Highlights

Towards a qualitative theory of the interruption of eating behavior change

Philippe Dague, Laurent Muller, Loïc Paulevé, Marc Irigoin-Guichandut

- A concrete example of how combining physiology, thermodynamics and psychology paves the way towards a formal theory of behavior change
- A formal causal model accounts for the phenomena of non-initiation, interruption or maintenance of eating behavior change during severe diets
- Qualitative abstraction and reasoning deals with the ignorance of individual model parameter values and the only knowledge of common evolution profiles over time of individual physiological variables
- Powerful methods of automatic reasoning on a discrete qualitative abstraction of the model allow automatic exhaustive exploration of its solution space and provides a superset of the real solution set, in terms of active causalities and their signs
- Analytical resolution of an under-approximation of the model in terms of linear ODEs provides a subset of the real solution set that happens to be equal to the previous superset, proving that it is the real solution set.
- The model provides information on which parameters are necessarily person-specific and which may be constant, and therefore handles inter-individual variability within a single formal framework.
- This work makes a new contribution to Mathematical, Computational, Biophysical and Psychological Modeling and Reasoning

Towards a qualitative theory of the interruption of eating behavior change

Philippe Dague[a,*], Laurent Muller[b], Loïc Paulevé[c], Marc Irigoin-Guichandut[d]

[a] *Université Paris-Saclay, CNRS, ENS Paris-Saclay, Inria, Laboratoire Méthodes Formelles, 4 avenue des Sciences, 91190, Gif-sur-Yvette, France*
[b] *Université de Lorraine, APEMAC, Ile du Saulcy, 57045, Metz, France*
[c] *Université de Bordeaux, CNRS, Bordeaux INP, LaBRI, 351 cours de la Libération, 33405, Talence, France*
[d] *Assistance Publique Hôpitaux de Paris, Hôpital de la Pitié-Salpêtrière, 47-83 Bd de l'Hôpital, 75013, Paris, France*

Abstract

The poor maintenance of eating behavior change is one of the main obstacles to minimizing weight regain after weight loss during diets for non-surgical care of obese or overweight patients. We start with a known informal explanation of interruption in eating behavior change during severe restriction and formalize it as a causal network involving psychological variables, which we extend with energetic variables governed by principles of thermodynamics. The three core phenomena of dietary behavior change, i.e., non-initiation, initiation followed by discontinuation and initiation followed by non-discontinuation, are expressed in terms of the value of the key variable representing mood or psychological energy, the fluctuation of which is the result of three causal relationships. Based on our experimental knowledge of the time evolution profile of the three causal input variables, we then proceed to a qualitative analysis of the resulting theory, i.e., we consider an over-approximation of it which, after discretization, can be expressed in the form of a finite integer-based model. Using Answer-Set Programming, we show that our formal model faithfully reproduces the three phenomena and, under a certain assumption, is minimal. We generalize this result by providing all the minimal models reproducing these phenomena when the possible causal relationships exerted on mood are extended to all the other variables (not just those assumed in the informal explanation), with arbitrary causality signs. Finally, by a direct analytical resolution of an under-approximation of our theory, obtained by assuming linear causalities, as a system of linear ODEs, we find exactly the same minimal models, proving that they are also equal to the actual minimal models of our theory since these are framed below and above by

*Corresponding author
Email addresses: philippe.dague@universite-paris-saclay.fr (Philippe Dague), laurent.muller@univ-lorraine.fr (Laurent Muller), loic.pauleve@labri.fr (Loïc Paulevé), marc.irigoin@aphp.fr (Marc Irigoin-Guichandut)
URL: https://orcid.org/0000-0003-1679-0804 (Philippe Dague)

the models of the under-approximation and the over-approximation. We determine which parameters need to be person-specific and which can be considered constant, i.e., we explain inter-individual variability. Our approach could pave the way for universally accepted theories in the field of behavior change and, more broadly, in other areas of psychology.

Keywords: behavior change, hypocaloric diet, formal theory, thermodynamics, causal network, over-/under-approximation, discretization, qualitative reasoning, timescale, inter-individual variability, answer-set programming

---

1. Introduction

For non-surgical care of obese or overweight patients, international recommendations advise a moderate reduction in food intake (deficit of 600-700 kcal/-day) adapted to individual food preferences of patients over a period of six months as part of a lifestyle intervention, to induce a weight loss of 5 to 10% that has been shown to improve comorbidities (Jensen et al., 2014). However, a severe reduction in food intake called a very-low-calorie diet, which was popular as a supervised diet in the '70s and '80s, because of the more considerable weight loss obtained in much less time, can still be advised nowadays to a limited number of patients, but tends to remain popular as an unsupervised diet. When supervised, this diet lasts from eleven to fourteen weeks (Jensen et al., 2014).

Weight regain after weight loss is the main pitfall of obesity medical care. This issue, already identified sixty years ago, persists in spite of the addition of cognitive-behavioral techniques and relapse prevention techniques, and of increased intensity and duration of comprehensive lifestyle interventions (Williamson, 2017), as repeatedly checked since then (e.g., Langeveld and DeVries, 2015).

In order to minimize weight regain after weight loss, a panel of experts met in 2014 at the initiative of the NIH and identified as one of the two main obstacles (the other being metabolic adaptation) the poor maintenance of behavior change, in particular of eating behavior change (MacLean et al., 2015). This is partly due to the behavioral adaptation through stimulation of feeding behavior. In order to investigate poor maintenance of eating behavior change, we propose to focus on severe restriction which differs from moderate restriction by its dynamics in weeks instead of months (here, a duration of twelve weeks is selected).

An informal explanation of the interruption of eating behavior change during severe restriction is familiar to professionals of obesity but has rarely been stated (Wadden et al., 1983; Foster and Wadden, 1995; Strasser et al., 2015). According to Strasser et al. (2015):

> *"[...] embarking upon a hypocaloric diet that was too extreme [...] At first, the dieter may experience elation at the thought of [weight] loss and pride of their rejection of food. With time, however, the limits imposed by such*

extreme diets cause effects such as depression or fatigue that make the diet impossible to sustain". (1)

All health behaviors are faced with this maintenance issue: "maintaining behavior change remains a tremendous challenge" (Nielsen et al., 2018) or "consistent, reliable behavior change remains an elusive target" (Sumner et al., 2018). To address this maintenance issue, the main strategy has been to target theories of behavior change in order to design more effective theory-based interventions (Hagger and Luszczynska, 2014) because the latter have generally proven to be superior to non-theory-based interventions, even though there is some inconsistency (Prestwich et al., 2015; Gourlan et al., 2016).

Statement (1) can be considered as a verbal psychological proto-theory, as it meets part of the definition of such a theory: "a systematic way of understanding events or situations. It is a set of concepts, definitions, and propositions that explain or predict these events or situations by illustrating the relationships between variables" (Glanz et al., 1995). The main missing parts are the definitions of its constructs.

This statement seems to be of interest primarily because of its specificity to maintenance of eating behavior change, but also because of four other characteristics: (i) it applies to a specific timescale, expressed in weeks (Zaheer et al., 1999; Vallacher et al., 2015) and is very parsimonious, (ii) it targets mood fluctuations which are associated with food restriction (Fond et al., 2013; Zhang et al., 2015), (iii) it does not include any cognitive constructs which are poorly associated with actual behavior change (e.g., Fishbein et al., 2003), and (iv) it is decontextualized because of the absence of social and environmental factors.

Recently, West et al. (2019) proposed the following formalism for representing behavior-change theories : "A given theory is represented in terms of (i) its component constructs [...] which are labeled and defined, and (ii) relationships between pairs of constructs, which may be causal, structural or semantic", which can be described in a graphical representation.

The aim of the present paper is to propose the premises of a theory of maintenance of eating behavior change based on statement (1) (which is therefore the only given conceptual input to the present study), capable of describing the interruption as well as the maintenance of eating behavior change. This postulate is described in Section 2 using the aforementioned formalism, restricting relationships to the causal type, i.e., stimulations or inhibitions, and mathematically defining constructs as variables. A first causal network is derived from statement (1), centered on the key psychological variable mood, together with another psychological variable. It is then linked to a thermodynamic model, also in the form of a causal network, the fusion of the two networks providing our combined causal theory. The timescale is specified, and the three core phenomena of dietary behavior change - non-initiation, initiation followed by discontinuation and initiation followed by non-discontinuation - are expressed in terms of mood fluctuation over time. In Section 3, we build on the fact that the qualitative evolution profiles over time of the two thermodynamic variables

that causally influence mood are known experimentally to radically simplify our causal theory as the sub-network limited to the three causal relationships exerted on mood by these two variables and the other psychological variable (whose evolution profile can be theoretically hypothesized). These three causal relationships and the three evolution profiles of the causal input variables, together with the conditions on mood of the three core phenomena, are then formally expressed in a qualitative framework, giving rise, after discretization, to an integer-based model that forms an over-approximation (i.e., less constrained) of our initial theory, its solution set therefore being a superset of the theory's actual solution set of the theory. The consistency of this discrete model, which covers all three phenomena, is then verified combinatorially by satisfiability-based methods, using Answer-Set Programming. In Section 4, we generalize our causal model by considering all possible causal relationships that could exert on mood (and not only the three derived from statement (1)) and any sign for these relationships, i.e., stimulation or inhibition, considered as parameters of the problem. By the same qualitative over-approximation, we find that there are other solutions (covering all three phenomena), in terms of causality, than the one provided by the statement and we enumerate all minimal solutions, i.e., a superset of real solutions. Then, with the idea of framing not only above but also below these real solutions, we consider in Section 5 an under-approximation (i.e., more constrained) of our theory obtained by linearization, i.e., assuming that causal relations are linear. Our entire causal network is therefore modeled as a set of linear ODEs, simple enough to be solved analytically by hand (note that, here, we do not use the experimental evolution profiles of the thermodynamic variables, whose behavior over time is in fact obtained by solving the ODEs). Given the solution function for mood behavior over time, and the conditions for the three core phenomena, a parametric mathematical analysis provides all the minimal solutions in terms of causal parameters, covering all three phenomena. It turns out that these solutions, which by construction form a subset of the actual solutions, coincide with the solutions found in Section 4, which form a superset of the actual solutions. This proves that these common solutions are exactly the real solutions of our initial problem. In addition, we take account of inter-individual variability by determining which parameters should be person-specific and which can be considered constant. Section 6 examines these results within the framework of psychological theories of behavior change. Finally Section 7 concludes.

2. Mood-focused thermodynamic-grounded theory

2.1. Formalization of statement (1) as a causal graph

Let's start by identifying the variables, i.e., the constructs of the theory.
Two variables are directly identified in (1): *Time* in "at first" and "with time", and *Weight loss*, which are both quantitative variables.
Then, five other variables, to be identified, require interpretations that are the authors' decisions. These interpretations are as follows:

1. "depression or fatigue" and "elation" are identified as mood disorder symptoms and are replaced with two other such symptoms, "loss of energy" and "increased energy", respectively: in the context of behavior change, these two symptoms are preferred because of their association with symptom "fatigue" in major depressive episodes ("fatigue or loss of energy"), and with symptom "increased activity" in hypomanic episodes ("increased energy and activity"), as reported in DSM-5 (American Psychiatric Association, DSM-5 Task Force, 2013); a variable labeled *Energy*$_\psi$, for "psychological energy" or "mood", is derived from these two symptoms, here mostly in physiological values;
2. "pride of their rejection of food" is identified as "pleasure of their rejection of food", pleasure referring to the symptom "diminished pleasure" in major depressive episodes, not mentioned in hypomanic episodes, and relabeled *Pleasure* (American Psychiatric Association, DSM-5 Task Force, 2013);
3. "such extreme diet" is relabeled *Reduced food intake* and allows the representation of the dietary behavior under study by its consequence;
4. a suitable variable labeled *Restriction* is required to produce *Reduced food intake*;
5. in the present case of a supervised diet, an exogenous variable labeled *Intervention* needs to be added.

Restriction and *Intervention* variables are binary and *Reduced food intake* is quantitative. On the other hand, the psychological variables *Energy*$_\psi$ and *Pleasure* are ordinal (Stevens, 1946) as they have to be measured using ordinal scales, e.g., items 15 and 4 respectively of Beck Depression Inventory (Beck et al., 1996), and, concerning the variable *Energy*$_\psi$, with a seven-point scale[1] as for appetite (item 8) to indicate an increase as well as a decrease.

Then, causal relationships (Pearl, 2009) between variables, that are implicit in (1), are made explicit as follows:

- "dieter may experience elation at the thought of [weight] loss" becomes *Energy*$_\psi$ is stimulated by *Weight loss*;
- "food deprivation causes effects such as depression or fatigue" becomes *Reduced food intake* inhibits *Energy*$_\psi$;
- "depression or fatigue that make the diet impossible to sustain" becomes, by contrast, *Energy*$_\psi$ stimulates *Restriction*, and provides some minimal threshold E_{min} of *Energy*$_\psi$, i.e., its lower physiological limit, that, if reached, causes the impossibility to sustain the diet, i.e., the impossibility of dietary behavior change.

[1] meaning: 1. I don't have enough energy to do anything, 2. I don't have enough energy to do very much, 3. I have less energy than I used to have, 4. I have as much energy as ever, 5. I have more energy than I used to have, 6. I have enough energy to do a lot, 7. I have enough energy to do anything

Finally, among the additional variables, *Intervention* stimulates *Restriction*, and *Restriction* stimulates *Reduced food intake*.

Furthermore, three more interpretations are required to identify additional causal relationships:

- a stimulation from *Pleasure* toward $Energy_{\psi}$, because of the positive valence of *Pleasure*;
- a stimulation from *Restriction* toward *Pleasure*, because *Pleasure* occurs when food intake is reduced;
- a stimulation from *Reduced food intake* toward *Weight loss* to represent the effect of behavior change on health outcome.

After these eight interpretations and with stimulations and inhibitions represented by $\longrightarrow$ and $\dashv$ respectively, the formal theory is presented in Fig. 1A. It is presented equivalently in Fig. 1B, where *Weight loss* is replaced with *Weight variation*, and *Reduced food intake* with *Food intake*, with a change in causality sign of three arcs because initial and replaced variables vary in opposite directions, i.e., *Weight variation* inhibits $Energy_{\psi}$, *Food intake* stimulates $Energy_{\psi}$, and *Restriction* inhibits *Food intake*.

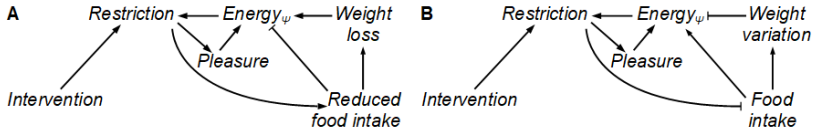

Figure 1: Causal graph formalizing statement (1): (A) initial and (B) final.

