causal model (SCM; [5,24,25]). In contrast to classical probabilistic models, an SCM induces not only an observational distribution but also a set of so-called interventional distributions. In other words, an SCM predicts how a system reacts under interventions [43]. We make use of a publicly available empirical dataset to test key aspects of the structure of the proposed SCM. Moreover, we use established methods for the estimation of average causal effects focusing on central aspects of the ASE theory.

MATERIALS AND METHODS

Empirical Dataset

In this work, we used data from a previous study conducted at the Translational Neuromodeling Unit (TNU) Zurich, the perception of breathing in the human brain (PBIHB) study; a detailed description of the dataset can be found elsewhere [14]. It comprises behavioural, questionnaire and neuroimaging data from 60 healthy individuals. The questionnaire data used for our analysis are freely available for download from the Zenodo open data repository at `https://doi.org/10.5281/zenodo.10992529`. Participants completed a battery of psychological questionnaires assessing subjective affective measures, both general and breathing-specific subjective interoceptive beliefs as well as measures of general positive and negative affect, resilience, self-efficacy and fatigue.
For our analysis, we focused on the following measures as representations of the central quantities of the ASE theory:

- **fatigue (F)**: Fatigue Severity Scale (FSS)
- **general self-efficacy (S)**: General Self-Efficacy Scale (GSES)
- **depression (D)**: Centre for Epidemiologic Studies Depression Scale (CES-D)
- **metacognition of allostatic control (M)**: Sum of the subscales 3 (not worrying) and 8 (trusting) of the Multidimensional Assessment of Interoceptive Awareness ($\text{MAIA}_{3,8}$).

One important caveat is that, to our knowledge, there does not yet exist a measure that was specifically developed for the construct of M (metacognition of allostatic control, i.e. the feeling of being in control over one's own bodily states). In this study, as a proxy measure, we use the sum of the subscales 3 and 8 of the MAIA questionnaire. These subscales reflect an individual's tendency not to experience distress in response to bodily inputs signalling dyshomeostasis and to perceive the body as a safe place, respectively. The sum of these subscales was used in a previous study testing predictions from ASE theory [33] and may currently represent the best approximation to M that is easily applied in practice.

SCM of the ASE theory

An SCM [5,25] over variables $\mathbf{X} = [X_1, ..., X_n]$ comprises a set of structural equations and distributions of the noise variables (a formal definition of an SCM is provided in Appendix A1). The structural equations together with the noise distributions induce the observational distribution $\mathbb{P}_{\mathbf{X}}$ as simultaneous solution to the structural equations [43]. In addition to the observational distribution, an SCM induces interventional distributions. Each intervention denotes a scenario in which we fix a certain subset of the variables to a certain value, e.g. $\mathbb{P}_{do(X_1:=x_1)}$.

Under assumptions of linearity and normality, the SCM of the ASE theory takes the following form:

$$A = N_a \tag{1}$$

$$G = N_g \tag{2}$$

$$M = \theta_1 A + \theta_2 G + N_m \tag{3}$$

$$F = \theta_3 M + \theta_4 A + \theta_5 G + N_f \tag{4}$$

$$S = \theta_6 A + \theta_7 G + N_s \tag{5}$$

$$D = \theta_8 F + \theta_9 S + \theta_{10} FS + \theta_{11} A + \theta_{12} G + N_d \tag{6}$$

where A stands for age, G for gender, M for metacognition of allostatic control, F for fatigue, S general self-efficacy, D for depression, and N_i are jointly independent noise variables. $\forall i \neq g$, N_i follows a normal distribution and N_g is a Bernoulli random variable.

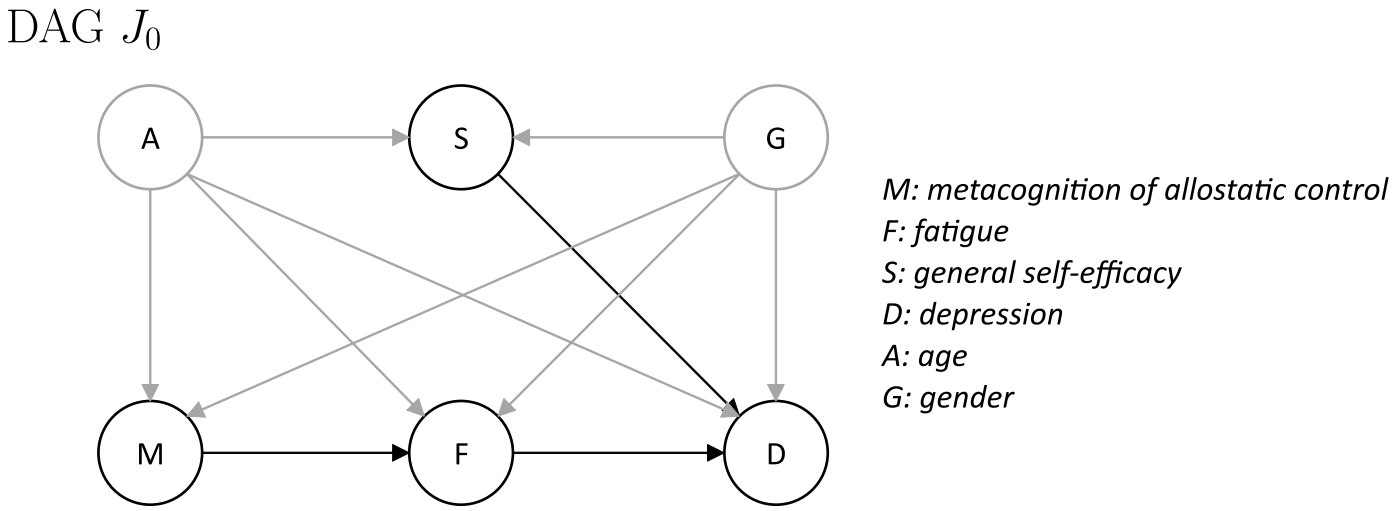

Figure 1: Directed acyclic graph (DAG) J_0 summarizing the key proposal of the allostatic self-efficacy theory (ASE; [40]). The DAG J_0 is representative for the induced observational distribution $\mathbb{P}$ and the interventional distributions induced by interventions on metacognition of allostatic control (M; $\mathbb{P}_{do(M:=m)}$), fatigue (F; $\mathbb{P}_{do(F:=f)}$) or general self-efficacy (S; $\mathbb{P}_{do(S:=s)}$). The other variables in the graph are depression (D), age (A) and gender (G). Black edges represent causal directions as proposed by the ASE theory, grey edges represent effects that are not explicitly part of the ASE theory but are likely to exist.

Figure 1 displays a graphical summary of the causal structure implied by the ASE theory in the form of a directed acyclic graph (DAG) J_0. The directed edge from metacognition of allostatic control (M) to fatigue (F) represents the prediction that fatigue arises as a consequence of a metacognitive diagnosis by the brain – i.e. the brain concludes that it has low control over its bodily states. When this low allostatic self-efficacy (for which fatigue is the accompanying feeling state) is combined with beliefs of lack of control in other domains than the body (low general self-efficacy), this is predicted to lead to the onset of depression. These effects are represented by the directed edges from fatigue (F) to depression (D) and from general self-efficacy (S) to depression (D). The variables age (A) and gender (G) are not explicitly part of the ASE theory, but are known to be associated with the central quantities of the theory. Hence, the DAG J_0 in Figure 1 is representative for the induced observational distribution $\mathbb{P}$ and the interventional distributions induced by interventions on metacognition of allostatic control (M; $\mathbb{P}_{do(M:=m)}$), fatigue (F; $\mathbb{P}_{do(F:=f)}$) or general self-efficacy (S; $\mathbb{P}_{do(S:=s)}$).