Finally, there are two last interpretations pertaining to causal relationships:

- due to the co-variation of *Pleasure* and $Energy_{\psi}$ when mood fluctuates (American Psychiatric Association, DSM-5 Task Force, 2013), an additional inhibition from *Weight variation* toward *Pleasure* could be expected, but the priority being the variation of $Energy_{\psi}$, the indirect effect of *Weight variation* on $Energy_{\psi}$ through *Pleasure* is included in its direct effect on $Energy_{\psi}$ while the effect of *Pleasure* on $Energy_{\psi}$ does not vary with mood, i.e., no additional causality is required, and *Pleasure* does not vary with mood;
- likewise, pleasure of eating, i.e., positive valence of food, represented by the stimulation from *Food intake* toward $Energy_{\psi}$ could be expected to be negatively modulated by *Weight variation*: here, a non-modulated stimulation is used as an approximation.

Of note, two other symptoms of major depressive episodes, "weight loss ... or weight gain", and "decrease or increase in appetite", that refer to variables *Weight variation* and *Food intake*, and do not report co-variation with mood (American Psychiatric Association, DSM-5 Task Force, 2013).

2.2. Thermodynamic model and combined causal graph

Let us now turn to the thermodynamic model. The living human organism is assimilated to an open thermodynamic system (Hall, 2012). The four energetic variables (standard quantitative variables) of the organism are defined as follows:

1. *Energy Intake* represents thermodynamic energy content of carbohydrates, fats, and proteins in food;
2. *Energy Expenditure*, energy (heat) expanded by the organism when not performing physical work, and energy expanded after a meal, called thermal effect of feeding;
3. *Work*, energy expanded during physical work performed by the organism on the environment;
4. *Energy Store variation*, variation of the thermodynamic energy content of carbohydrate, fat, and protein stores of the organism.

According to the principle of the conservation of energy, we have the following thermodynamic balance equation:

$$\textit{Energy Store variation} = \textit{Energy Intake} - \textit{Energy Expenditure} - \textit{Work} \quad (2)$$

In order to make this model compatible with the causal psychological theory, we use causal relationships, instead of energy fluxes, to relate variables. Therefore, in causal terms, *Energy Store variation* receives a stimulation from *Energy Intake* and inhibitions from *Energy Expenditure* and from *Work*.

Then, physiological adaptation is added. It includes metabolic adaptation, also called diet-induced adaptive thermogenesis, which decreases *Energy Expenditure* through an increase of mitochondrial efficiency, when *Energy Store* is depleted, and behavioral adaptation, which increases *Energy Intake* through an increase in response to food-related cues, in the same conditions (Hall, 2012). *Energy Store* needs to be distinguished from its variation because the effect of both adaptations persists when a reduced *Energy Store* is maintained, i.e., when *Energy Store variation* is null (Rosenbaum et al., 2008). Metabolic adaptation is therefore represented by a stimulation from *Energy Store* toward *Energy Expenditure* (which also includes the decrease in *Energy Expenditure* due to muscle loss), and behavioral adaptation by an inhibition from *Energy Store* toward *Energy Intake*. This thermodynamic causal model is presented in Fig. 2A.

Two assumptions are made: (i) *Work* decrease due to protein and fat mass loss is compensated by an increase in physical activity so that *Work* is considered as constant, i.e., no stimulation is required from *Energy Store* toward *Work*; (ii) the thermal effect of feeding is considered as constant when *Energy Intake* varies, because it is by far the smallest component of *Energy Expenditure*, i.e., no stimulation is required from *Energy Intake* toward *Energy Expenditure* (Hall, 2012). Of note, the relationship between *Energy Store variation* and *Energy Store*, i.e., integral causality, is kept implicit.

Finally, two more assumptions are made in the theory, (iii) *Weight variation* can be assimilated to *Energy Store variation* (Hall, 2012), and (iv) *Food intake* can be assimilated to *Energy Intake*, a common practice in a weight loss diet.

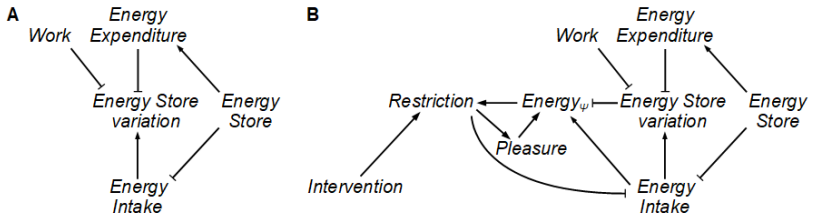

Figure 2: (A) Thermodynamic model and (B) combined theory.

So, the stimulation from *Food intake* toward *Weight variation* in Fig. 1B can be explained by the thermodynamic model and replaced by it, giving the combined theory which is presented in Fig. 2B. An ultimate assumption is made: (v) *Work* could be expected to influence $Energy_\psi$, but being constant, it would not induce $Energy_\psi$ fluctuations: therefore, this causality is not further considered here.

Causal relationships between two variables are interpreted as monotonic functions, strictly increasing in the case of stimulations, and strictly decreasing in the case of inhibitions, that we can think a priori as person-specific to represent inter-individual variability (in particular their intensities, i.e., magnitude ratios between the cause and the effect), but have the same monotonic property for all individuals. Whenever multiple causal relationships influence a given variable, a single function combining these multiple causal relationships must be devised, i.e., for those influencing $Energy_\psi$, *Energy Intake* and *Restriction*. The additivity of the individual effects will be assumed and therefore the sum function will be used. These comments do not apply to the three functions derived from the principle of the conservation of energy which are identical for all individuals, linear, and additive on the common target variable *Energy Store variation*.

The central role played by $Energy_\psi$ in the behavior change is enlightened by the combined theory: (i) $Energy_\psi$ value conditions dietary behavior change, i.e., when the value of $Energy_\psi$ is not above the given positive threshold labeled E_{min}, such behavior change is either not initiated or discontinued, and (ii) $Energy_\psi$ fluctuations are due to inter-related variables depending on such behavior change. It is therefore clear that our theory unfolds over time and formalizes what is actually a dynamical system (Katok and Hasselblatt, 1995).

2.3. Timescale and the three core dynamic phenomena

Statement (1) has the peculiarity to explain dynamic phenomena of dietary behavior change at a specific timescale that is given by a dynamic of weight loss in weeks induced by "such extreme diets". Note that most behavior-change theories identified by a multidisciplinary literature review are not dynamic, and among the few that are dynamic, none of them specify a timescale (Michie et al., 2014). Nevertheless, this specification might be critical because dynamic phenomena may be explained by distinct theories when they unfold e.g., over days vs. weeks or months (Zaheer et al., 1999).

Dynamic phenomena of severely hypocaloric dietary behavior change occur, in the present case of a supervised diet, during an intervention[2] period, which lasts twelve weeks, that is subdivided into an initiation period and a discontinuation period.

The specification of a timescale has been reviewed in organizational research and conceptualized with five types of time interval: recording, aggregation, and observation intervals associated with the engagement of the researcher, and existence and validity intervals associated with the phenomena under study (Zaheer et al., 1999). Even though interventions are not considered, these types of time interval can be applied to behavior-change theories. In our case, the recording interval, the interval over which a variable is measured, is 24 hours, food intake being commonly measured in kcal/24 hours. The aggregate interval, the interval over which measures are to be aggregated for theorizing, is here one week and provides the length of the initiation period. The observation interval, which is self-explanatory, is equal here to the length of the intervention period, i.e., twelve weeks (Jensen et al., 2014). The existence interval, the length of time needed for one instance of the phenomenon to occur, is equal here to the observation interval, but can be equal to a fraction of it when the existence interval is unknown or when more than one instance is to be studied.

Event-based dynamic phenomena of dietary behavior change are identified by means of two successive events, initiation and discontinuation, defined by a taxonomy of adherence to medications (Vrijens et al., 2012) and potentially occurring during a prescription period (our intervention period). These dynamic phenomena can then be exhaustively categorized into three core event-based phenomena:

$$\begin{aligned} &\text{Ph.1} \quad \textit{non-initiation} \\ &\text{Ph.2} \quad \textit{initiation followed by discontinuation} \\ &\text{Ph.3} \quad \textit{initiation followed by non-discontinuation} \end{aligned} \tag{3}$$

The selection of a timescale, i.e., the characterization of these three phenomena as single-timescale, requires that (i) the intervention period is subdivided into two periods, the first period with a duration corresponding to the smallest time interval considered for theorizing (that is how one week has been chosen) and (ii) initiation can only occur during the first period and discontinuation during the second period, non-occurrence of these events being established at the end of their respective period. Consequently, event-based dynamic phenomena of the shortest timescale, i.e., when discontinuation occurs during the first period instead of the second, and of the longest timescale, i.e., when initiation occurs during the second period instead of the first, are discarded. Only the three

[2]Interventions are defined as "coordinated sets of activities designed to change specified behavior patterns" (Michie et al., 2011); here, the intervention is conceptualized as the most basic type, "provide instruction: telling the person how to perform a behavior" (Abraham and Michie, 2008).

identified phenomena remain, corresponding respectively to: (1) no event during first and second period, (2) initiation during first period and discontinuation during second period, and (3) initiation during first period and no event during second period.

Now, the objective is to show that our theoretical model covers all three core phenomena (3). According to the model, we saw that $\mathit{Energy}_\psi \leq E_{min}$ is a sufficient condition of non-initiation or discontinuation of hypocaloric dietary behavior change. This sufficient condition is not necessary in general, as other processes than those described in the model can also produce non-initiation or discontinuation of such behavior change. But these other processes, such as socio-environmental and cognitive factors, operate at timescales which differ from the specific timescale considered here, with an existence interval shorter than one week or a few times longer than twelve weeks (think for example of Energy_ψ decrease throughout working days with recovery during the weekend, and throughout Fall, with recovery in Spring, with respective existence intervals of one week and one year). So, Energy_ψ value decreasing below E_{min}, or equal to it, is a condition which becomes necessary and sufficient when this condition is restricted to this specific timescale. With such a restriction, non-initiation or discontinuation of behavior change from one side, and initiation or non-discontinuation of behavior change from the other side, are therefore equivalent to $\mathit{Energy}_\psi \leq E_{min}$ and $\mathit{Energy}_\psi > E_{min}$, respectively.

3. Qualitative analysis of the causal model

We now turn to an analysis of the causal model derived in Section 2, with the aim of establishing its ability to reproduce the three core phenomena indicated by (3). To do this, we will first rely on the fact that the evolution profiles of the two energetic variables that exert a direct causal influence on Energy_ψ are known experimentally. This greatly simplifies the model, since only the three causal relationships having a direct effect on Energy_ψ in Fig. 2B need be taken into account. In particular, this avoids making predictions between time points, i.e., the dynamic model can be analyzed statically at each time slice. Next, we discretize the time and variables involved to obtain a qualitative abstraction of our model in the form of a finite integer-based model. As this is an over-approximation of our initial model, i.e., a less constrained model, its solutions contain all those of this initial model (plus, possibly, a few spurious ones). Finally, Answer-Set Programming will be used to solve the resulting combinatorial problem.

To simplify notation, we will hereafter refer to the variables Energy_ψ, *Energy Store*, *Energy Store variation*, *Energy Intake*, *Energy Expenditure*, *Work*, *Pleasure*, *Restriction* and *Intervention* as E_ψ, ES, dES, EI, EE, W, P, R, I respectively. With these notations, the causal graph in Fig. 2B becomes that of Fig. 4A.

3.1. Evolution profiles of the causal input variables of Energy$_\psi$

Consider the three variables which are the direct causal inputs of E_ψ, i.e., EI, dES, and P (see Fig. 4A). The evolution profiles of EI and ES (and consequently of dES) over time are known experimentally (see Fig. 3), from the evolution profiles of food intake and weight respectively, because the diet is supervised:

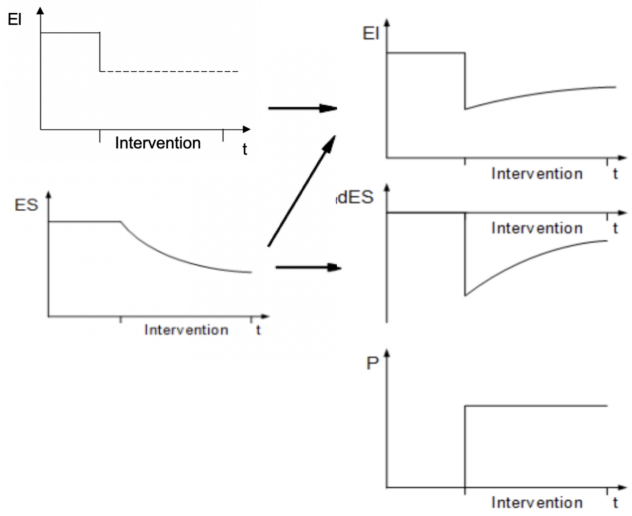

Figure 3: On the left, experimental evolution profiles over time of *Energy Intake* EI and *Energy Store* ES. On the right, the profile of EI derived from its profile on the left, corrected by behavioral adaptation (inhibition from ES), the profile of the variation dES derived from that of ES, and the theoretically hypothesized profile of *Pleasure* P (arrows indicate derivations).

- EI, which is constant during the basal period which precedes the intervention (we will call basal values the constant values of variables during this basal period), undergoes a sudden drop of more than 50% of its initial value at the beginning of the intervention, due to the severe reduction in food intake. EI should theoretically stay constant during all the intervention period following the diet instruction but actually, due to the behavioral adaptation (inhibition from ES to EI), it increases slightly with time (albeit less than 50% of the amount of the drop) during the intervention, following ES decrease.

- ES, which is constant during the basal period that precedes the intervention, decreases during the intervention period, but its value at the end of the intervention is greater than 50% of its basal value (the loss of weight is less than 50%). In addition this decrease is slower and slower, with equilibrium at the end of the intervention period. Actually, ES variation, i.e., the derivative variable dES, which is null during the basal period, undergoes a sudden drop at the beginning of the intervention (to compensate

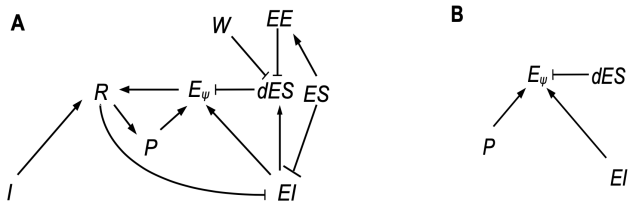

Figure 4: (A) The complete causal graph of the combined theory (renaming of Fig. 2B) is replaced by (B) the extracted sub-graph, together with the evolution profiles of Fig. 3 right.

the drop of EI, due to the energy conservation (2)) and then increases during the intervention period, until it returns to zero at the end of this period.