When taking a closer look at the causal graph in Figure 1, there are a number of points worth highlighting. (i) There is no direct link between metacognition of allostatic control (M) and general self-efficacy (S). (ii) There is no direct link from metacognition of allostatic control (M) to depression (D). All of its influence is mediated by fatigue (F). (iii) There is no direct link between fatigue (F) and general self-efficacy (S). While these three links are, in principle, plausible causal influences, they were not included in the original formulation of the ASE theory[40]. Whether these links should be included in a revision of the ASE theory can, in principle, be tested using methods of causal inference, given appropriate readouts of the involved quantities and relying on the assumption of the Markov condition.

Statistical Analysis

Our hypotheses as well as the entire analysis were pre-registered in a time-stamped analysis plan that is publicly available on the Zenodo open data repository at `https://doi.org/10.5281/zenodo.10559656`. Below, we explicitly highlight any deviations from the pre-specified analysis plan. The analysis code is available at `https://github.com/alexjhess/pbihb-ase-causality`. The analysis pipeline underwent an internal code review by a researcher not involved in the initial data analysis to identify errors and ensure the reproducibility of our results.

Causal structure of ASE theory in the PBIHB dataset

Learning causal structure from observational data is inherently difficult. One reason is the existence of models that are observationally but not interventionally equivalent[6,25,26,43]. This has several implications (e.g. see[43]), one of them being that without assumptions, it is impossible to learn causal structure from observational data.

In graphical models, the Markov condition (see e.g.[16]) is a formalisation of the following principle (sometimes referred to as Reichenbach's common cause principle): If two random variables X and Y are dependent, then there must be some cause-effect structure that explains the observed dependence. That is, either X causes Y, or Y causes X, or another unobserved variable H causes both X and Y, or some combination of the aforementioned[29]. A formal definition of the Markov condition is presented in Appendix A2. The Markov condition establishes a connection from graphical separation properties (d-separation; see Appendix A3 for a formal definition) to conditional independencies in the distribution. Any distribution induced by an acyclic SCM satisfies the Markov condition with respect to the corresponding graph[17,25]. Hence, the Markov condition is typically considered to be a mild assumption.

Assuming that the observational distribution $\mathbb{P}$ induced by the SCM of the ASE theory (equations 1-6) is Markov with respect to the DAG J_0, we tested whether we find any contradictions to the structure of the DAG J_0 in the PBIHB dataset. More precisely, we examined the three predictions described in the last paragraph of and formalised as part of our pre-registered **Hypothesis 1**: Data from the PBIHB study satisfy the following conditional independence statements:

(i) $M \perp\!\!\!\perp S \mid A, G$

(ii) $M \perp\!\!\!\perp D \mid F, A, G$ and $M \perp\!\!\!\perp D \mid F, A, G, S$

(iii) $F \perp\!\!\!\perp S \mid A, G$ and $F \perp\!\!\!\perp S \mid A, G, M$

As a statistical test for conditional independence, we used the asymptotic χ^2 test on the mutual information for conditional Gaussians (MI_{cg}) for mixed discrete and normal variables as implemented in the R package **bnlearn**[35], using a significance level $\alpha = 0.01$ (Bonferroni corrected).

Since conditional independence testing is a difficult statistical problem[38], we validated our results using two alternative methods: a kernel conditional independence test (KCI;[45]) as implemented in the R package **CondIndTests**, and a test based on the generalised covariance measure (GCM;[38]) as implemented in the R package **GeneralisedCovarianceMeasure**. These additional tests of conditional independence were not part of our pre-specified analysis. We decided to conduct these additional tests to evaluate the robustness of our results across different methods of conditional independence testing (i.e. a sensitivity analysis). We used the same significance level $\alpha = 0.01$ for the KCI as well as the GCM based tests to ensure compatibility with the pre-specified tests.

Estimating the average causal effect from M to F

ASE theory predicts that fatigue is a feeling state that is triggered by a metacognitive diagnosis of loss of control over bodily states. We aimed to test this prediction as part of our **Hypothesis 2**: There is a negative average causal effect from metacognition of allostatic control (M) to fatigue (F)

$$\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}\left[F\right] = \theta_3. \tag{7}$$

Adjusting for covariates is one of various methods for estimating causal effects from observational data. Suppose we are interested in finding the effect of M on F and assume the factors deemed relevant to the problem are structured as in Figure 1. In other words, we are interested in calculating the intervention distribution $\mathbb{P}_{do(M:=m)}(f)$. Given a valid adjustment set (VAS) $\mathbf{Z}$, here e.g. $\mathbf{Z} = (A, G)$, the intervention distribution can be calculated (see [23,30,39]) as $\mathbb{P}_{do(M:=m)}(f) = \sum_z \mathbb{P}(f \mid m, \mathbf{z})\mathbb{P}(\mathbf{z})$, since

$$\begin{aligned}
\mathbb{P}_{do(M:=m)}(f) &= \sum_{\mathbf{z}} \mathbb{P}_{do(M:=m)}(f, m, \mathbf{z}) && (8)\\
&= \sum_{\mathbf{z}} \mathbb{P}_{do(M:=m)}(f \mid m, \mathbf{z})\mathbb{P}_{do(M:=m)}(m, \mathbf{z}) && (9)\\
&= \sum_{\mathbf{z}} \mathbb{P}_{do(M:=m)}(f \mid m, \mathbf{z})\mathbb{P}_{do(M:=m)}(\mathbf{z}) && (10)\\
&= \sum_{\mathbf{z}} \mathbb{P}(f \mid m, \mathbf{z})\mathbb{P}(\mathbf{z}) && (11)
\end{aligned}$$

where in the last step one can use the fact that causal relationships are autonomous under interventions (this property is sometimes referred to as "autonomy") [26].

In linear Gaussian systems, a causal effect from M to F can be approximated by $\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}\left[F\right]$ (see e.g. [26]). Assuming that $\mathbf{Z}$ is a VAS for $\{M, F\}$ and $\{M, F\}, \mathbf{Z}$ follow a Gaussian distribution, then the conditional $F \mid M = m, \mathbf{Z} = \mathbf{z}$ follows a Gaussian distribution as well. Hence, the mean of the distribution is given by

$$\mathbb{E}\left[F \mid M = m, \mathbf{Z} = \mathbf{z}\right] = \theta_3 m + \mathbf{b}^t\mathbf{z} \tag{12}$$

for some θ_3 and $\mathbf{b}$. It follows from equation 11 that

$$\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}\left[F\right] = \theta_3 \tag{13}$$

One can estimate the conditional mean (eq. 12) by regressing F on M and $\mathbf{Z}$ and subsequently reading off the regression coefficients for M. Alternatively, more sophisticated techniques for estimation of the average causal effect can be used, such as the propensity score method [32] and double/debiased machine learning (DML; [8]). In Appendix B, the two methods are described in more detail.

As pre-specified in our analysis plan, we conducted linear regression in combination with a one-sided t-test on the regression coefficient of M to evaluate Hypothesis 2. We compared our estimate of the causal effect from M to F obtained via linear regression with the results obtained from using more sophisticated estimation techniques, i.e. the propensity score method [32] and DML [8], following our pre-registered analysis plan.

Estimating the average causal effect from F^*S on D

Another prediction of ASE theory is that fatigue, in combination with a generalisation of low self-efficacy beliefs beyond bodily control, induces depression. We formalised this prediction as part of our **Hypothesis 3**: There is a negative average causal effect of the interaction term between fatigue and general self-efficacy (F^*S) on depression (D)

$$\frac{\partial}{\partial f \partial s}\mathbb{E}_{do(F:=f,S:=s)}\left[D\right] = \theta_{10}. \tag{14}$$

Evaluation of Hypothesis 3 followed the same line of reasoning as for Hypothesis 2. We used linear regression in combination with a one-sided t-test on the regression coefficient of F^*S. Subsequently, we compared the resulting estimate to the results obtained

using the propensity score method and DML.

RESULTS

Raw Data

Figure 2 shows a scatter plot matrix of the raw data. Displayed are the measures for all variables A, G, M, F, S, D used in the analysis.

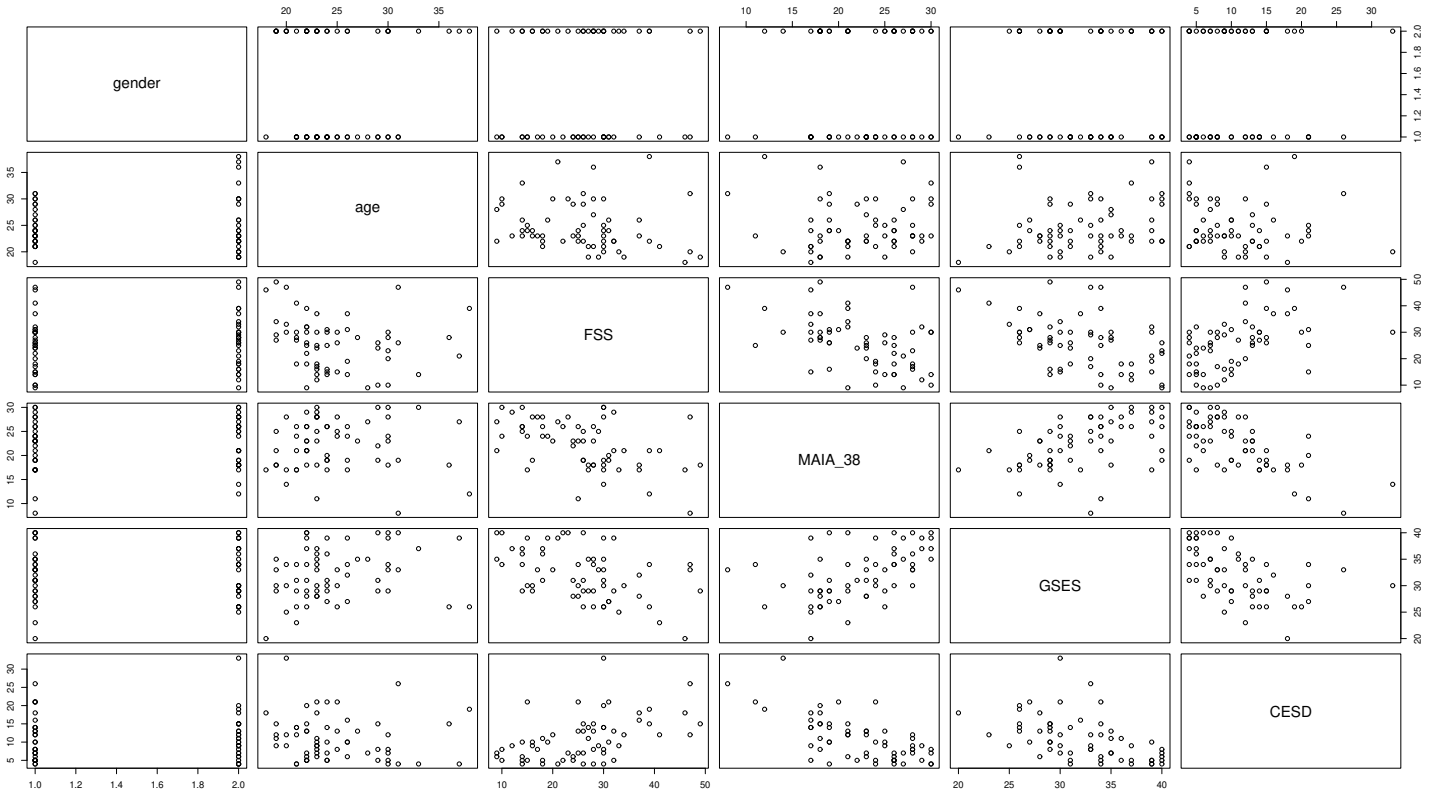

Figure 2: Scatter plot matrix of raw data used in the analysis. Displayed are all the pairwise scatter plots of the variables used for the analysis in a matrix format. For example, the scatter plot located on the intersection of row 3 and column 2 is a plot of variables age versus fatigue (as measured by the FSS). The variables displayed are gender, age, fatigue (assessed by the FSS), metacognition of allostatic control (assessed by the $\text{MAIA}_{3,8}$), self-efficacy (assessed by the GSES) and depression (assessed by the CES-D).

Results from the Statistical Analysis

Causal structure of ASE theory in the PBIHB dataset

Table 1 displays the results from conditional independence testing to evaluate the three predictions formulated as part of Hypothesis 1. The results can be summarised as follows:

Table 1: Results from different conditional independence test methods (MI_{cg}, GCM, KCI) for the three predictions formulated as part of Hypothesis 1. Results are presented for three different test methods. An asterisk indicates statistically significant evidence against the null hypothesis (H_0: variables are conditionally independent) using the pre-specified level $\alpha = 0.01$, which corresponds to a threshold of $p < 0.05$ Bonferroni corrected for the multiple comparisons of the five tests, p-values are shown in parentheses.

Hypothesis 1	d**-separation statement**	MI_{cg} (p-value)	**GCM** (p-value)	**KCI** (p-value)
(i)	$M \perp\!\!\!\perp_{J_0} S \mid A, G$	22.044* (1.634e-05)	4.254* (2.104e-05)	26.451* (5.194e-06)
(ii)	$M \perp\!\!\!\perp_{J_0} D \mid F, A, G$	24.167* (5.652e-06)	-3.131* (0.001743)	8.513* (0.001346)
	$M \perp\!\!\!\perp_{J_0} D \mid F, A, G, S$	16.883* (0.000216)	-2.574 (0.010064)	2.992 (0.022626)
(iii)	$F \perp\!\!\!\perp_{J_0} S \mid A, G$	13.010* (0.001496)	-3.390* (0.000700)	13.613* (0.001279)
	$F \perp\!\!\!\perp_{J_0} S \mid A, G, M$	4.057 (0.131500)	-2.088 (0.036799)	2.013 (0.118908)

(i) $M \not\perp\!\!\!\perp S \mid A, G$. We find significant evidence that metacognition of allostatic control (M) and general self-efficacy (S) are not independent conditional on age (A) and gender (G) across all three different conditional independence test methods. In other words, we find a contradiction regarding the conditional independence of M and S given A, G, within the DAG J_0.

(ii) $M \not\perp\!\!\!\perp D \mid F, A, G$ and $M \perp\!\!\!\perp D \mid F, A, G, S$. We find significant evidence that M and depression (D) are not independent conditional on fatigue (F), A, G across all three methods for conditional independence testing. This result is consistent with our findings for (i) in the sense that if we add a directed edge from M to S in the DAG J_0 (Figure 1), the only set of variables that d-separates M and D is the set F, A, G, S (and not F, A, G). However, the results for conditional independence tests of M and D conditional on F, A, G, S are mixed with 2 out of 3 tests (GCM and KCI) not reaching the pre-specified significance level $\alpha = 0.01$. Hence further evidence is needed to draw conclusions regarding the statement $M \perp\!\!\!\perp D \mid F, A, G, S$.

(iii) $F \not\perp\!\!\!\perp S \mid A, G$ and $F \perp\!\!\!\perp S \mid A, G, M$. When looking at the conditional independence between F and S, the results depend on the set of variables that we condition on. We find significant evidence that F and S are not independent conditional on A, G across all three different test methods. However, we fail to reject the null hypothesis that F and S are independent conditional on the set M, A, G consistently across all three different test methods. This result is also in line with our findings for (i) in the sense that if we add a directed edge from M to S in the DAG J_0 (Figure 1), the only set of variables that d-separates F and S is the set M, A, G.