It is important to note that these profiles are qualitative (we do not rely on any numeric value of these variables) and, from observation, verified by all patients without exception (including the percentage ratios indicated which have been chosen broad enough to encompass all cases).

Finally, P, as the pleasure of food rejection, is by nature zero during the basal period and positive during the intervention period, and its evolution profile is theoretically hypothesized to be as simple as possible, i.e., a step function (see Fig. 3):

- P (pleasure of rejection of food) is null during the basal period and positive, assumed constant, during the intervention period.

3.2. The simplified model and its qualitative analysis

3.2.1. Simplified model

The fact that the evolution over time of E_ψ's three causal input variables is known greatly simplifies our model. Indeed, since the only thing that matters is the relative value over time of E_ψ with respect to E_{min}, which is entirely determined by the three direct causal influences exerted on it by EI, dES and P, and by the evolution over time of these three input variables, it is not necessary to simulate over time the entire causal network of Fig. 4A to compute this evolution, since it is known experimentally. We can thus get rid of causal loops in the thermodynamic part of the network (due to the integral causality from dES to ES) and use only the simplified model given by the sub-graph of Fig. 4B limited to the causal relationships of these three variables with E_ψ, which is sufficient to decide about initiation and about discontinuation of behavior change, given known evolution profiles of EI, dES and P over time.

So, the complete model given by Fig. 4A (together with the initial qualitative values of the variables in the basal period and of EI at the beginning of the intervention period) is replaced by the simplified model given by the very simple causal graph of Fig. 4B together with the qualitative evolution profiles over time of EI, dES and P of Fig. 3.

We will now proceed to the qualitative analysis of this simplified causal graph. Let's consider first an individual causal relationship from variable X to variable Y. This can be represented as a function F of X and *Time* t, i.e., $Y(t) = F(X(t), t)$, with a given sign, i.e., $\partial F/\partial X$ positive for a stimulation and negative for an inhibition. By analogy with a transfer function in Control Theory, if we consider its response to a step function, F is characterized by its propagation delay (time between the rising edge of the input and activation of the output), its settling time (time between activation of the output and stabilization at its equilibrium value) and its gain (ratio between the equilibrium value and the input step value). Considering that the recording interval is equal to 24 hours, we can, at this qualitative scale of time discretization, assume that the causal response is instantaneous (i.e., propagation delay and settling time are negligible) and thus F is independent of *Time* t: $Y(t) = F(X(t))$. Its gain (i.e., the intensity of the causality) is person-specific. Taken into account the additivity of the individual effects in a multiple causal relationship, the simplified model of Fig. 4B can thus be expressed as: $E_\psi(t) = F_{EI}(EI(t)) + F_{dES}(dES(t)) + F_P(P(t))$.

Now, due to the strict monotonicity of functions F_{EI}, F_{dES}, and F_P, functions $F_{EI}(EI(t))$, $-F_{dES}$ $(dES(t))$, and $F_P(P(t))$ have the same evolution profiles over time as the functions $EI(t)$, $dES(t)$, and $P(t)$ from the increase or decrease point of view. Nevertheless, we will not require that the additional qualitative constraints imposed on $EI(t)$ in terms of value ratios be preserved by F_{EI} as the causal functions may be non-linear. Therefore the qualitative evolution profiles of the three effects $F_{EI}(EI(t))$, $-F_{dES}(dES(t))$, and $F_P(P(t))$ are identical over time to those of the input variables in terms of monotonicity (but obviously not in terms of magnitude, which is person-specific). We will note these effects as $\overline{EI}(t)$, $\overline{dES}(t)$ and $\overline{P}(t)$, respectively, i.e., for any causal input variable X of E_ψ, $\overline{X}(t)$ is defined as the effect $F_X(X(t))$ on $E_\psi(t)$ of $X(t)$ if the causality is stimulation and its opposite $-F_X(X(t))$ if it is inhibition (i.e., we consider the positive component of the effect, and keep the sign of the causality separate), and $\overline{X}(t)$ has the same evolution profile over time as $X(t)$ in terms of monotonicity and discontinuity. Finally, we can write:

$$E_\psi(t) = \overline{EI}(t) - \overline{dES}(t) + \overline{P}(t) \tag{4}$$

where $\overline{EI}(t)$, $\overline{dES}(t)$, and $\overline{P}(t)$, now being the effects of the three input variables, have qualitative evolution profiles given by Fig. 3, with no constraint on their absolute and relative magnitudes (so, for our qualitative analysis, working with $\overline{X}(t)$ is the same as working with $X(t)$). In addition, we have the condition that the global causal effect (right-hand side of (4)) is greater than E_{min} during the basal period, which is expressed, since $dES(t)$ and $P(t)$ are both null during this period, by:

$$E_0 > E_{min} \quad \textit{where } E_0 = \overline{EI}(t) \textit{ for } t \textit{ in the basal period} \tag{5}$$

3.2.2. Discretization and qualitative analysis

Consider now the discretization of the variables in a semi-quantitative framework of qualitative reasoning (Kuipers, 1994) based on intervals arithmetic. Concerning *Time*, the most obvious choice is to discretize it in a finite number (which can be increased if needed) of time intervals of equal size, numbered by integers. But we can look at the coarsest discretization possible in order to limit the size of the solution space, thus the whole complexity. We saw that the time period is broken down into the basal period and the intervention period, the latter being subdivided into an initiation period and a discontinuation period. We can adopt a single discrete time value for the basal period, as all variables are constant there. The same thing can be done for the initiation period, which is short (one tenth of the whole intervention period) and for which only the sudden discontinuity of $EI(t)$, $dES(t)$, and $P(t)$ (and thus of their effects) at the beginning of this period qualitatively matters. Now, the qualitative trends of these three input variables (and thus of their effects) remain unchanged during the whole discontinuation period, i.e., $\overline{EI}(t)$ is increasing, $-\overline{dES}(t)$ decreasing, and $\overline{P}(t)$ constant. So, we can also adopt here a single discrete time value for this period, to test if $E_\psi(t) \leq E_{min}$ or $E_\psi(t) > E_{min}$. Finally, it is enough to discretize *Time* into three intervals corresponding to basal, initiation, and discontinuation periods, labeled arbitrarily by the integers -1, 0 and 1 (i.e., the intervention period begins at time 0), corresponding to $[-1, 0[$, $[0, 1[$, and $[1, 2[$, respectively. With such a coarse discretization of *Time*, it will obviously be impossible to express the continuity of all variables for $t \in [0, 2[$. In particular we cannot certify that a solution to $E_\psi(0) > E_{min}$ and $E_\psi(1) > E_{min}$ actually never satisfies $E_\psi(t) \leq E_{min}$ during the intervention period (i.e., for some real number $t \in [0, 2[$). But this is quite standard for any qualitative abstraction, which provides only an over-approximation, that is, a less constrained model, of the initial problem, therefore with more solutions, which means possible spurious solutions. In fact we will compare the results of this coarse discretization of *Time* with those of a fine-grain discretization in terms of a number of small time intervals, allowing the concept of continuity of the variables to be expressed, and show they are identical.

The common discretization of $\overline{EI}$, $\overline{dES}$, and $\overline{P}$ is as follows. An interval $[-n, n]$ is defined, which includes all possible values (in energy unit), for all patients, of $\overline{EI}(t)$, $\overline{dES}(t)$, and $\overline{P}(t)$ during basal and intervention periods (n is not known but this does not matter). This interval is divided into a certain number $2m$ (where m is a given integer parameter which can be chosen arbitrarily large) of sub-intervals of equal size n/m (plus interval [0]). Integer values from $-m$ to m are assigned to the endpoints of these sub-intervals (which means adopting n/m as new energy unit, which can always be done since the conditions $E_\psi(t) \leq E_{min}$ or $E_\psi(t) > E_{min}$, with $E_\psi(t)$ given by (4), are invariant by a same homothety on the value scale). Therefore, the qualitative values of $\overline{EI}(t)$, $\overline{dES}(t)$, and $\overline{P}(t)$, corresponding to some specific curves with evolution profiles given by Fig. 3, over a discretized *Time* interval $[T, T + 1[$, with $T \in \{-1, 0, 1\}$, are given by some integer intervals $[x^-, x^+]$, $[y^-, y^+]$, and $[z^-, z^+]$, respectively,

with $x^-, x^+, y^-, y^+, z^-, z^+ \in \{-m, \ldots, m\}$, that encompass the real continuous values of $\overline{EI}(t)$, $\overline{dES}(t)$, and $\overline{P}(t)$ along these curves for $t \in [T, T+1[$. As these curves are monotonic on each time interval $[T, T+1[$ from Fig. 3, we have: $x^-, x^+ \in \{\overline{EI}(T), \overline{EI}(T+1)\}$ and the analog for the two other variables. As the right-hand side of (4) is monotonic wrt. each of the three variables, its extreme values on $[T, T+1[$ are reached at T and $T+1$, so can be computed in terms of $x^-, x^+, y^-, y^+, z^-, z^+$ and are themselves integers. This means that, in order to check if $E_\psi(t) \leq E_{min}$ or $E_\psi(t) > E_{min}$ over the time intervals, only the qualitative values of the three variables at the time points $\{-1, 0, 1\}$ matter. Finally, all solutions in the discrete approximation (so, maybe some spurious ones, but none missing for the initial, non-approximated, model) will be obtained by assigning any integer between $-m$ and m to any of the three variables at any of the three time points such that all the constraints defining the evolution profiles in Fig. 3 are satisfied. This extends straightforwardly to the fine-grain discretization of *Time*. In this case, the continuity of the variables is naturally expressed in the discretized framework by restricting the evolution of the variables to stepwise changes at successive time points. Since each of the variables can take a maximum of m values (each with a constant sign), we can represent this restriction (i.e., require that two of its successive values differ by one at most) over its entire range of possible variation during the intervention period (where the variables vary monotonically in time) by adopting m as the number of discrete time points in this period. This means that $Time = \{-1, 0, 1, \ldots, m-1\}$ in the fine-grain discretization, with -1 always designating the basal period.

The qualitative analysis therefore changes the initial continuous problem into a simpler constraint based integer program where a (finite) exhaustive search can be performed. With the coarse discretization of *Time*, the size of the search space is in $O\left(m^9\right)$ (number of possible value assignments for each of the input variables to the power of the number of assignments, i.e., the number of variables times the number of time points). With the fine-grain *Time* discretization, the complexity increases a lot, as it is now in $O\left(m^{3(m+1)}\right)$.

3.2.3. Expressing the evolution profiles and the three core phenomena conditions

The integer-based constraints formalizing the evolution profiles, in terms of increase and decrease, as described in subsection 3.1, are the following in the case of the coarsest discretization of *Time* (we quantified the sudden drop of $\overline{dES}(t)$ and rise of $\overline{P}(t)$ at the beginning of the intervention period by at least one unit, which is the least restrictive possible; in the same way, the increase of $\overline{EI}(t)$ after the beginning of the intervention period as well as the difference between its final value at the end of this period and its basal value, are quantified by at least one unit, so the sudden drop of $\overline{EI}(t)$ at the beginning of the intervention period is at least two units):

$$\begin{aligned} \overline{EI}(-1) - 1 \geq \overline{EI}(1) \geq \overline{EI}(0) + 1 \\ \overline{dES}(-1) = \overline{dES}(1) = 0 \quad \overline{dES}(0) \leq -1 \end{aligned} \tag{6}$$

$$\overline{P}(-1) = 0 \quad \overline{P}(0) = \overline{P}(1) \geq 1$$

In the case of the fine-grain discretization of *Time* with $m+1$ time points ($-1, 0, m-1$ identical respectively to $-1, 0, 1$ in the coarsest discretization and $1, 2, \ldots, m-2$ discretizing the intervention period between 0 and 1 in the coarsest discretization), we add constraints to express the monotonicity of the variables and the restriction to stepwise changes at successive time points between 0 and $m-1$.

For what concerns E_ψ, we must focus on the conditions $E_\psi(t) \leq E_{min}$ and $E_\psi(t) > E_{min}$ which determine the three core phenomena. Taking into account that $E_0 = E_\psi(-1) > E_{min}$ (5), and denoting the positive range $E_0 - E_{min}$ by ΔE, these conditions can be rewritten as:

$$\begin{array}{ll} \textit{interruption:} & E_\psi(t) - E_\psi(-1) \leq -\Delta E \\ \textit{no interruption:} & E_\psi(t) - E_\psi(-1) > -\Delta E \end{array} \tag{7}$$

The value of ΔE is unknown but actually $E_\psi(t)$ has an upper physiological value E_{max} and thus varies physiologically between E_{min} and E_{max} and this range is not very wide. As $0 < \Delta E < E_{max} - E_{min}$, ΔE can be expressed as an interval between two small fractions of m, say integers ΔE^- and ΔE^+. So conditions of non-initiation or discontinuation from one side and of initiation or non-discontinuation from the other side at *Time* T can be expressed as:

$$(\overline{EI}(T) - \overline{dES}(T) + \overline{P}(T)) - (\overline{EI}(-1) - \overline{dES}(-1) + \overline{P}(-1)) \leq -\Delta E, \text{ and}$$
$$(\overline{EI}(T) - \overline{dES}(T) + \overline{P}(T)) - (\overline{EI}(-1) - \overline{dES}(-1) + \overline{P}(-1)) > -\Delta E$$

respectively and must be verified for all integer values $\Delta E \in \{\Delta E^-, \ldots, \Delta E^+\}$.

Finally, taken into account that $\overline{dES}(-1) = 0$ and $\overline{P}(-1) = 0$, and that $> -\Delta E$ is equivalent to $\geq 1 - \Delta E$ as all values are integers, the conditions of the three core phenomena, given by (3), are, for the coarse discretization of *Time* with three time points, as follows:

$$\begin{array}{ll} \text{Ph.1} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \leq -\Delta E \\ \text{Ph.2} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \geq 1 - \Delta E \\ & (\overline{EI}(1) - \overline{EI}(-1)) - \overline{dES}(1) + \overline{P}(1) \leq -\Delta E \\ \text{Ph.3} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \geq 1 - \Delta E \\ & (\overline{EI}(1) - \overline{EI}(-1)) - \overline{dES}(1) + \overline{P}(1) \geq 1 - \Delta E \end{array} \tag{8}$$

Furthermore, from (5), since $E_0 = \overline{EI}(-1) = \Delta E + E_{min}$, we necessarily have:

$$\overline{EI}(-1) \geq 1 + \Delta E \tag{9}$$

Note that we are looking for the existence of a solution for any of the three phenomena Ph.i, i $\in \{1, 2, 3\}$ and any integer value $\Delta E \in \{\Delta E^-, \ldots, \Delta E^+\}$. We can also consider a stronger condition by requiring, for any Ph.i, the existence

of a solution valid for all integer values ΔE. This is expressed by the following stronger constraints:

$$\begin{array}{ll}
\text{Ph.1} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \leq -\Delta E^{+} \\
\text{Ph.2} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \geq 1 - \Delta E^{-} \\
 & (\overline{EI}(1) - \overline{EI}(-1)) - \overline{dES}(1) + \overline{P}(1) \leq -\Delta E^{+} \\
\text{Ph.3} & (\overline{EI}(0) - \overline{EI}(-1)) - \overline{dES}(0) + \overline{P}(0) \geq 1 - \Delta E^{-} \\
 & (\overline{EI}(1) - \overline{EI}(-1)) - \overline{dES}(1) + \overline{P}(1) \geq 1 - \Delta E^{-}
\end{array} \tag{10}$$

and:

$$\overline{EI}(-1) \geq 1 + \Delta E^{+} \tag{11}$$

For the fine-grain discretization of *Time* with m time points in the intervention period, the conditions of the core phenomena can be refined: discontinuation (second constraint of Ph.2) is expressed by

$$\exists T \in \{1, \ldots, m-1\} \quad (\overline{EI}(T) - \overline{EI}(-1)) - \overline{dES}(T) + \overline{P}(T) \leq -\Delta E \tag{12}$$

and non-discontinuation (second constraint of Ph.3) by

$$\forall T \in \{1, \ldots, m-1\} \quad (\overline{EI}(T) - \overline{EI}(-1)) - \overline{dES}(T) + \overline{P}(T) \geq 1 - \Delta E \tag{13}$$

In the same way as above, the stronger conditions can be used.

3.2.4. Assessing the model capacity of faithful reproduction of the three core phenomena

Assessing our simplified discrete model capacity of a faithful reproduction of the three core phenomena means checking that it owns solutions, in terms of the nine (resp., $3(m+1)$) variable values taken into $\{0, 1, \ldots, m\}$ that represent $\overline{EI}(T)$, $-\overline{dES}(T)$ and $\overline{P}(T)$ for $T \in \{-1, 0, 1\}$ (resp., $T \in \{-1, 0, \ldots, m-1\}$), that are consistent with constraints (6) representing the evolution profiles of the input variables (or their effects) and with any of the three core phenomena conditions (8) (or the stronger ones (10)), whatever it is. This is an integer-based constraint satisfaction problem, whose complete processing chain can be synthetically expressed by Fig. 5.

Constraints expressing the evolution profiles (6) and any of the three core phenomena (8) (or (10)) define a polyhedral cone in $\mathbb{R}^9$ (or $\mathbb{R}^{3(m+1)}$). Now a cone is non-empty (which means the existence of a solution) if and only if it contains a point with integer coordinates (which means the existence of an integer solution). In other words, our constraint satisfaction problem has a solution (in real numbers) if and only if it has an integer solution, ensuring that a sufficiently large search space is considered (a ball centered on the origin with a radius large enough to contain the smallest integer point in the cone, if there is one). This shows that we are sure to obtain the right answer about the existence or not of a real solution to our set of qualitative constraints by our integer-based computing by choosing a large enough m.

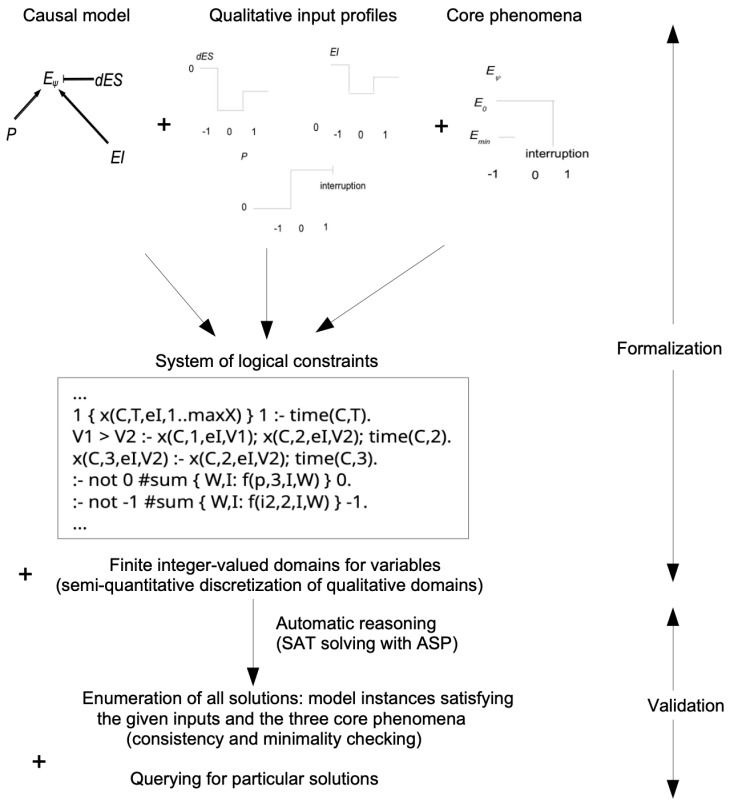

Figure 5: Processing chain.

Discretization being an over-approximation of the continuous problem (with real-based curves for the evolution profiles), the nonexistence of a discrete solution would prove that the model cannot reproduce all three core phenomena and has to be rejected. But, as spurious solutions may well exist, it is in general impossible to conclude the validity of the model from the existence of a discrete solution. This is particularly the case with the coarse discretization of *Time* with just three time points. However, when using the fine-grain discretization of *Time* with m time points for the intervention period, then all discretized variables, including *Time*, approximate more and more closely the corresponding real variables when m increases, i.e., (6) get closer and closer to the real curves of Fig. 3, and (8) (or (10)) get closer and closer to the phenomena (3). We can conclude that, for a large enough m, any integer solution of our set of qualitative constraints is arbitrarily close to a real solution of our abstract model. Nevertheless, the fact that such a solution could be spurious in reality cannot be completely ruled out, due to the over-approximation of our model itself, in particular of the evolution profiles of the input variables that might be submitted in reality to stronger constraints not represented in our model that

could falsify the solution found. That is why we will consider and solve in Section 5 an under-approximation of the problem that does not use these evolution profiles and will compare the results.

3.3. Results and minimality checking

We used ASP (Answer-Set Programming) (Baral, 2003; Gebser et al., 2012; Lifschitz, 2019), a simple and powerful modeling language to solve combinatorial problems described as logic programs, and more specifically *clingo* (Gebser et al., 2018; Potassco), to implement our qualitative analysis. This language is widely used in bioinformatics, particularly in symbolic systems biology, it allows concise programming close to the problem specifications and is very efficient at enumerating all solutions, which will be useful in Section 4. This is why we have preferred ASP to integer linear programming, which could be another choice.

An extended version of the ASP code will be presented in detail in Section 4 (see Listing 1). The present code is obtained by fixing the signs of the causalities exerted on E_ψ, i.e., by adding the following to the code in Listing 1: `csgn(ei,1).`, `csgn(des,-1).`, `csgn(p,1).`, `csgn(es,0).`, `csgn(ee,0).`. The code response is *Satisfiable*, which shows that our causal model of Fig. 4B, together with constraints (6) and any of the three core phenomena (8) (or (10)), is satisfiable (and this result is obtained as well with *Time* discretization with three time points as with $m + 1$ time points), which proves the adequacy of our simplified causal model, in that it reproduces any of the three core phenomena, under the assumption that our abstract representation of the evolution profiles of the input variables is consistent with reality.

We can also study the minimality of this model in terms of the causal relationships involved, i.e., if this causal graph could not be simplified by removing a certain causal relationship while remaining a solution, i.e., reproducing the three core phenomena. We successively remove one of the three causal relationships in the graph of Fig. 4B (i.e., by writing successively in the code `csgn(ei,0).`, `csgn(des,0).`, `csgn(p,0).`). The results show that the program is still *Satisfiable* with `csgn(p,0).`, which proves that the model remains a solution when removing the stimulation of E_ψ by P (i.e., by simply removing the variable P as it has no other causal influence), which proves that it is not minimal. This can be easily understood from a purely mathematical point of view: the effect on E_ψ of the stimulation by P, which is a positive (constant) integer from $T = 0$, can be simulated by merging it with the effect of the stimulation by EI (if the negative drop of $\overline{EI}(0)$ is big enough with regards to the positive drop of $\overline{P}(0)$, the latter can be absorbed by the first and the evolution profile of $\overline{EI}$ is still admissible). Now, if we look at the evolution profiles in Fig. 3, we see that, if $dES(t)$ restores its basal value, i.e., 0, at the end of the intervention period, it is not the case for $EI(t)$, whose increase during the intervention period is less than half its drop at the beginning of this period. Therefore, in the absence of P, and in the case of the phenomenon Ph.3 of initiation and non-discontinuation, the basal value of E_ψ cannot be restored at the end of the intervention period, where it is necessarily smaller (between E_{min} and E_0). In this case, it can be considered preferable to ensure that the mood level of the patient at the end of

this successful diet has returned to its initial value before the intervention, i.e., to require the satisfaction of the constraint:

$$(\overline{EI}(1) - \overline{EI}(-1)) - \overline{dES}(1) + \overline{P}(1) = 0 \tag{14}$$

(with 1 replaced by $m - 1$ with the fine-grain discretization of *Time*) instead of the second constraint of Ph.3 in (8) and (10). With this stronger constraint (lines 71-72 of the code in Listing 1), the result of the automatic analysis is that our causal model of Fig. 4B. is now a minimal solution.

4. Minimal solutions when allowing any causal relationships

Up to now, we have considered statement (1) as the only valid proto-theory, that we have formalized as the causal graph of Fig. 4A, and we have shown that the extracted sub-graph of Fig. 4B, together with the qualitative evolution profiles of the three input variables of Fig. 3, satisfies any of the three core phenomena (8) (or (10)) and is minimal if we require restoration of the basal mood level at the end of the intervention period, if reached (14). Now, from our causal graph of Fig. 4A, we see that, in all generality, five variables can be considered as direct causal input variables for the psychological variable E_ψ: in addition to the two thermodynamic variables EI and dES and the psychological variable P considered so far because our first objective was to build a formal theory based on statement (1) and to check its ability to reproduce the three core phenomena), the thermodynamic variables EE and ES can also be taken into account (and in fact must be in view of our second objective, which is to investigate whether other causal networks, in terms of direct influences exerted on E_ψ, could also reproduce the three core phenomena, and to enumerate them all, if there are any). This means we are now dealing with the general extracted causal graph of Fig. 6, where each of the five causal relationships may be a stimulation, an inhibition, or absent, i.e., is parameterized by a sign in $\{-1, 0, 1\}$.

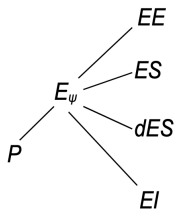

Figure 6: The general extracted causal graph, with signs of the causalities taken arbitrarily in $\{-1, 0, 1\}$ (inhibition, inactive, stimulation, respectively).

The qualitative evolution profile of ES (and thus of its effect $\overline{ES}$) is known (see Fig. 3 bottom left) and, as EE is submitted to a single causal relationship, i.e., a stimulation by ES, we can reasonably assume that EE (and therefore

$\overline{EE}$) has the same qualitative evolution profile as ES:

$$\begin{aligned} &\overline{ES}(-1) \geq \overline{ES}(0) \geq \overline{ES}(-1) - 1 \qquad \overline{ES}(-1) > \overline{ES}(1) \\ &\overline{EE}(-1) \geq \overline{EE}(0) \geq \overline{EE}(-1) - 1 \qquad \overline{EE}(-1) > \overline{EE}(1) \end{aligned} \tag{15}$$

Equation (4) is generalized by the following:

$$E_\psi(t) = s_{EI}\overline{EI}(t) + s_{dES}\overline{dES}(t) + s_{EE}\overline{EE}(t) + s_{ES}\overline{ES}(t) + s_P\overline{P}(t) \tag{16}$$

where s_X' are the signs of the causalities of X' to E_ψ, taken in $\{-1, 0, 1\}$, not all zero, and the evolution profiles of $\overline{EI}(t), \overline{dES}(t), \overline{ES}(t), \overline{EE}(t)$, and $\overline{P}(t)$ are given by (6) and (15).

Finally, the constraints (8) or (10) expressing the conditions of the three core phenomena are generalized using (7) and (16). For example, the strong constraints are given by:

$$\begin{aligned} \text{Ph.1} \quad & s_{EI}(\overline{EI}(0) - \overline{EI}(-1)) + s_{dES}\overline{dES}(0) + s_{EE}(\overline{EE}(0) - \overline{EE}(-1)) \\ & + s_{ES}(\overline{ES}(0) - \overline{ES}(-1)) + s_P\overline{P}(0) \leq -\Delta E^+ \\ \text{Ph.2} \quad & s_{EI}(\overline{EI}(0) - \overline{EI}(-1)) + s_{dES}\overline{dES}(0) + s_{EE}(\overline{EE}(0) - \overline{EE}(-1)) \\ & + s_{ES}(\overline{ES}(0) - \overline{ES}(-1)) + s_P\overline{P}(0) \geq 1 - \Delta E^- \\ & s_{EI}(\overline{EI}(1) - \overline{EI}(-1)) + s_{dES}\overline{dES}(1) + s_{EE}(\overline{EE}(1) - \overline{EE}(-1)) \\ & + s_{ES}(\overline{ES}(1) - \overline{ES}(-1)) + s_P\overline{P}(1) \leq -\Delta E^+ \\ \text{Ph.3} \quad & s_{EI}(\overline{EI}(0) - \overline{EI}(-1)) + s_{dES}\overline{dES}(0) + s_{EE}(\overline{EE}(0) - \overline{EE}(-1)) \\ & + s_{ES}(\overline{ES}(0) - \overline{ES}(-1)) + s_P\overline{P}(0) \geq 1 - \Delta E^- \\ & s_{EI}(\overline{EI}(1) - \overline{EI}(-1)) + s_{dES}\overline{dES}(1) + s_{EE}(\overline{EE}(1) - \overline{EE}(-1)) \\ & + s_{ES}(\overline{ES}(1) - \overline{ES}(-1)) + s_P\overline{P}(1) \geq 1 - \Delta E^- \end{aligned} \tag{17}$$

Similarly, the strong constraint (11) is generalized by:

$$s_{EI}\overline{EI}(-1) + s_{EE}\overline{EE}(-1) + s_{ES}\overline{ES}(-1) \geq 1 + \Delta E^+ \tag{18}$$

Finally, the restoration of the basal value of E_ψ at the end of the intervention period in the case of phenomenon Ph.3, is obtained by replacing the second constraint of Ph.3 above by the generalization of (14), i.e.:

$$\begin{aligned} s_{EI}(\overline{EI}(1) - \overline{EI}(-1)) + s_{dES}\overline{dES}(1) + s_{EE}(\overline{EE}(1) - \overline{EE}(-1)) \\ + s_{ES}(\overline{ES}(1) - \overline{ES}(-1)) + s_P\overline{P}(1) = 0 \end{aligned} \tag{19}$$