Estimating the average causal effect from M to F

As predicted by the ASE theory, we find significant evidence for a negative average causal effect from metacognition of allostatic control (M) to fatigue (F) $\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}[F] = \theta_3$ across all three different estimation methods. The resulting estimates $\hat{\theta}_3$ for the VAS $\mathbf{Z} = (A, G)$ are displayed in Table 2 alongside lower and upper bounds of a 95% confidence interval for $\hat{\theta}_3$, the corresponding value of the t-statistic as well as the p-value for the one-sided t-test.

Table 2: Average causal effect from M to F using $\mathbf{Z} = (A, G)$. Displayed are estimates of the average causal effect from M to F $\hat{\theta}_3$ across three different methods to adjust for the covariates $\mathbf{Z} = (A, G)$. We report a point estimate $\hat{\theta}_3$, the lower and upper bounds of a 95% confidence interval for $\hat{\theta}_3$, the value of the t-statistic as well as the p-value for the one-sided t-test. An asterisk indicates a statistical significance using the pre-specified level $\alpha = 0.017$ (Bonferroni-corrected).

estimation method	$\hat{\theta}_3$	confidence interval		t value	p-value
linear regression	-0.4845*	-0.712	-0.257	-4.259	3.968e-05
propensity score	-0.4816*	-0.717	-0.246	-4.092	6.689e-05
DML	-0.3872*	-0.6481	-0.1262	-2.9082	0.0018

The results from our sensitivity analysis, i.e. estimating θ_3 using a different VAS $\mathbf{Z} = (A, G, S)$, are listed in Table 3. They confirm the finding of a negative average causal effect from M to F when using $\mathbf{Z} = (A, G)$ as a VAS. The main difference between the results of the two analyses are that the second analysis using $\mathbf{Z} = (A, G, S)$ yields a slightly lower absolute value for $\hat{\theta}_3$ as well as a non-significant p-value using the DML method.

Table 3: Average causal effect from M to F using $\mathbf{Z} = (A, G, S)$. Displayed are estimates of the average causal effect from M to F $\hat{\theta}_3$ across three different methods to adjust for the covariates $\mathbf{Z} = (A, G, S)$. We report a point estimate of $\hat{\theta}_3$, the lower and upper bounds of a 95% confidence interval for $\hat{\theta}_3$, the value of the t-statistic as well as the p-value for the one-sided t-test. An asterisk indicates a statistical significance using the pre-specified level $\alpha = 0.017$ (Bonferroni-corrected).

estimation method	$\hat{\theta}_3$	confidence interval		t value	p-value
linear regression	-0.3545*	-0.610	-0.099	-2.785	0.0037
propensity score	-0.3775*	-0.692	-0.063	-2.400	0.0098
DML	-0.2049	-0.563	0.153	-1.122	0.1309

Estimating the average causal effect from F^*S to D

We do not find evidence for the predicted negative average causal effect of the interaction term between fatigue and general self-efficacy (F^*S) on depression (D) $\frac{\partial}{\partial f \partial s}\mathbb{E}_{do(F:=f,S:=s)}[D] = \theta_{10}$ across all three different estimation methods for both VAS $\mathbf{Z} = (A, G)$ and $\mathbf{Z} = (A, G, M)$. Tables containing the resulting estimates for $\hat{\theta}_{10}$ including a 95% confidence interval and the value of the t-statistic as well as the p-value for the one-sided t-test are listed in the Appendix C.

DISCUSSION

In this paper, we proposed a formulation of the allostatic self-efficacy (ASE) theory of fatigue and depression in the language of causal inference. Specifically, we identified the variables of central interest to the ASE theory and formulated a structural causal model (SCM) under assumptions of linearity and normality. The SCM as well as the induced directed acyclic graph (DAG) describe the direction of causality among these variables. Using data of 60 healthy individuals from a previous study on interoception of breathing and its relation with several psychopathological constructs[14], we tested the proposed causal model empirically. Relying on the assumption of the Markov condition, we used the dataset to search for contradictions to conditional independence statements (Hypothesis 1) that are implied by the graph structure (d-separation). In a second and third step, we estimated the value of two causal effects that are predicted by the ASE theory using methods of covariate adjustment, propensity scores and double/debiased machine learning. As predicted by the ASE theory, we found a statistically significant negative average causal effect from metacognition of allostatic control (M) to fatigue (F) $\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}[F] = \theta_3$ across all three methods of estimation. Our sensitivity analysis using a different valid adjustment set largely confirmed this finding with two out of three estimation methods yielding a significant result.

The assumption of the Markov condition establishes a connection from d-separation statements in a causal graph to conditional independence statements in the distribution. In the analysis of Hypothesis 1, we tested concrete predictions implied by the DAG J_0 (Figure 1). (i) Using the the data from the PBIHB study, we were able to reject the null hypothesis of $M \perp\!\!\!\perp S \mid A, G$ at the pre-specified level $\alpha = 0.01$. (ii) We found significant evidence against $M \perp\!\!\!\perp D \mid F, A, G$ in the empirical data set. However, in line with the graph structure J_0 implied by the ASE theory, we did not find clear evidence against $M \perp\!\!\!\perp D \mid F, A, G, S$. That is, only one out of three conditional independence tests rejected the null hypothesis of metacognition of allostatic control (M) being independent from depression (D) conditional on the set F, A, G, S. (iii) We also found significant evidence against $F \perp\!\!\!\perp S \mid A, G$ in the empirical data. Yet, we did not find any evidence against (iii) $F \perp\!\!\!\perp S \mid A, G, M$. All three conditional independence test methods consistently failed to reject the null hypothesis of fatigue (F) and general self-efficacy (S) being independent given the set A, G, D, M.

There are a number of potential explanations for the results related to Hypothesis 1. The most straightforward explanation is that the proposed causal model is incorrect. This can include the presence of additional edges between nodes as well as variables that were not considered acting as mediators or confounds or a combination of all of the aforementioned. For example, although the ASE theory does not make an explicit statement about a direct link between metacognition of allostatic control (M) and general self-efficacy (S), it is plausible to assume the existence of a directed edge from M (the feeling of control over bodily states) to S (an individuals general expectation of personal mastery and control[4]). The construct of S is closely related to concepts of metacognition (see e.g.[9]) and represents a "global" construct of self-beliefs about one's capacity to achieve goals and overcome adversity; this can be understood as including more "local" domain-specific forms of self-efficacy, such as metacognition of allostatic control. From this view, the idea that metacognition of allostatic control (M) may contribute to (an thus influence) beliefs of general self-efficacy (S) is therefore not entirely unreasonable and would be a potential explanation for the results of (i) and (iii). More precisely, a directed edge from M to S would render M and S d-connected, since there would always exist a path between M and S that is not blocked by any set of variables. This cause-effect structure would explain the observed dependence between the two variables in the empirical data set according to Reichenbach's common cause principle[29]. Another consequence of introducing and edge from M to S would be that the set of variables that d-separates F and S would consist of variables A, G, M and not A, G only, which corresponds to our findings for (iii). The same is true for the set of variables d-separating M and D, which would consist of variables F, A, G, S and not F, A, G in this case, potentially explaining our findings for (ii). However, since the evidence for (ii) $M \perp\!\!\!\perp D \mid F, A, G, S$ was mixed, further research needs to bring clarity to the question of (conditional) independence of M and D.