The ASP code is given in Listing 1 in its version with strong constraints (10) and three time points (it is available, along with the version with $m+1$ time points at `https://github.com/phdague/eating-behavior-change`). In this code, the variables EI, dES, ES, EE, P should be interpreted as $\overline{EI}, \overline{dES}, \overline{ES}, \overline{EE}, \overline{P}$ respectively. Lines 2,4 and 6-7 define the parameter values, here 21 for m (in

fact, stability of the solution set is observed for smaller values of m), 4 for ΔE^-, 7 for ΔE^+ and 1 for the minimum absolute value of discontinuities (which is the least constrained value possible). Line 10 lists the five possible causal input variables for E_ψ and line 12 stipulates that the sign of each causality is a single integer taken from $\{-1, 0, 1\}$. The aim of the program is to find all solution assignments for these signs. Lines 15-17 assert the temporal extent of each of the three core phenomena (non-initiation (`ni`), initiation followed by discontinuation (`id`) and initiation followed by non-discontinuation (`ind`)) along the three time points -1 (basal period), 0 (initiation period), 1 (discontinuation period). The main predicate is `v(Ph,T,I,V)` which takes the value $True$ when `V` is the value at time point `T` of the input variable `I` in the phenomenon `Ph` scenario. Lines 22-25, 28-30, 33-36, 39-42, and 45-47 express the evolution profiles of variables $\overline{EI}, \overline{dES}, \overline{ES}, \overline{EE}$, and $\overline{P}$ respectively, given by (6) and (15). Line 50 applies the sign of the causality to each value of a variable $\overline{X}$ to obtain the effect of X on E_ψ, and stores the result in the predicate `eff`. According to (16), the sum of the values obtained is therefore equal to E_ψ. Lines 54-55, 57-60, and 62-65 express the three core phenomena Ph.1, Ph.2, and Ph.3 respectively and implement (17) (as a goal `:-` is interpreted as asserting negation, the presence of `not` inside the goal is globally equivalent to an assertion). Lines 67-68 implement the constraint (18) concerning the overall causal effect during the basal period. Lines 71-72 implement the restoration of the basal value of E_ψ in phenomenon Ph.3, given by (19). Line 76 requires at least two causalities to be inactive, i.e., at most three to be active, in order to enumerate only minimal models (in the case where lines 71-72 are enabled; otherwise, it is sufficient to consider at most two to be active, i.e., we replace 2 by 3 in line 76). Finally, line 79 displays all possible sign assignments to the causalities in the solutions found, i.e., these are projected onto the five sign variables only. Note that after this, we prove that there is no other minimal model, by obtaining the unsatisfiability of the program to which we have added the sign negations of the active causalities in every minimal solution obtained so far.

```
1  % maximal absolute value of variables
2          #const m=21.
3  % minimal absolute value of discontinuity of EI, dES, and P
4          #const disc = 1.
5  % bounds of the positive gap between E0 and Emin
6          deltaem(m/5).
7          deltaep(m/3).
8
9  % the five possible causal input variables to psychological energy
10         cause((ei;des;es;ee;p)).
11 % signs of the causalities
12         1{csgn(I,(-1;0;1))}1 :- cause(I).
13 % the three core phenomena and their temporal extent
14 % (-1 for basal, 0 for initiation, 1 for discontinuation periods)
15         time(ni,-1..0).
16         time(id,-1..1).
17         time(ind,-1..1).
18
```

```
% evolution profile of EI > 0 with negative discontinuity of
% at least disc units between -1 and 0 then incresae of at least
% one unit without reaching its basal value
        1 { v(Ph,T,ei,1..m) } 1 :- time(Ph,T).
        V2 <= V1-disc :- v(Ph,-1,ei,V1); v(Ph,0,ei,V2).
        V3 > V2 :- v(Ph,0,ei,V2); v(Ph,1,ei,V3).
        V3 < V1 :- v(Ph,-1,ei,V1); v(Ph,1,ei,V3).
% evolution profile of dES <= 0 with negative discontinuity of
% at least disc units between -1 and 0, zero in -1 and in 1
        v(Ph,-1,des,0) :- time(Ph,-1).
        1 { v(Ph,0,des,-m..-disc) } 1 :- time(Ph,0).
        v(Ph,1,des,0) :- time(Ph,1).
% evolution profile of ES > 0 decreasing by at least one unit with
% a variation of at most one unit (continuity) between -1 and 0
        1 { v(Ph,T,es,1..m) } 1 :- time(Ph,T).
        V1 >= V2  :- v(Ph,-1,es,V1); v(Ph,0,es,V2).
        V2 >= V1-1  :- v(Ph,-1,es,V1); v(Ph,0,es,V2).
        V1 > V3 :- v(Ph,-1,es,V1); v(Ph,1,es,V3).
% evolution profile of EE > 0 decreasing by at least one unit with
% a variation of at most one unit (continuity) between -1 and 0
        1 { v(Ph,T,ee,1..m) } 1 :- time(Ph,T).
        V1 >= V2  :- v(Ph,-1,ee,V1); v(Ph,0,ee,V2).
        V2 >= V1-1   :- v(Ph,-1,ee,V1); v(Ph,0,ee,V2).
        V1 > V3 :- v(Ph,-1,ee,V1); v(Ph,1,ee,V3).
% evolution profile of P >= 0, zero in -1, with discontinuity
% of at least disc units between -1 and 0
        v(Ph,-1,p,0) :- time(Ph,-1).
        v(Ph,T,p,V) :- time(Ph,T); T >= 0; val(p,Ph,V).
        1 { val(p,Ph, disc..m) } 1 :- time(Ph,_).

% application of the signs of the causalities to get the effect
        eff(Ph,T,I,V*W) :- v(Ph,T,I,V); csgn(I,W).

% the three core phenomena
% Ph.1: non-initiation (ni)
        :- deltaep(G);
   not #sum { W,I,0: eff(ni,0,I,W); -W,I,-1: eff(ni,-1,I,W) } -G.
% Ph.2: initiation followed by discontinuation (id)
        :- deltaem(G);
   not 1-G #sum { W,I,0: eff(id,0,I,W); -W,I,-1: eff(id,-1,I,W)}.
        :- deltaep(G);
   not #sum { W,I,1 : eff(id,1,I,W); -W,I,-1: eff(id,-1,I,W)} -G.
% Ph.3: initiation followed by non-discontinuation (ind)
        :- deltaem(G);
   not 1-G #sum { W,I,0: eff(ind,0,I,W); -W,I,-1: eff(ind,-1,I,W)}.
        :- deltaem(G);
   not 1-G #sum { W,I,1: eff(ind,1,I,W); -W,I,-1: eff(ind,-1,I,W)}.
% the global causal effect on E at -1 is greater than deltaep
        :- time(Ph,_); deltaep(G);
   not G+1 #sum { W,I: eff(Ph,-1,I,W) }.

% restoration at 1 of the basal value of E
        :- not
   #sum { W,I,1: eff(ind,1,I,W); -W,I,1: eff(ind,-1,I,W)} = 0.

% minimality: at most three active causalities
% (change 2 by 3 if restoration disabled)
```

```
76          :- not 2 #count { I : csgn(I,0) }.
77
78 % display of the signs of the causalities
79          #show csgn/2.
```

Listing 1: ASP code with three time points (code3.asp)

We look for all the minimal models in terms of the signs of the causalities (for the order $0 < 1$, $0 < -1$), i.e., among $3^5 - 1 = 242$ potential candidates. Note that we want to enumerate all the projections of the solutions onto the five variables representing the signs of causality, not just count them. The results are given in Listing 2 (restoring the basal value of E_ψ at the end of the Ph.3 intervention period is not required) and Listing 3 (restoring the basal value of E_ψ at the end of the Ph.3 intervention period is required).

This enumeration is the main bottleneck in terms of computation time. With m the number of qualitative values in the quantity space of each of the five causal input variables, p ($1 \leq p \leq 5$) a given maximum number of active causalities in the causal sign assignments whose satisfiability is checked (p is taken equal to 2 in the case where the basal value of E_ψ is not restored at the end, where all minimal models are found to have two active causalities, and equal to 3 in the case where the basal value of E_ψ is restored, where all minimal models are found to have three active causalities), and q the number of time points, complexity is in $O\left(m^{pq}\right)$. And therefore, in $O\left(m^{5q}\right)$ to determine afterwards that there are no other minimal solutions. This is why it is important to design a coarse-grained qualitative model, with small values for its parameters, which we have achieved with $q = 3$ time points, allowing a not too high overall complexity in $O\left(m^{15}\right)$, which remains manageable by ASP even with a value as large as 21 for m (see Listings 2 and 3). In fact, we observe that the execution time is perfectly admissible, less than a minute in total for the version with restoration of the basal mood value (40s to enumerate solutions with at most three active causalities, and 15s to show that there is no other minimal solution). The key was therefore to be able to model the three core phenomena using just three time points, representing the basal, initiation, and discontinuation periods. The comparison with our other, more refined, qualitative over-approximation with $q = m+1$ time points speaks for itself. The complexity increases to $O\left(m^{3(m+1)}\right)$ to enumerate all minimal solutions with three active causalities (with $m = 21$, the execution time is around two hours with the restoration of the basal mood value) and reaches $O\left(m^{5(m+1)}\right)$ to prove there are no other minimal solutions, which is not achievable by the program with $m = 21$.

```
./clingo-5.2.2-macos-10.9/clingo --project 0  code3.asp
clingo version 5.2.2
Reading from test1.asp
Solving...
Answer: 1
csgn(ei,1) csgn(des,0) csgn(es,0) csgn(ee,1) csgn(p,0)
Answer: 2
csgn(ei,0) csgn(des,1) csgn(es,0) csgn(ee,1) csgn(p,0)
Answer: 3
csgn(ei,1) csgn(des,-1) csgn(es,0) csgn(ee,0) csgn(p,0)
```

```
Answer: 4
csgn(ei,1) csgn(des,0) csgn(es,1) csgn(ee,0) csgn(p,0)
Answer: 5
csgn(ei,0) csgn(des,1) csgn(es,1) csgn(ee,0) csgn(p,0)
Answer: 6
csgn(ei,0) csgn(des,0) csgn(es,1) csgn(ee,0) csgn(p,-1)
Answer: 7
csgn(ei,0) csgn(des,0) csgn(es,0) csgn(ee,1) csgn(p,-1)
SATISFIABLE
Models       : 7
Calls        : 1
Time         : 0.173s (Solving: 0.15s 1st Model: 0.01s Unsat: 0.01s)
CPU Time     : 0.173s
```

Listing 2: Results without (lines 71-72 disabled) restoration of the basal mood value at the end of the intervention period (with $m+1$ time points CPU Time is 19s)

Without requiring (19), i.e., the restoration of the basal state mood value at the end of the intervention for phenomenon Ph.3, the automatic analysis provides the seven following minimal solutions (the same with (17) or the weak form generalizing (8) and with three or $m+1$ time points), all with exactly two active causal relationships (see Listing 2):

$$\begin{array}{lll} (m1) & EI \longrightarrow E_\psi & dES \longrightarrow\!\!\!\!\!\!\mid\ E_\psi \\ (m2) & EI \longrightarrow E_\psi & EE \longrightarrow E_\psi \\ (m3) & EI \longrightarrow E_\psi & ES \longrightarrow E_\psi \\ (m4) & P \longrightarrow\!\!\!\!\!\!\mid\ E_\psi & EE \longrightarrow E_\psi \\ (m5) & P \longrightarrow\!\!\!\!\!\!\mid\ E_\psi & ES \longrightarrow E_\psi \\ (m6) & dES \longrightarrow E_\psi & EE \longrightarrow E_\psi \\ (m7) & dES \longrightarrow E_\psi & ES \longrightarrow E_\psi \end{array} \tag{20}$$

$m1$ is the minimal model found in subsection 3.3. The others can be grouped in twos, as EE and ES, having identical qualitative evolution profiles, play the same role. Each of these new solutions can a posteriori be roughly qualitatively explained: $m2/m3$ are derived from $m1$, as ES has an evolution profile similar to that of $-dES$, except the discontinuity at $T = 0^-$ of the latter but it is not required for the initiation; $m4/m5$ are derived from $m2/m3$ as EI and $-P$ have similar evolution profiles; $m6/m7$ are derived from $m2/m3$ as EI and dES have similar evolution profiles (see Fig. 3).

```
./clingo-5.2.2-macos-10.9/clingo --project 0  code3.asp
clingo version 5.2.2
Reading from test2.asp
Solving...
Answer: 1
csgn(ei,-1) csgn(des,1) csgn(es,1) csgn(ee,0) csgn(p,0)
Answer: 2
csgn(ei,-1) csgn(des,1) csgn(es,0) csgn(ee,1) csgn(p,0)
Answer: 3
csgn(ei,0) csgn(des,1) csgn(es,-1) csgn(ee,1) csgn(p,0)
Answer: 4
csgn(ei,0) csgn(des,1) csgn(es,0) csgn(ee,1) csgn(p,1)
Answer: 5
```

```
csgn(ei,0) csgn(des,1) csgn(es,1) csgn(ee,0) csgn(p,1)
Answer: 6
csgn(ei,0) csgn(des,1) csgn(es,1) csgn(ee,-1) csgn(p,0)
Answer: 7
csgn(ei,1) csgn(des,-1) csgn(es,0) csgn(ee,0) csgn(p,1)
Answer: 8
csgn(ei,1) csgn(des,-1) csgn(es,0) csgn(ee,-1) csgn(p,0)
Answer: 9
csgn(ei,1) csgn(des,-1) csgn(es,-1) csgn(ee,0) csgn(p,0)
Answer: 10
csgn(ei,0) csgn(des,0) csgn(es,1) csgn(ee,-1) csgn(p,-1)
Answer: 11
csgn(ei,-1) csgn(des,0) csgn(es,1) csgn(ee,0) csgn(p,-1)
Answer: 12
csgn(ei,1) csgn(des,0) csgn(es,1) csgn(ee,0) csgn(p,1)
Answer: 13
csgn(ei,1) csgn(des,0) csgn(es,1) csgn(ee,-1) csgn(p,0)
Answer: 14
csgn(ei,1) csgn(des,0) csgn(es,-1) csgn(ee,1) csgn(p,0)
Answer: 15
csgn(ei,1) csgn(des,0) csgn(es,0) csgn(ee,1) csgn(p,1)
Answer: 16
csgn(ei,-1) csgn(des,0) csgn(es,0) csgn(ee,1) csgn(p,-1)
Answer: 17
csgn(ei,0) csgn(des,0) csgn(es,-1) csgn(ee,1) csgn(p,-1)
SATISFIABLE
Models       : 17
Calls        : 1
Time         : 40.371s (Solving: 40.36s 1st Model: 0.00s Unsat: 4.53s)
CPU Time     : 40.361s
```

Listing 3: Results with restoration of the basal mood value at the end of the intervention period (with $m+1$ time points CPU Time is 7381s)

With the stronger constraint (19), requiring the restoration of the basal mood value at the end of the intervention for phenomenon Ph.3, the automatic analysis provides the seventeen following minimal solutions (the same using (17) or the weak form generalizing (8) and using three or $m+1$ time points), all with exactly three active causal relationships (see Listing 3):

$$
\begin{array}{lccc}
(M1) & P \longrightarrow E_\psi & EI \longrightarrow E_\psi & dES \longrightarrow\!\!\!\!\mid\ E_\psi \\
(M2) & P \longrightarrow\!\!\!\!\mid\ E_\psi & EI \longrightarrow\!\!\!\!\mid\ E_\psi & EE \longrightarrow E_\psi \\
(M3) & P \longrightarrow\!\!\!\!\mid\ E_\psi & EI \longrightarrow\!\!\!\!\mid\ E_\psi & ES \longrightarrow E_\psi \\
(M4) & P \longrightarrow E_\psi & EI \longrightarrow E_\psi & EE \longrightarrow E_\psi \\
(M5) & P \longrightarrow E_\psi & EI \longrightarrow E_\psi & ES \longrightarrow E_\psi \\
(M6) & P \longrightarrow E_\psi & dES \longrightarrow E_\psi & EE \longrightarrow E_\psi \\
(M7) & P \longrightarrow E_\psi & dES \longrightarrow E_\psi & ES \longrightarrow E_\psi \\
(M8) & EI \longrightarrow E_\psi & EE \longrightarrow E_\psi & ES \longrightarrow\!\!\!\!\mid\ E_\psi \\
(M9) & EI \longrightarrow E_\psi & EE \longrightarrow\!\!\!\!\mid\ E_\psi & ES \longrightarrow E_\psi \\
(M10) & EI \longrightarrow E_\psi & dES \longrightarrow\!\!\!\!\mid\ E_\psi & EE \longrightarrow\!\!\!\!\mid\ E_\psi \\
(M11) & EI \longrightarrow E_\psi & dES \longrightarrow\!\!\!\!\mid\ E_\psi & ES \longrightarrow\!\!\!\!\mid\ E_\psi \\
(M12) & EI \longrightarrow\!\!\!\!\mid\ E_\psi & dES \longrightarrow E_\psi & EE \longrightarrow E_\psi \\
(M13) & EI \longrightarrow\!\!\!\!\mid\ E_\psi & dES \longrightarrow E_\psi & ES \longrightarrow E_\psi
\end{array}
\tag{21}
$$

$$
\begin{array}{lccc}
(M14) & dES \longrightarrow E_\psi & EE \longrightarrow\!\!\!\!\mid\, E_\psi & ES \longrightarrow E_\psi \\
(M15) & dES \longrightarrow E_\psi & EE \longrightarrow E_\psi & ES \longrightarrow\!\!\!\!\mid\, E_\psi \\
(M16) & P \longrightarrow\!\!\!\!\mid\, E_\psi & EE \longrightarrow\!\!\!\!\mid\, E_\psi & ES \longrightarrow E_\psi \\
(M17) & P \longrightarrow\!\!\!\!\mid\, E_\psi & EE \longrightarrow E_\psi & ES \longrightarrow\!\!\!\!\mid\, E_\psi
\end{array}
$$

$M1$ is the model formalized from statement (1) throughout this paper. The others can again be grouped in twos, as EE and ES play the same role, i.e., we obtain eight new couples of models. Looking at the qualitative evolution profiles of Fig. 3, we can empirically convince ourselves that each of these models verifies any of the three core phenomena. Now, it is up to the experts, after examination, to provide reasons to rule out if necessary some or all of these extra couples of models. Nevertheless, some general causality principles can be applied. As dES is the derivative of ES, it appears natural to exclude models with an active causality on E_ψ of both of these variables. And, as EE is directly caused by ES with no other causal influence, to also exclude an active causality of both of these variables and consequently of EE and dES as well, i.e., finally to authorize only one active input causality from $\{dES, EE, ES\}$. After these exclusions, only two extra couples of models in addition to $M1$ remain, namely $M2/M3$ and $M4/M5$ (accordingly, after the same exclusions applied to (20), only the models $m1$, $m2/m3$ and $m4/m5$ remain). It is worthwhile noting that all five remaining models require an active causality from the psychological variable P in addition to causalities from EI and another of the three remaining energetic variables. $M4/M5$ is derived from $M1$ by replacing the inhibition from dES by a stimulation from EE or ES, which is easy to understand, and a priori acceptable, from the energy balance $-dES = EE - EI + W_0$. $M2/M3$ derives from $M4/M5$ by inverting the signs of the causalities of P and EI (inhibitions instead of stimulations), which can be explained mathematically but is in conflict with the observed influences of *Pleasure* and *Energy Intake* on mood.

5. Analytical resolution of the model

In whet follows, we will give a direct proof that the previous minimal solutions found by the qualitative analysis of our discretized model (thus an over-approximation of our original causal model including all its solutions and possibly others) and our integer programming, are valid, non-spurious, solutions of our original model. To do this, we linearize our model, i.e., we impose that the causal relations be affine functions. The model obtained is obviously more constrained than the original model (in particular, there is no a priori reason for causal influences to be linear in reality, non-linearities being frequent in biology, and this was not assumed in our qualitative analysis), i.e., it is an under-approximation of it, whose solutions are guaranteed to be solutions of the original model, but some of which might be missing. The resulting parameterized symbolic model (the unknown coefficients of the linear causalities and the thresholds of the psychological variables are treated as parameters) is expressed as a system of ordinary differential equations that are simple enough to

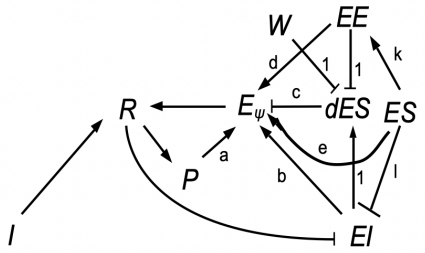

Figure 7: The complete causal graph with all possible causal relationships exerted on E_ψ. Causalities are assumed to be linear and their intensities are displayed on the corresponding arcs.

be solved analytically by hand, from the sole knowledge of the initial conditions at $t = 0$ (unlike our qualitative analysis, we do not rely here on the experimental evolution profiles of EI and ES given in Fig. 3, which will in fact be derived as solutions of the differential equations). Then, from these solutions, a direct analysis provides all the minimal solutions in terms of active causal relationships. We will then see that these minimal solutions are exactly those found by our qualitative analysis. As they are respectively a subset and a superset of the set of minimal solutions of our original model, this will prove that the latter is equal to both. Finally, a parametric analysis determines which parameters need to be person-specific and which parameters can be assumed constant, thus shedding light on inter-individual variability.

5.1. Equations underlying the causal model

Under linearity assumption of the causal relationships and with the simplified energetic model (in particular, work assumed constant), the causal model is given in Fig. 7 and underlain by the following equations (the causal relationships, parameterized in superscript by the linear coefficient representing their intensity, from where each equation is derived, is written in comments below the equation):

$$E_\psi(t) - E_0 = b(EI(t) - EI_0) - c\frac{dES}{dt}(t) + d(EE(t) - EE_0) + e(ES(t) - ES_0) + aP(t) \tag{22}$$
$$/^*\ EI \longrightarrow^b E_\psi \quad dES \longrightarrow\!\!\!\!\mid^c E_\psi \quad EE \longrightarrow^d E_\psi \quad ES \longrightarrow^e E_\psi \quad P \longrightarrow^a E_\psi$$

$$\frac{dES}{dt}(t) = EI(t) - EE(t) - W(t) \tag{23}$$
$$/^*\ EI \longrightarrow^1 dES \quad EE \longrightarrow\!\!\!\!\mid^1 dES \quad W \longrightarrow\!\!\!\!\mid^1 dES$$

$$EI(t) - EI_0 = -l(ES(t) - ES_0) - R(t) \tag{24}$$
$$/^*\ ES \longrightarrow\!\!\!\!\mid^l EI \quad R \longrightarrow\!\!\!\!\mid^1 EI$$

$$EE(t) - EE_0 = k(ES(t) - ES_0) \tag{25}$$

$$/^* \; ES \longrightarrow^k EE$$

$$W(t) = W_0 \tag{26}$$

$$/^* \; W \text{ is assumed constant}$$

$$R(t) = \text{if } E_\psi > E_{min} \text{ then } I(t) \text{ else } 0 \tag{27}$$

$$/^* \; I \longrightarrow^1 R \quad E_\psi \longrightarrow R \text{ acts as a switch}$$

$$\text{(if } E_\psi > E_{min} \text{ no effect, else forces } R \text{ to } 0)$$

$$P(t) = sign(R(t))P_0 \tag{28}$$

$$/^* \; R \longrightarrow P \text{ binary } (0 \text{ if } R = 0, P_0 \text{ if } R > 0)$$

$$I(t) = \text{if } t \geq 0 \text{ then } I_0 \text{ else } 0 \tag{29}$$

$$/^* \text{ the intervention is constant for } t \geq 0$$

where $EE_0, ES_0, EI_0, E_0, E_{min}, P_0, W_0, I_0$ are positive constants.

(22) expresses the causal effect on E_ψ as an affine function of all possible causal variables (according to the extracted causal graph of Fig. 6), where a, b, c, d, e are constant parameters (positive, negative or zero, as we assume neither the existence nor the sign of these causalities). This is the linearized form of (16), i.e., the effect $\overline{X}(t)$ of the causal input variable $X(t)$ is given by $\overline{X}(t) = q_X X(t) + r_X$ where the coefficient q_X is positive and we have $a = s_P q_P$, $b = s_{EI} q_{EI}$, $c = -s_{dES} q_{dES}$ (as we have kept as default sign an inhibition from dES as in the initial model of Fig. 4B), $d = s_{EE} q_{EE}$ and $e = s_{ES} q_{ES}$. (23) is the thermodynamic balance equation (2), (24) and (25) express the causalities of ES on EI and EE as affine functions, where k, l are positive constant parameters because the signs of these causalities are fixed by thermodynamics. Note that only the counterpart (16) of (22) has been taken into account in our qualitative model, as well as the qualitative evolution profiles of the five causal input variables (6,15), replaced here by equations (23,24,25).

We assume the equilibrium (i.e., null derivatives) during the basal period ($t < 0$) with the following constant values for the variables: $I(t) = R(t) = P(t) = 0$, $EI(t) = EI_0$, $ES(t) = ES_0$, $EE(t) = EE_0$, and $E_\psi(t) = E_0$, with $E_0 > E_{min}$. The equations above are satisfied by this basal state with the following constraint between constants:

$$EI_0 - EE_0 = W_0 \tag{30}$$

Since $E_\psi(t)$ is entirely created by the combined effects of the five causal input variables, it is natural to assign it a zero value when the values of the five variables are zero, which amounts to saying that equation (22) is linear, i.e., its constant term $E_0 - bEI_0 - dEE_0$ - eES_0 is zero. This means that the basal value E_0 of $E_\psi(t)$ is equal to the global effect $bEI_0 + dEE_0 + eES_0$ of the three causalities active in this basal state. As E_0 is positive, so is this overall effect (stated as (18) in the qualitative case):

$$bEI_0 + dEE_0 + eES_0 > 0 \tag{31}$$

which implies that at least one of the parameters b, d, e is positive.

We will now see that the above system of ordinary differential equations can be solved analytically from this initial basal state. This means we do not rely on the evolution profiles given by Fig. 3 as we did for the qualitative analysis, but rather we compute analytically these profiles as solutions of the system of ODEs with the basal state as the initial state, in the case of affine causal relationships. Note that the discretized qualitative model we constructed in Sections 3 and 4 as an over-approximation of our initial model is not the discretization of the present ODE system constructed as an under-approximation of that initial model. In particular, our discretized model does not assume linear causal relations, which is the prerequisite for obtaining linear ODEs that are easy to solve. And an over-approximation of an under-approximation does not in general provide an over-approximation of the initial problem.

5.2. Continuity and discontinuities

Context. The exogenous interruption $I(t)$ is constant (≥ 0) during each phase: $I(t) = 0$ during the basal phase ($t < 0$), $I(t) = I_0 > 0$ during the strong restriction phase ($0 \leq t$). Therefore all variables are continuous (even derivable) during each phase and the only possible discontinuities are in 0^-.
Assumption. We assume that $ES(t)$ is continuous in 0^- (and then $EE(t)$ too by (25)). Indeed, it is natural to assume that thermodynamic variables are by essence continuous, the only discontinuities being exogenous and imposed (case of the discontinuity of $EI(t)$ resulting from the imposed discontinuity of $I(t)$). In fact, continuous variables can accommodate exogenous discontinuities if their derivatives can be discontinuous in 0^- (which is typically the case of $\frac{dES}{dt}(t)$).

In $t = 0$, $I(t)$ undergoes a discontinuity with positive value I_0 (29), thus $R(t)$ undergoes the discontinuity I_0 (27), $P(t)$ the discontinuity P_0 (28), $EI(t)$ the discontinuity $-I_0$ (24), $\frac{dES}{dt}(t)$ the discontinuity $-I_0$ (23), and thus $E_\psi(t)$ the discontinuity $(c-b)I_0+aP_0$ (22). This means that: $E_\psi(0)-E_0 = (c-b)I_0+aP_0$.

There is thus interruption in $t = 0$ (then non-initiation) if and only if:

$$(b - c)I_0 - aP_0 \geq E_0 - E_{min}$$

Note that this requires $b > c$ or $a < 0$ and depends neither on d nor on e.

We will consider from now on the behavior on $[0, +\infty[$ when there is no interruption in $t = 0$, that is if $E_\psi(0) > E_{min}$.

5.3. Possible equilibrium values

Equilibrium values, denoted $X^=$ for a variable X, are obtained by setting the derivatives equal to 0 in the equations and by solving the algebraic system obtained (we will see later that equilibrium occurs asymptotically for $t \to +\infty$).

$$EI^= - EE^= = W_0$$
$$EI^= - EI_0 = -l(ES^= - ES_0) - I_0$$

$$EE^{=} - EE_0 = k(ES^{=} - ES_0)$$
$$E_{\psi}^{=} - E_0 = b(EI^{=} - EI_0) + d(EE^{=} - EE_0) + e(ES^{=} - ES_0) + aP_0$$

It follows that:

$$\begin{aligned} ES^{=} - ES_0 &= -\frac{1}{k+l} I_0 \\ EI^{=} - EI_0 = EE^{=} - EE_0 &= -\frac{k}{k+l} I_0 \\ E_{\psi}^{=} - E_0 &= -\frac{kb + kd + e}{k+l} I_0 + aP_0 \end{aligned} \tag{32}$$

Interruption being equivalent to $E_{\psi} \leq E_{min}$, a sufficient condition of interruption is thus $E_{\psi}^{=} < E_{min}$ (strict inequality as equilibrium is asymptotically reached), i.e.:

$$\frac{kb + kd + e}{k+l} I_0 - aP_0 > E_0 - E_{min}$$

which requires in particular: $aP_0 < \dfrac{kb + kd + e}{k+l} I_0$. We will see, by computing $ES(t), EI(t), EE(t)$ by the equations (23,24,25), that $E_{\psi}(t)$ is strictly monotonic or constant on $[0, +\infty[$, thus necessarily strictly decreasing in case of interruption. This implies that this condition is also necessary.