The revised DAG J_1 (Figure 3) provides a graphical summary of the above considerations regarding the results related to Hypothesis 1. From DAG J_0 to J_1, we added a directed edge from M to S. However, there are several other potential explanations for the observed results, so this example should by no means be taken as "the correct model". If anything, this should be regarded as an updated hypothesis to be tested in future investigations.

DAG J_1

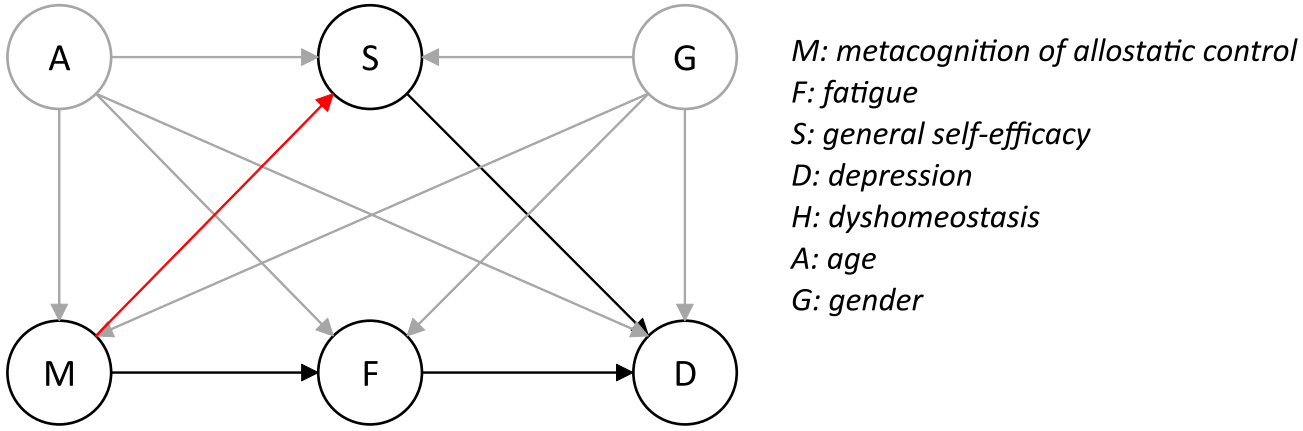

Figure 3: Updated directed acyclic graph (DAG) J_1 of the allostatic self-efficacy theory (ASE;[40]) providing one potential explanation for the observed results from analysis of Hypothesis 1. Modifications from DAG J_0 to J_1 are shown in red.

Concerning Hypothesis 2, we found evidence for a negative average causal effect from metacognition of allostatic control (M) to fatigue (F) $\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}[F] = \theta_3$ across all three estimation methods (covariate adjustment, propensity scores, DML) for two different VAS. This is in line with the prediction by the ASE theory that the subjective experience of fatigue arises as a consequence of a metacognitive diagnosis that the brain's control over bodily states is failing (low allostatic control). This also confirms findings from previous research, which identified metacognition of allostatic control (M) (operationalised by the sum of the subscales 3 and 8 of the MAIA questionnaire) to be associated with fatigue (F) scores[33]. Our new results go beyond this previous finding, in the sense that the current study suggests a direction of the effect as opposed to purely associative statements. It is worth highlighting that the estimation of the causal effect from M to F would not be affected by the proposed additional link between M to S as suggested by the analysis results concerning Hypothesis 1 (i) (see Figure 3) since the set A, G would still be a valid adjustment set (VAS).

With regard to Hypothesis 3, we did not find evidence for a negative average causal effect of the interaction term between fatigue and general self-efficacy (F^*S) on depression (D) $\frac{\partial}{\partial f \partial s}\mathbb{E}_{do(F:=f,S:=s)}[D] = \theta_{10}$. The present work is, to the best of our knowledge, the first attempt to investigate the predicted influence of the interaction between fatigue and general self-efficacy on depression. Across all three different estimation methods and using different VAS, we found, if anything, very small effects. However, one may rightfully question whether the sample in this study was adequate for testing Hypothesis 3, at least in the context of the ASE theory. This is because our participants were drawn from the general population and, not surprisingly, did not show pronounced levels of depression (compare Figure 2). By contrast, predictions of the ASE theory concerning depression assume a clinically relevant state of depression[40]. Therefore, the potential interaction effect F^*S on D remains an open question that should be addressed in the future, using samples with clinically relevant levels of depression.

The present study has a number of limitations. First of all, we are limited by certain features of the dataset at hand. In addition to the low levels of depression discussed above, the sample size (N=60) is relatively small. Therefore, it will be crucial to see whether our findings can be reproduced in larger population samples. Moreover, the dataset is purely observational, meaning that there are no interventions on any of the variables of interest. This makes the problem of causal inference (even more) challenging. A logical aim for future studies would be to use variables like M as targets for cognitive interventions.

Additionally, our analysis relies on the assumption that we have access to valid measurements of the variables in our SCM. While we employed validated and widely used measures for fatigue, depression, and general self-efficacy, there does not yet exist a val-

idated measurement tool that was specifically developed for the construct of metacognition of allostatic control (M). Here, as in previous research [33], we used a plausible proxy measure, the sum of the MAIA subscales 3 and 8 (for a detailed motivation, please see the Methods section). An important goal for future research is the development and validation of easily applicable readouts for metacognition of allostatic control (M).

Beyond the limitations of the dataset, our proposed SCM of the ASE theory is arguably only a crude approximation to reality. The most obvious concern is the one of unobserved confounds, which we articulated in more detail in the discussion of the results from Hypothesis 1. More specifically, one important limitation of the present study is that our SCM does not include sleep. While sleep is not an explicit component of the ASE theory, previous work has repeatedly demonstrated the importance of sleep quality for fatigue (e.g. [19,33]). In the present study, we did not examine the potential influence of sleep since the available dataset did not include any measures of sleep quality.

A second limitation is that we adopt the common assumption that all effects are linear and that all of the random variables follow a normal distribution (except gender). These assumptions of linearity and normality should be kept in mind when interpreting our findings for Hypotheses 2 and 3. Another potential drawback of our SCM is that we did not explicitly consider the role of time. Most of the variables in our SCM are plausibly considered to be dynamic states, i.e. their values are likely to change over time. In this work, we used a dataset representing a snapshot in time and implicitly assumed that the causal effects take place instantaneously. However, it is plausible to assume that, for example, the effects of elevated fatigue levels do not immediately lead to elevated symptoms of depression, but that this effect evolves over timescales of weeks, months or even years.

Finally, there are numerous assumptions underlying our statistical tests. Conditional independence testing, which lies at the heart of causal discovery [39], is one of its most challenging tasks [34]. For Hypothesis 1, we additionally rely on the assumption of the Markov condition. Without going into details for any of these assumptions, we highlight that the strongest of all the assumptions made throughout the entire analysis is the assumption of unconfoundedness. In other words, our results are based on the assumption that our proposed SCM contains all variables relevant for the phenomenon under consideration. However, it is likely that further variables exist that influence those in the proposed SCM (e.g. sleep, see above). The omission of these (partially unknown) variables may affect the results for all three hypotheses that we tested.

Despite the numerous limitations, this work also has several strengths worth highlighting. Foremost, we provided the first concrete formulation of the ASE theory in the language of causal inference. Our proposal of an SCM brings the content of a verbally formulated theory into the realm of concrete mathematical equations. Together with the induced DAG, they provide a formal basis for analysis and allowed us to identify a set of empirically testable hypotheses which may guide future research. Secondly, we used multiple independent methods for both conditional independence testing (Hypothesis 1) as well as the estimation of causal effects (Hypotheses 2 and 3). In this way, we are able to draw conclusions in that they do not depend on assumptions and properties of any single method. Last but not least, all of our hypotheses and statistical analysis procedures were pre-registered and specified in detail in an ex ante analysis plan (`https://doi.org/10.5281/zenodo.10559656`). Preregistration is an important and effective protection for the robustness of research, given the many degrees of freedom and the numerous cognitive biases that scientists may inadvertently be affected by [21].