5.4. Solving the equations

From (23,24,25), taking into account (26,27,29,30), we obtain: $\dfrac{dES}{dt} = -(k+l)(ES(t) - ES_0) - I_0$, and thus:

$$\begin{aligned} ES(t) - ES_0 &= Ae^{-(k+l)t} - \frac{1}{k+l} I_0 \\ EE(t) - EE_0 &= kAe^{-(k+l)t} - \frac{k}{k+l} I_0 \\ EI(t) - EI_0 &= -lAe^{-(k+l)t} - \frac{k}{k+l} I_0 \\ E_{\psi}(t) - E_0 &= -KAe^{-(k+l)t} - \frac{kb + kd + e}{k+l} I_0 + aP_0 \\ &\text{with} \quad K = lb - kc - lc - kd - e \end{aligned}$$

where A is a constant. So, as long as $E_{\psi}(t) > E_{min}$, $ES(t), EE(t), EI(t)$ and $E_{\psi}(t)$ are all four equal to the exponential function $e^{-(k+l)t}$ with multiplicative coefficients $A, kA, -lA$ and $-KA$ respectively and with additive constants that represent the limit values when $t \to +\infty$, equal to the equilibrium values computed previously. We conclude that $ES(t), EE(t), EI(t)$ and $E_{\psi}(t)$ are monotonic and that $ES(t)$ and $EE(t)$ vary in the same direction and $ES(t)$ and

that $EI(t)$ vary in opposite directions.

The assumed continuity of $ES(t)$ in $t = 0$ provides: $A = \frac{1}{k+l} I_0$. As $A > 0$, we obtain: $ES(t)$ and $EE(t)$ decrease, $EI(t)$ increases, $E_\psi(t)$ increases if $K > 0$, decreases if $K < 0$ and is constant if $K = 0$.

Finally, we obtain, as long as $E_\psi(t) > E_{min}$:

$$
\begin{aligned}
ES(t) - ES_0 &= -\frac{I_0}{k+l}\left(1 - e^{-(k+l)t}\right) \\
\frac{dES}{dt}(t) &= -I_0 e^{-(k+l)t} \\
EE(t) - EE_0 &= -\frac{kI_0}{k+l}\left(1 - e^{-(k+l)t}\right) \\
EI(t) - EI_0 &= -\frac{I_0}{k+l}\left(le^{-(k+l)t} + k\right) \\
E_\psi(t) - E_0 &= \frac{KI_0}{k+l}\left(1 - e^{-(k+l)t}\right) + (c-b)I_0 + aP_0
\end{aligned}
$$

5.5. Qualitative behavior of the solutions

The variations of the different variables for $t \geq 0$ are as follows.

- $ES(t)$ decreases, starting from ES_0 in $t = 0$ with a slope equal to $-I_0$, and tends asymptotically towards $ES_0 - \frac{I_0}{k+l}$ when $t \to +\infty$. Note that, in terms of monotonicity, it agrees with the experimental evolution profile of ES given in Fig. 3. The additional qualitative constraint set in subsection 3.1, and not retained in our qualitative analysis, namely $ES_0 - ES^{=} < 0.5ES_0$, can be imposed on the profile of $ES(t)$, as we are dealing with an under-approximation, and is expressed by: $\frac{2I_0}{k+l} < ES_0$.

  $\frac{dES}{dt}(t)$ undergoes a negative discontinuity in $t = 0^-$, from 0 to $-I_0$, then increases with a slope equal to $(k+l)I_0$ at $t = 0$ and tends asymptotically towards 0 when $t \to +\infty$, which agrees with the experimental evolution profile of dES given in Fig. 3.

- $EE(t)$ decreases, starting from EE_0 in $t = 0$ with a slope equal to $-kI_0$, and tends asymptotically towards $EE_0 - \frac{kI_0}{k+l}$ when $t \to +\infty$. The additional qualitative constraint that we impose on the profile of $EE(t)$ is similar to that on $ES(t)$, i.e., $EE_0 - EE^{=} < 0.5EE_0$, and is expressed by: $\frac{2kI_0}{k+l} < EE_0$.

- $EI(t)$ undergoes a negative discontinuity in $t = 0^-$, from EI_0 to $EI_0 - I_0$, then increases with a slope equal to lI_0 at $t = 0$ and tends asymptot-

ically towards $EI_0 - \frac{kI_0}{k+l}$ when $t \to +\infty$. Again it agrees, in terms of monotonicity, with the experimental evolution profile of EI given in Fig. 3.The additional qualitative constraints set in subsection 3.1, and not retained in our qualitative analysis, namely $EI_0 - I_0 < 0.5EI_0$ and $EI^{=} - (EI_0 - I_0) < 0.5I_0$, can be imposed on the profile of $EI(t)$ and are expressed by: $2I_0 > EI_0$ and $k > l$.

In total, all the additional qualitative constraints set in subsection 3.1 and imposed on the profiles of the thermodynamic variables are expressed by:

$$k > l \quad \text{and} \quad EI_0 < 2I_0 < \frac{k+l}{k} min(EE_0, kES_0) \tag{33}$$

- $E_\psi(t)$ undergoes a discontinuity in $t = 0^-$ (except if $(b-c)I_0 = aP_0$, in which case it is continuous), from E_0 to $E_\psi(0) = E_0 + (c-b)I_0 + aP_0$.

  If $(b-c)I_0 - aP_0 \geq E_0 - E_{min}$, then there is interruption in $t = 0$ (and thus non-initiation).

  Else $E_\psi(t)$ has a slope equal to KI_0 in $t = 0$ and:

  - increases from $E_\psi(0)$ if $K > 0$, and tends asymptotically towards $E_\psi(0) + \frac{K}{k+l} I_0$ when $t \to +\infty$,
  - is constant equal to $E_\psi(0)$ if $K = 0$,
  - decreases from $E_\psi(0)$ if $K < 0$ and tends asymptotically towards $E_\psi(0) + \frac{K}{k+l} I_0$ when $t \to +\infty$ if $\frac{-K}{k+l} I_0 \leq E_\psi(0) - E_{min}$, i.e., $\frac{kb + kd + e}{k+l} I_0 - aP_0 \leq E_0 - E_{min}$, or reaches E_{min} with interruption if $\frac{-K}{k+l} I_0 > E_\psi(0) - E_{min}$, i.e., $\frac{kb + kd + e}{k+l} I_0 - aP_0 > E_0 - E_{min}$.

These different behaviors of $E_\psi(t)$ are depicted in Fig. 8.

5.6. The three core phenomena of dietary behavior change

1. There is interruption in $t = 0$ (and thus non-initiation) if and only if:

$$(b-c)I_0 - aP_0 \geq E_0 - E_{min} \tag{34}$$

2. There is interruption in $t > 0$ if and only if:

$$E_0 - E_{min} + \frac{K}{k+l} I_0 < (b-c)I_0 - aP_0 < E_0 - E_{min} \tag{35}$$

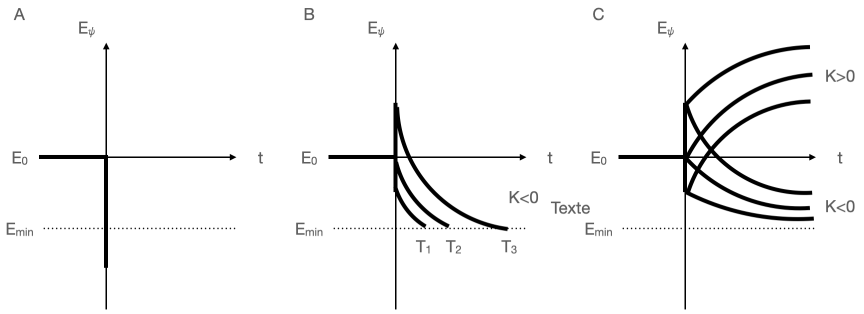

Figure 8: The three core phenomena according to the different possible behaviors of E_ψ: (A) interruption in $t = 0$ and therefore non-initiation; (B) interruption in $t = T > 0$ and therefore either non-initiation if $T < 1$ week or initiation followed by discontinuation if $1 \leq T \leq 12$ weeks or initiation followed by non-discontinuation if $T > 12$ weeks; (C) no interruption and therefore initiation followed by non-discontinuation.

which requires $K < 0$. The interruption occurs at positive time T given by:

$$T = \frac{-1}{k+l} Log\left(1 + \frac{k+l}{KI_0}(E_0 - E_{min} - (b-c)I_0 + aP_0)\right) \tag{36}$$

and is characterized as non-initiation, initiation followed by discontinuation or initiation followed by non-discontinuation depending on whether T is smaller than one week, between one and twelve weeks or greater than twelve weeks, respectively.

3. There is no interruption (and thus initiation followed by non-discontinuation) if and only if:

$$\begin{aligned} (b-c)I_0 - aP_0 &< E_0 - E_{min} \quad and \\ (b-c)I_0 - aP_0 &\leq E_0 - E_{min} + \frac{K}{k+l} I_0 \end{aligned} \tag{37}$$

The three possible behaviors of E_ψ and the related three core phenomena are depicted in Fig. 8.

5.7. Minimal models of (35)

We note that the phenomenon Ph.2 of initiation followed by discontinuation is only covered by the second case above, i.e., constraint (35). And also that this case covers the two other phenomena Ph.1 and Ph.3. Thus, to get a model of the three phenomena it is enough to get a model of the second case, i.e., satisfying (35). We are thus interested in computing the minimal causal models of constraints (23–33) that satisfy constraint (35), i.e. the interruption at some positive time. A causal model is given by signs of the causalities exerted on

E and provided by the values in $\{-, 0, +\}$ of the signs of a, b, c, d, e. Minimal models are defined with respect to the order relationship: $0 < -$ and $0 < +$.

The satisfaction of (35) imposes $K < 0$, thus such a model has to satisfy: $lb-(k+l)c-kd-e < 0$, i.e., $[b] \ominus [c] \ominus [d] \ominus [e] = -$ in the sign algebra (where $[x]$ denotes the sign of x and $\oplus$, $\ominus$ the sign addition and the sign subtraction, respectively).

1. *One active causality*
   Among the four minimal models (i.e., only one of the parameters b, c, d, e not null and a null) of $K < 0$, only two satisfy also (31), i.e., $[b] \oplus [d] \oplus [e] = +$, namely $d > 0$ and $e > 0$, which provide, concerning the satisfaction of (35):

   (a) $d > 0, a = b = c = e = 0$. $0 < E_0 - E_{min} < \dfrac{kd}{k+l} I_0$ is satisfiable for appropriate values of d.

   (b) $e > 0, a = b = c = d = 0$. $0 < E_0 - E_{min} < \dfrac{e}{k+l} I_0$ is satisfiable for appropriate values of e.

   We thus obtain two minimal models with only one active causality on E_ψ:

$$(M2.1) \quad d > 0, \quad EE \longrightarrow^{d} E_\psi \tag{38}$$

$$\text{with } \frac{kdI_0}{k+l} > E_0 - E_{min} \text{ and } E_\psi \text{ and } T \text{ given by:}$$

$$E_\psi(t) - E_0 = \frac{-kdI_0}{k+l}\left(1 - e^{-(k+l)t}\right),$$

$$T = \frac{-1}{k+l} Log\left(1 - \frac{k+l}{kdI_0}(E_0 - E_{min})\right)$$

$$(M2.2) \quad e > 0, \quad ES \longrightarrow^{e} E_\psi \tag{39}$$

$$\text{with } \frac{eI_0}{k+l} > E_0 - E_{min} \text{ and } EE_\psi \text{ and } T \text{ given by:}$$

$$E_\psi(t) - E_0 = \frac{-eI_0}{k+l}(1 - e^{-(k+l)t}),$$

$$T = \frac{-1}{k+l} Log\left(1 - \frac{k+l}{eI_0}(E_0 - E_{min})\right)$$

   We check that in both cases T may take any positive value, depending on the value of d or e.
   As we look for only minimal models, it is sufficient in the following to restrict ourselves to the case where $d, e \leq 0$.

2. *Two active causalities*
   With $d, e \leq 0$, (31) imposes $b > 0$. Constraint $K < 0$ thus provides $c > 0$. Let's check the possibility of a model with only two active causalities. This means that $b, c > 0$ and $a = d = e = 0$. Constraint (35) becomes:

$(b-c)I_0 < E_0 - E_{min} < \frac{bk}{k+l} I_0$ which is satisfiable for appropriate positive values of b, c with $\frac{b}{c} < 1 + \frac{k}{l}$. We obtain the unique following model with two active causalities:

$$(M2.3) \quad b, c > 0 \quad EI \longrightarrow^{b} E_\psi \quad dES \longrightarrow\!\!\!\!\mid^{c} E_\psi \tag{40}$$

with $(b-c)I_0 < E_0 - E_{min} < \frac{bk}{k+l} I_0$ and E_ψ and T given by:

$$E_\psi(t) - E_0 = \left(\frac{l}{k+l} b - c \right) I_0 \left(1 - e^{-(k+l)t} \right) + (c-b) I_0,$$

$$T = \frac{-1}{k+l} Log \left(1 - \frac{k+l}{((k+l)c - lb) I_0} (E_0 - E_{min} - (b-c) I_0) \right)$$

We check that T may take any positive value, depending on the values of b and c.

This model is thus the only minimal one with $d, e \leq 0$, since if we take some of a, d, e that are not null we would produce a model containing this one, thus non-minimal.

In total, there are thus three minimal causal models of (35), given by: $M2.1$, $M2.2, M2.3$ (38,39,40).

5.8. Minimal models of (34,35,37)

Though a model of (35), i.e., interruption in $t > 0$ (Fig. 8B) covers the three core phenomena Ph.1, Ph.2 and Ph.3, according to the value of T with respect to the thresholds $t = 1$ and $t = 12$ (in weeks), it is more appropriate to cover all possible behaviors of E_ψ, that is also the case where non-initiation (Ph.1) is due to interruption in $t = 0$ (Fig. 8A) and the case where initiation followed by non-discontinuation (Ph.3) is due to no interruption (Fig. 8C). Equivalently, this is as qualitatively abstracting the numerical thresholds of 1 and 12 weeks into $t = 0$ and $t = +\infty$ respectively and consider that actually (34) is prototypical of the phenomenon Ph.1 of non-initiation, (35) is prototypical of the phenomenon Ph.2 of initiation followed by discontinuation and (37) is prototypical of the phenomenon Ph.3 of initiation followed by non-discontinuation. We are thus interested in computing the minimal causal models of constraints (23–33) that satisfy constraints (34,35,37) together (i.e., with different values of the parameters a, b, c, d, e but the same signs for these values).