CONCLUSIONS

In summary, our work provides a formal basis for testing predictions by the ASE theory of fatigue and depression in the context of causal inference. We evaluated central aspects of our proposed SCM using a publicly available dataset and provided an updated version of the SCM that accounts for our empirical findings. In addition, we were able to confirm previous findings regarding the association between metacognition of allostatic control (M) and fatigue (F). Our analysis enabled us to quantify the direction as well as the sign of the causal effect, i.e. we found a negative average causal effect from M to F $\frac{\partial}{\partial m}\mathbb{E}_{do(M:=m)}[F] = \theta_3$, as predicted by the ASE theory. Finally, we identified a number of open questions that remain to be addressed in future research and that may help unravel the mechanisms behind fatigue and depression.

ACKNOWLEDGMENTS

We wish to thank Jonas Peters for helpful discussions.

AUTHOR CONTRIBUTIONS

Conceptualization, A.J.H., D.W., J.H. and K.E.S.; methodology, A.J.H., J.H. and K.E.S.; software, A.J.H. and D.W.; validation, D.W.; formal analysis, A.J.H.; investigation, A.J.H. and O.K.H.; resources, O.K.H. and K.E.S.; data curation, O.K.H.; writing—original draft preparation, A.J.H.; writing—review and editing, A.J.H., D.W., O.K.H., J.H. and K.E.S.; visualization, A.J.H.; supervision, K.E.S.; project administration, A.J.H.; funding acquisition, O.K.H. and K.E.S. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the René and Susanne Braginsky Foundation, the ETH Foundation and the University of Zurich. O.K.H. (née Faull) was supported by a Marie Sklodowska-Curie Postdoctoral Fellowship from the European Unions Horizon 2020 research and innovation program under the grant agreement 793580, and a Rutherford Discovery Fellowship from the Royal Society Te Aprangi.

AUTHOR COMPETING INTERESTS

The authors declare no conflicts of interest.

AI ASSISTED TECHNOLOGIES

We did not use any AI assisted technologies, neither for data analysis nor during writing of the manuscript.

REFERENCES

[1] V. Ainley, M. A. J. Apps, A. Fotopoulou, and M. Tsakiris. Bodily precision: a predictive coding account of individual differences in interoceptive accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160003, Nov. 2016. doi: 10.1098/rstb.2016.0003. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2016.0003. Publisher: Royal Society.

[2] A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5 edition, 2013. URL https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596.

[3] P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. DoubleML – An Object-Oriented Implementation of Double Machine Learning in R, Jan. 2023. URL http://arxiv.org/abs/2103.09603. arXiv:2103.09603 [cs, econ, stat].

[4] A. Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191–215, Mar. 1977. ISSN 0033295X. doi: 10.1037/0033-295X.84.2.191. URL /record/1977-25733-001.

[5] K. A. Bollen. *Structural equations with latent variables*. Structural equations with latent variables. John Wiley & Sons, Oxford, England, 1989. ISBN 978-0-471-01171-2. doi: 10.1002/9781118619179. Pages: xiv, 514.

[6] S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, Oct. 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2064. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-5/Foundations-of-structural-causal-models-with-cycles-and-latent-variables/10.1214/21-AOS2064.full. Publisher: Institute of Mathematical Statistics.

[7] A. Chaudhuri and P. O. Behan. Fatigue in neurological disorders. *The Lancet*, 363(9413):978–988, Mar. 2004. ISSN 0140-6736. doi: 10.1016/S0140-6736(04)15794-2. URL https://www.sciencedirect.com/science/article/pii/S0140673604157942.

[8] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, Feb. 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[9] I. Clark and G. Dumas. The regulation of task performance: A trans-disciplinary review. *Frontiers in Psychology*, 6(JAN):1862, Jan. 2016. ISSN 16641078. doi: 10.3389/fpsyg.2015.01862. URL www.frontiersin.org. Publisher: Frontiers Media S.A.

[10] J. D. Fisk, P. G. Ritvo, L. Ross, D. A. Haase, T. J. Marrie, and W. F. Schlech. Measuring the Functional Impact of Fatigue: Initial Validation of the Fatigue Impact Scale. *Clinical Infectious Diseases*, 18(Supplement_1):S79–S83, Jan. 1994. ISSN 1058-4838. doi: 10.1093/clinids/18.Supplement_1.S79. URL https://doi.org/10.1093/clinids/18.Supplement_1.S79.

[11] S. M. Fleming and R. J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012. ISSN 14712970. doi: 10.1098/rstb.2011.0417. Publisher: Royal Society.

[12] K. Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, Apr. 2005. ISSN 0962-8436. doi: 10.1098/rstb.2005.1622. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2005.1622. Publisher: Royal Society.

[13] K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel. Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3):227–260, Mar. 2010. ISSN 1432-0770. doi: 10.1007/s00422-010-0364-z. URL https://doi.org/10.1007/s00422-010-0364-z.

[14] O. K. Harrison, L. Köchli, S. Marino, R. Luechinger, F. Hennel, K. Brand, A. J. Hess, S. Frässle, S. Iglesias, F. Vinckier, F. H. Petzschner, S. J. Harrison, and K. E. Stephan. Interoception of breathing and its relationship with anxiety. *Neuron*, 0(0):1–14, Oct. 2021. ISSN 0896-6273. doi: 10.1016/J.NEURON.2021.09.045. URL http://www.cell.com/article/S0896627321007182/fulltext. Publisher: Elsevier.

[15] S. S. Khalsa, R. Adolphs, O. G. Cameron, H. D. Critchley, P. W. Davenport, J. S. Feinstein, J. D. Feusner, S. N. Garfinkel, R. D. Lane, W. E. Mehling, A. E. Meuret, C. B. Nemeroff, S. Oppenheimer, F. H. Petzschner, O. Pollatos, J. L. Rhudy, L. P. Schramm, W. K. Simmons, M. B. Stein, K. E. Stephan, O. Van den Bergh, I. Van Diest, A. von Leupoldt, M. P. Paulus, V. Ainley, O. Al Zoubi, R. Aupperle, J. Avery, L. Baxter, C. Benke, L. Berner, J. Bodurka, E. Breese, T. Brown, K. Burrows, Y. H. Cha, A. Clausen, K. Cosgrove, D. Deville, L. Duncan, P. Duquette, H. Ekhtiari, T. Fine, B. Ford, I. Garcia Cordero, D. Gleghorn, Y. Guereca, N. A. Harrison, M. Hassanpour, T. Hechler, A. Heller, N. Hellman, B. Herbert, B. Jarrahi, K. Kerr, N. Kirlic, M. Klabunde, T. Kraynak, M. Kriegsman, J. Kroll, R. Kuplicki, R. Lapidus, T. Le, K. L. Hagen, A. Mayeli, A. Morris, N. Naqvi, K. Oldroyd, C. Pané-Farré, R. Phillips, T. Poppa, W. Potter, M. Puhl, A. Safron, M. Sala, J. Savitz, H. Saxon, W. Schoenhals, C. Stanwell-Smith, A. Teed, Y. Terasawa, K. Thompson, M. Toups, S. Umeda, V. Upshaw, T. Victor, C. Wierenga, C. Wohlrab, H. w. Yeh, A. Yoris, F. Zeidan, V. Zotev, and N. Zucker. Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(6):501–513, June 2018. ISSN 24519030. doi: 10.1016/j.bpsc.2017.12.004. Publisher: Elsevier Inc.

[16] S. L. Lauritzen and S. L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York, May 1996. ISBN 978-0-19-852219-5.

[17] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990. ISSN 1097-0037. doi: 10.1002/net.3230200503. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230200503.

[18] Z. M. Manjaly, N. A. Harrison, H. D. Critchley, C. T. Do, G. Stefanics, N. Wenderoth, A. Lutterotti, A. Müller, and K. E. Stephan. Pathophysiological and cognitive mechanisms of fatigue in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry*, 90(6):642–651, June 2019. ISSN 1468330X. doi: 10.1136/jnnp-2018-320050. URL http://jnnp.bmj.com/. Publisher: BMJ Publishing Group.

[19] V. Nociti, F. A. Losavio, V. Gnoni, A. Losurdo, E. Testani, C. Vollono, G. Frisullo, V. Brunetti, M. Mirabella, and G. Della Marca. Sleep and fatigue in multiple sclerosis: A questionnaire-based, cross-sectional, cohort study. *Journal of the Neurological Sciences*, 372:387–392, Jan. 2017. ISSN 0022-510X. doi: 10.1016/j.jns.2016.10.040. URL https://www.sciencedirect.com/science/article/pii/S0022510X16306840.

[20] C. L. Nord and S. N. Garfinkel. Interoceptive pathways to understand and treat mental health conditions. *Trends in Cognitive Sciences*, 0(0), Apr. 2022. ISSN 1364-6613. doi: 10.1016/J.TICS.2022.03.004. URL http://www.cell.com/article/S1364661322000626/fulltext. Publisher: Elsevier.

[21] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, Mar. 2018. doi: 10.1073/pnas.1708274114. URL https://www.pnas.org/doi/10.1073/pnas.1708274114. Publisher: Proceedings of the National Academy of Sciences.

[22] W. H. Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization, 2004. ISBN 978-92-4-154649-2. URL https://iris.who.int/handle/10665/42980. Accepted: 2012-06-16T14:40:38Z Journal Abbreviation: ICD-10.

[23] J. Pearl. Belief networks revisited. *Artificial Intelligence*, 59:49–56, 1993.

[24] J. Pearl. Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688, 1995. ISSN 0006-3444. doi: 10.2307/2337329. URL https://www.jstor.org/stable/2337329. Publisher: [Oxford University Press, Biometrika Trust].

[25] J. Pearl. *Causality*. Cambridge University Press, Sept. 2009. ISBN 978-0-521-89560-6. Google-Books-ID: f4nuexsNVZIC.

[26] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. URL https://library.oapen.org/handle/20.500.12657/26040. Accepted: 2019-01-20 23:42:51.

[27] F. H. Petzschner, L. A. Weber, T. Gard, and K. E. Stephan. Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. *Biological Psychiatry*, 82(6):421–430, Sept. 2017. ISSN 18732402. doi: 10.1016/j.biopsych.2017.05.012. Publisher: Elsevier USA.

[28] G. Pezzulo, F. Rigoli, and K. Friston. Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134:17–35, Nov. 2015. ISSN 18735118. doi: 10.1016/j.pneurobio.2015.09.001. Publisher: Elsevier Ltd.

[29] H. Reichenbach. *The Direction of Time*. Dover Publications, Mineola, N.Y., 1956.

[30] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, Jan. 1986. ISSN 0270-0255. doi: 10.1016/0270-0255(86)90088-6. URL https://www.sciencedirect.com/science/article/pii/0270025586900886.

[31] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550, Sept. 2000. ISSN 1044-3983. URL https://journals.lww.com/epidem/fulltext/2000/09000/marginal_structural_models_and_causal_inference_in.11.aspx.

[32] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, Apr. 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

[33] M. Rouault, I. Pereira, H. Galioulline, S. M. Fleming, K. E. Stephan, and Z.-M. Manjaly. Interoceptive and metacognitive facets of fatigue in multiple sclerosis. *European Journal of Neuroscience*, 58(2):2603–2622, 2023. ISSN 1460-9568. doi: 10.1111/ejn.16048. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.16048.

[34] J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, Mar. 2018. URL https://proceedings.mlr.press/v84/runge18a.html. ISSN: 2640-3498.

[35] M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35:1–22, July 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03. URL https://doi.org/10.18637/jss.v035.i03.

[36] A. K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, Nov. 2013. ISSN 13646613. doi: 10.1016/j.tics.2013.09.007. Publisher: Elsevier Current Trends.

[37] A. K. Seth, K. Suzuki, and H. D. Critchley. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 3(JAN):395, Jan. 2012. ISSN 16641078. doi: 10.3389/fpsyg.2011.00395. URL www.frontiersin.org. Publisher: Frontiers.

[38] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1857. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-3/The-hardness-of-conditional-independence-testing-and-the-generalised-covariance/10.1214/19-AOS1857.full. Publisher: Institute of Mathematical Statistics.

[39] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000. ISBN 978-0-262-19440-2.

[40] K. E. Stephan, Z. M. Manjaly, C. D. Mathys, L. A. Weber, S. Paliwal, T. Gard, M. Tittgemeyer, S. M. Fleming, H. Haker, A. K. Seth, and F. H. Petzschner. Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10(NOV2016):550, Nov. 2016. ISSN 16625161. doi: 10.3389/fnhum.2016.00550. URL www.frontiersin.org. Publisher: Frontiers Media S. A.

[41] B. Toussaint, J. Heinzle, and K. E. Stephan. A computationally informed distinction of interoception and exteroception. *Neuroscience & Biobehavioral Reviews*, 159: 105608, Apr. 2024. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2024.105608. URL https://www.sciencedirect.com/science/article/pii/S0149763424000770.

[42] W. M. v. d. Wal and R. B. Geskus. ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software*, 43:1–23, Sept. 2011. ISSN 1548-7660. doi: 10.18637/jss.v043.i13. URL https://doi.org/10.18637/jss.v043.i13.

[43] S. Weichwald and J. Peters. Causality in Cognitive Neuroscience: Concepts, Challenges, and Distributional Robustness. *Journal of Cognitive Neuroscience*, 33(2):226–247, Feb. 2021. ISSN 1530-8898. doi: 10.1162/jocn_a_01623.

[44] S. Wessely. Chronic Fatigue: Symptom and Syndrome. *Annals of Internal Medicine*, 134(9_Part_2):838–843, May 2001. ISSN 0003-4819. doi: 10.7326/0003-4819-134-9_Part_2-200105011-00007. URL https://www.acpjournals.org/doi/10.7326/0003-4819-134-9_part_2-200105011-00007. Publisher: American College of Physicians.

[45] K. Zhang, J. Peters, D. Janzing, and B. Schoelkopf. Kernel-based Conditional Independence Test and Application in Causal Discovery, Feb. 2012. URL http://arxiv.org/abs/1202.3775. arXiv:1202.3775 [cs, stat].

APPENDIX A. DEFINITIONS

A1. Structural Causal Model

We adopt the definition of SCMs according to [43]:

Definition 1. *An SCM over variables* $\boldsymbol{X} = [X_1, ..., X_n]$ *comprises*

- *structural equations which relate each variable* X_k *to its parents* $\boldsymbol{PA}(X_k) \subseteq \{X_1, ..., X_n\}$ *and a noise variable* N_k *via a function* f_k *such that* $X_k := f_k\left(\boldsymbol{PA}\left(X_k\right), N_k\right)$*, as well as a*
- *noise distribution* $\mathbb{P}_N$ *of the noise variables* $\boldsymbol{N} = [N_1, ..., N_n]^T$.

In a directed causal graph associated with an SCM, the nodes correspond to the variables $X_1, ..., X_n$ *and there is an edge from* X_i *to* X_j *whenever* X_i *appears on the right hand side of the equation* $X_j := f_j\left(\boldsymbol{PA}\left(X_j\right), N_j\right)$*. In other words, if* $X_i \in \boldsymbol{PA}(X_j)$ *the graph contains the edge* $X_i \to X_j$*. For this work, we assume that the graph does not contain any cycles. The structural equations together with the noise distributions induce the observational distribution* $\mathbb{P}_X$ *of* $X_1, ..., X_n$ *as simultaneous solution to the equations.*

A2. Markov condition

Definition 2. *Given a DAG* G *over nodes* $\boldsymbol{X}$*, we say that the distribution* $\mathbb{P}_{\boldsymbol{X}}$ *satisfies*

(i) the ***global Markov property*** *(MP) with respect to* G *if* $\forall$ *disjoint* $A, B, C \subseteq \boldsymbol{X}$
A *d-sep* $B \mid C \implies A \perp\!\!\!\perp B \mid C$

(ii) the ***local Markov property*** *(MP) if* $\forall j\ X_j \perp\!\!\!\perp ND_j \mid PA_j$

(iii) the ***factorisation property*** *if* $\mathbb{P}_{\boldsymbol{X}}$ *is absolutely constant with respect to a product measure and* $\forall x \forall j,\ p(x_{PA_j}) > 0 : p(\boldsymbol{x}) = p(x_1, ..., x_d) = \prod_j^d p(x_j | x_{PA_j})$

In the above definition, we used the following notation: ND_j represent the non-descendants of node X_j and PA_j denotes all nodes that have a directed edge to node X_j.

A3. d-separation

Definition 3. *d-separation is a graphical criterion whether two nodes are connected or not. Let* $\boldsymbol{X}$*,* $\boldsymbol{Y}$*,* $\boldsymbol{Z}$ *disjoint.*

(i) A path $X = i_1, ..., i_m = Y$ *is blocked by* $\boldsymbol{Z} \iff \exists$ *node* i_k *with* $i_{k-1} \to i_k \to i_{k+1}$ *and* $i_k \in \boldsymbol{Z}$
OR $\exists$ *node* i_k *with* $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ *and* $i_k \in \boldsymbol{Z}$
OR $\exists$ *node* i_k *with* $i_{k-1} \leftarrow i_k \to i_{k+1}$ *and* $i_k \in \boldsymbol{Z}$
OR $\exists$ *node* i_k *with* $i_{k-1} \to i_k \leftarrow i_{k+1}$ *and* $i_k \notin \boldsymbol{Z}$ *and* $DE(i_k) \cap \boldsymbol{Z} = \emptyset$

(ii) $\boldsymbol{X}$*,* $\boldsymbol{Y}$ *are d-connected given* $\boldsymbol{Z} \iff \exists X \in \boldsymbol{X}, Y \in \boldsymbol{Y}$ *s.t.* $\exists$ *path between* X *and* Y *that is not blocked*

(iii) if $\boldsymbol{X}$*,* $\boldsymbol{Y}$ *are not d-connected, then they are d-separated. We sometimes write* $\boldsymbol{X}$ *d-sep* $\boldsymbol{Y} \mid \boldsymbol{Z}$ *or* $\boldsymbol{X} \perp\!\!\!\perp_G \boldsymbol{Y} \mid \boldsymbol{Z}$

APPENDIX B. ESTIMATING CAUSAL EFFECTS USING COVARIATE ADJUSTMENT

B1. The "propensity score" method

In a point treatment situation one can adjust for a set of confounders $\mathbf{Z} = (A, G)$ when estimating the effect of exposure M by weighting observations i by the inverse probability weights

$$w_i = \frac{1}{P(M_i = m_i \mid \mathbf{Z}_i = \mathbf{z}_i)} \tag{15}$$

To increase statistical efficiency, one can use stabilised weights, e.g.

$$sw_i = \frac{P(M_i = m_i)}{P(M_i = m_i \mid \mathbf{Z}_i = \mathbf{z}_i)} \tag{16}$$

When dealing with a continuous exposure variable M, one can use stabilised weights

$$sw_i = \frac{f(m_i)}{f(m_i \mid \mathbf{z}_i)} \tag{17}$$

where $f(m_i)$ is the marginal density function of M, evaluated at the observed value in unit i, m_i, and $f(m_i \mid \mathbf{z}_i)$ conditional density function of M given $\mathbf{Z}$, evaluated at the observed values in unit i, $\{m_i, \mathbf{z}_i\}$ [42]. Weighting observations i by sw_i, one can fit a causal model, for instance a marginal structural model (MSM)

$$\mathbb{E}\left[F_m\right] = \beta_0 + \theta_3 m \tag{18}$$

with continuous outcome fatigue F. The response variable F_m is the potential outcome that could have been observed in a unit under study, when that unit would have received a specific treatment level m [31]. The expectation $\mathbb{E}\left[F_m\right]$ is the mean outcome, when all units under study would have received a specific treatment level m. Parameter θ_3 then quantifies the causal effect of M on F [42].

B2. Double/Debiased Machine Learning

DML removes the impact of regularisation bias and overfitting on estimation of the parameter of interest θ_3 by using Neyman-orthogonal moments and cross-fitting [8]. One application of DML is in the context of a partial linear regression model,

$$F = M\theta_3 + J_0(\mathbf{Z}) + N_f, \quad \mathbb{E}(N_f \mid M, \mathbf{Z}) = 0, \tag{19}$$

$$M = m_0(\mathbf{Z}) + V, \quad \mathbb{E}(V \mid \mathbf{Z}) = 0, \tag{20}$$

with fatigue F, metacognition of allostatic control M, a VAS $\mathbf{Z} = (A, G)$ consisting of confounding covariates and stochastic error terms N_f and V. The confounding covariates $\mathbf{Z}$ affect M and F via the functions m_0 and J_0, respectively. DML can be used to estimate θ_3, i.e. the main regression coefficient that we would like to infer, which can be interpreted as the average causal effect from M to F [3].

APPENDIX C. RESULTS FROM ESTIMATING THE AVERAGE CAUSAL EFFECT FROM F^*S TO D

Table 4: Average causal effect of the interaction term F^*S to D using $\mathbf{Z} = (A, G)$. Displayed are estimates of the average causal effect of the interaction term F^*S to D $\hat{\theta}_{10}$ across three different methods to adjust for the covariates $\mathbf{Z} = (A, G)$. We report a point estimate of $\hat{\theta}_{10}$, the lower and upper bounds of a 95% confidence interval for $\hat{\theta}_{10}$, the value of the t-statistic as well as the p-value for the one-sided t-test. An asterisk indicates a statistical significance using the pre-specified level $\alpha = 0.017$ (Bonferroni-corrected).

estimation method	$\hat{\theta}_{10}$	confidence interval		t value	p-value
linear regression	0.0281	-0.191	0.247	0.257	0.6010
propensity score	0.0142	-0.151	0.180	0.172	0.5680
DML	-0.2051	-0.476	0.066	-1.482	0.0691

Table 5: Average causal effect of the interaction term F^*S to D using $\mathbf{Z} = (A, G, M)$. Displayed are estimates of the average causal effect of the interaction term F^*S to D $\hat{\theta}_{10}$ across three different methods to adjust for the covariates $\mathbf{Z} = (A, G, M)$. We report a point estimate of $\hat{\theta}_{10}$, the lower and upper bounds of a 95% confidence interval for $\hat{\theta}_{10}$, the value of the t-statistic as well as the p-value for the one-sided t-test. An asterisk indicates a statistical significance using the pre-specified level $\alpha = 0.017$ (Bonferroni-corrected).

estimation method	$\hat{\theta}_{10}$	confidence interval		t value	p-value
linear regression	0.0671	-0.122	0.257	0.711	0.7599
propensity score	0.0337	-0.159	0.227	0.350	0.6362
DML	0.0153	-0.283	0.314	0.100	0.5399