There are three minimal models of constraint (34) alone: $(b-c)I_0 - aP_0 \geq E_0 - E_{min}$, i.e., $[b] \ominus [c] \ominus [a] = +$, all with only one active causality: $a < 0; b > 0; c < 0$. Taking into account the satisfaction of constraint (31), we obtain the five following minimal causal models of (34):

$$(M1.1) \quad b > 0 \tag{41}$$

$$(M1.2) \quad a < 0, d > 0 \tag{42}$$

$$(M1.3) \quad a < 0, e > 0 \tag{43}$$
$$(M1.4) \quad c < 0, d > 0 \tag{44}$$
$$(M1.5) \quad c < 0, e > 0 \tag{45}$$

The minimal causal model of constraint (37) alone: $(b-c)I_0 - aP_0 < E_0 - E_{min}$ and $(b-c)I_0 - aP_0 \leq E_0 - E_{min} + \frac{K}{k+l}I_0$ is the empty model (no active causality). Taking into account the satisfaction of constraint (31), we thus obtain the three following minimal causal models of (37):

$$(M3.1) \quad b > 0 \tag{46}$$
$$(M3.2) \quad d > 0 \tag{47}$$
$$(M3.3) \quad e > 0 \tag{48}$$

Each of the three minimal causal models of (35), i.e., $M2.1(38)$, $M2.2(39)$, $M2.3(40)$ subsumes a minimal causal model of (37), namely $M3.2(47)$, $M3.3$ (48), $M3.1(46)$, respectively. We are therefore left to combine the minimal causal models of (34) and of (35).

The minimal causal model $M2.3(40)$ of (35) subsumes the model $M1.1(41)$ of (34) and the minimal causal models $M1.2(42)$, $M1.3(43)$, $M1.4(44)$, $M1.5$ (45) of (34) subsume the models $M2.1(38)$, $M2.2(39)$, $M2.1(38)$, $M2.2(39)$ of (35), respectively. These five models are thus minimal causal models of (34) and (35). We are therefore left to combine $M1.1(41)$ with $M2.1(38)$ and with $M2.2(39)$, which gives a sixth and a seventh minimal causal model of (34) and (35):

$$(M1.1.2.1) \quad b, d > 0 \tag{49}$$
$$(M1.1.2.2) \quad b, e > 0 \tag{50}$$

In total, there are thus seven minimal causal models of (34,35,37), i.e., of the three qualitative core phenomena, all with two active causalities, given by (40,49,50,42,43,44,45), that we rename $m1$ to $m7$ (and we set $a' = -a$ and $c' = -c$):

$(m1) \quad b, c > 0 \quad EI \longrightarrow^{b} E_\psi \quad dES \longrightarrow\!\!\!\!\mid^{c} E_\psi$

$$E_\psi(t) - E_0 = \left(\frac{lb}{k+l} - c\right) I_0 \left(1 - e^{-(k+l)t}\right) + (c-b)I_0$$

$(m2) \quad b, d > 0 \quad EI \longrightarrow^{b} E_\psi \quad EE \longrightarrow^{d} E_\psi$

$$E_\psi(t) - E_0 = \left(\frac{lb - kd}{k+l}\right) I_0 \left(1 - e^{-(k+l)t}\right) - bI_0$$

$(m3) \quad b, e > 0 \quad EI \longrightarrow^{b} E_\psi \quad ES \longrightarrow^{e} E_\psi$

$$E_\psi(t) - E_0 = \left(\frac{lb-e}{k+l}\right) I_0 \left(1 - e^{-(k+l)t}\right) - bI_0$$

$(m4)$ $\quad a < 0, d > 0 \quad P \longmapsto\!\!\!\!\mid^{a'} E_\psi \quad EE \longrightarrow^{d} E_\psi$

$$E_\psi(t) - E_0 = \frac{-kd}{k+l} I_0 \left(1 - e^{-(k+l)t}\right) - a'P_0 \tag{51}$$

$(m5)$ $\quad a < 0, e > 0 \quad P \longmapsto\!\!\!\!\mid^{a'} E_\psi \quad ES \longrightarrow^{e} E_\psi$

$$E_\psi(t) - E_0 = \frac{-e}{k+l} I_0 \left(1 - e^{-(k+l)t}\right) - a'P_0$$

$(m6)$ $\quad c < 0, d > 0 \quad dES \longrightarrow^{c'} E_\psi \quad EE \longrightarrow^{d} E_\psi$

$$E_\psi(t) - E_0 = \left(\frac{kd}{k+l} - c'\right) I_0 e^{-(k+l)t} - \frac{kd}{k+l} I_0$$

$(m7)$ $\quad c < 0, e > 0 \quad dES \longrightarrow^{c'} E_\psi \quad ES \longrightarrow^{e} E_\psi$

$$E_\psi(t) - E_0 = \left(\frac{e}{k+l} - c'\right) I_0 e^{-(k+l)t} - \frac{e}{k+l} I_0$$

These seven models, that are consistent with additional constraints (33), are exactly those (20) found by the qualitative analysis and its programming in ASP (with three or $m+1$ time points). In consequence, they are exactly the minimal solutions of our original problem.

5.9. Minimal models of (34,35,37) with restoration of the initial level of E_ψ in the case of (37)

It appears natural to impose additional constraint (14), namely that, for the phenomenon Ph.3 of initiation followed by non-discontinuation (37), the value $E_\psi^=$ of the psychological energy $E_\psi(t)$ at the final equilibrium (for $t = +\infty$) be equal to its initial value during the basal period, i.e.,

$$E_\psi^= = E_0 \tag{52}$$

From (32), this constraint can be rewritten as:

$$\frac{kb + kd + e}{k+l} I_0 = aP_0 \tag{53}$$

which is translated in the sign algebra into: $[b] \oplus [d] \oplus [e] = [a]$.

We see that none of the previous seven minimal causal models ($m1$–7) subsists (they actually all satisfy $E_\psi^= < E_0$).

Let's compute the minimal causal models of (37-53), as those of (34) and (35) are obviously unchanged. The three minimal causal models of (37), i.e., (46,47,48), give rise, in order to satisfy (53), to nine minimal causal models of

(37-53), all with two active causalities:

$$
\begin{aligned}
&(M'3.1) \quad a, b > 0 && (54)\\
&(M'3.2) \quad a, d > 0 && (55)\\
&(M'3.3) \quad a, e > 0 && (56)\\
&(M'3.4) \quad b > 0, d < 0 && (57)\\
&(M'3.5) \quad b > 0, e < 0 && (58)\\
&(M'3.6) \quad d > 0, b < 0 && (59)\\
&(M'3.7) \quad d > 0, e < 0 && (60)\\
&(M'3.8) \quad e > 0, b < 0 && (61)\\
&(M'3.9) \quad e > 0, d < 0 && (62)
\end{aligned}
$$

Finally, by combining the minimal causal models of (34), given by (41,42,43, 44,45), of (35), given by (38,39,40), and of (37-53), given by (54,55,56,57,58,59, 60,61,62), we obtain the following seventeen minimal causal models of (34,35,37-53), all with three active causalities:

$$
\begin{array}{llllll}
(M1) & a,b,c>0 & P \longrightarrow^{a} E_\psi & EI \longrightarrow^{b} E_\psi & dES \longdashv^{c} E_\psi & \\
(M2) & a,b<0,d>0 & P \longdashv^{a'} E_\psi & EI \longdashv^{b'} E_\psi & EE \longrightarrow^{d} E_\psi & \\
(M3) & a,b<0,e>0 & P \longdashv^{a'} E_\psi & EI \longdashv^{b'} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M4) & a,b,d>0 & P \longrightarrow^{a} E_\psi & EI \longrightarrow^{b} E_\psi & EE \longrightarrow^{d} E_\psi & \\
(M5) & a,b,e>0 & P \longrightarrow^{a} E_\psi & EI \longrightarrow^{b} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M6) & a,d>0,c<0 & P \longrightarrow^{a} E_\psi & dES \longrightarrow^{c'} E_\psi & EE \longrightarrow^{d} E_\psi & \\
(M7) & a,e>0,c<0 & P \longrightarrow^{a} E_\psi & dES \longrightarrow^{c'} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M8) & b,d>0,e<0 & EI \longrightarrow^{b} E_\psi & EE \longrightarrow^{d} E_\psi & ES \longdashv^{e'} E_\psi & \\
(M9) & b,e>0,d<0 & EI \longrightarrow^{b} E_\psi & EE \longdashv^{d'} E_\psi & ES \longrightarrow^{e} E_\psi & (63)\\
(M10) & b,c>0,d<0 & EI \longrightarrow^{b} E_\psi & dES \longdashv^{c} E_\psi & EE \longdashv^{d'} E_\psi & \\
(M11) & b,c>0,e<0 & EI \longrightarrow^{b} E_\psi & dES \longdashv^{c} E_\psi & ES \longdashv^{e'} E_\psi & \\
(M12) & b,c<0,d>0 & EI \longdashv^{b'} E_\psi & dES \longrightarrow^{c'} E_\psi & EE \longrightarrow^{d} E_\psi & \\
(M13) & b,c<0,e>0 & EI \longdashv^{b'} E_\psi & dES \longrightarrow^{c'} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M14) & c,d<0,e>0 & dES \longrightarrow^{c'} E_\psi & EE \longdashv^{d'} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M15) & c,e<0,d>0 & dES \longrightarrow^{c'} E_\psi & EE \longrightarrow^{d} E_\psi & ES \longdashv^{e'} E_\psi & \\
(M16) & a,d<0,e>0 & P \longdashv^{a'} E_\psi & EE \longdashv^{d'} E_\psi & ES \longrightarrow^{e} E_\psi & \\
(M17) & a,e<0,d>0 & P \longdashv^{a'} E_\psi & EE \longrightarrow^{d} E_\psi & ES \longdashv^{e'} E_\psi &
\end{array}
$$

Only one of these models verifies $d = e = 0$, i.e., with active causality neither from EE nor from ES, namely $M1$, which is the model described in this paper.

The sixteen other models have an active causality coming from EE or from ES or from both of them (in this case with opposite signs). These seventeen models, that are consistent with additional constraints (33), are exactly those (21) found by the qualitative analysis and its programming in ASP (with three or $m+1$ time points). In consequence, they are exactly the minimal solutions of our original problem when additional constraint (14) is assumed.

This gives us the following result:

Proposition 1. *There are seven minimal models, in terms of signs of all possible causalities exerted on mood E_ψ (stimulation, inhibition, or inactivity), of our causal theory presented in Fig. 2B or Fig. 4A, which cover the three core phenomena of non-initiation, initiation followed by discontinuation, and initiation followed by non-discontinuation. All have two active causalities and are given by* (20) *or* (51)*.*

If we add the restoration of the basal mood value at the end in the case of initiation followed by non-discontinuation, then there are seventeen minimal models, all with three active causalities, given by (21) *or* (63)*, and the model derived from statement* (1) *studied in Section 2 is one of them.*

We interpret the identity of the minimal models of our under- and over-approximations as a sign of the robustness of the actual minimal models of our theory, in the sense that these causal solutions of our problem of covering the three core phenomena do not change when we constrain the theory more (by assuming linear causalities) or less (by relying solely on the qualitative evolution profiles of the causal input variables and adopting qualitative modeling and reasoning): no causal solution is excluded in the first case, and no other causal solution appears in the second case.

5.10. Study of inter-individual variability

If we consider the intensities of the causalities, whereas for those of the thermodynamic model, i.e., k and l, they can be considered constant between individuals because fixed by thermodynamics, for those exerted on the psychological variable E_ψ, we have so far tacitly admitted that they are specific to each individual, whatever the causal input variable - psychological, in the case of a, or thermodynamic, in the case of b, c, d, e. This inter-individual variability might seem necessary to explain the three core phenomena, the occurrence of one or other of which depends precisely on each individual. However, if we examine the conditions (34, 35 37) of these phenomena, we find that the quantity $(E_0 - E_{min}) + aP_0$, plays a particular role, alongside causal intensities. Both $E_0 - E_{min}$, the basal state mood level above its minimal threshold, and P_0, the pleasure level of food rejection during the intervention period, are psychological parameters that can be assumed with a high degree of certainty to be person-specific. A natural question, then, is whether these person-specific parameters could suffice to explain the three core phenomena, i.e., without assuming the inter-individual variability in the intensities of the causalities, but quite the opposite, assuming that they could be physiological constants virtually identical for all individuals?

Let us denote the quantity $(E_0 - E_{min}) + aP_0$ by G. $E_0 - E_{min}$ and P_0 are positive quantities, as large as necessary (at least theoretically, physiologically they are necessarily upper bounded). So, if $a \geq 0$, G can take on any positive value and, if $a < 0$, any value (positive, negative or zero), whatever the value of $|a|$, just by varying the person-specific parameters $E_0 - E_{min}$ and P_0. The conditions (34, 35 37) relating to the time behavior of $E_\psi(t)$, which determines the three core phenomena, can be rewritten as:

$$\textit{interruption in } t = 0 \quad G \leq (b-c)I_0 \tag{64}$$

$$\textit{interruption in } t > 0 \quad (b-c)I_0 < G < (b-c)I_0 - \frac{K}{k+l}I_0 \tag{65}$$

$$\textit{no interruption} \qquad G > (b-c)I_0 \quad \textit{and} \quad G \geq (b-c)I_0 - \frac{K}{k+l}I_0 \tag{66}$$

Given the possible values of G, a constraint like $G > x$ or $G \geq x$ is always achievable by G and a constraint like $G < x$ or $G \leq x$ is always achievable if $a < 0$ and achievable if and only if $x > 0$ if $a \geq 0$. We can then conclude the following.

(64) (which implies non-initiation) is always feasible by G if $a < 0$ and feasible if and only if $b > c$ if $a \geq 0$.

(65) is feasible by G if and only if $K < 0$ (the left-hand side has to be smaller than the right-hand side) if $a < 0$ and if and only if $K < 0$ and $b - c > \frac{K}{k+l}$ if $a \geq 0$. Furthermore, if these conditions are satisfied, then G can take any value between $(b-c)I_0$ and $(b-c)I_0 - \frac{K}{k+l}I_0$, and so, from the expression of the time T of interruption (36), it follows that T can take any positive value, thus covering the phenomena of non-initiation (when $T < 1$ week), initiation followed by discontinuation (when $1 \leq T \leq 12$ weeks) and initiation followed by non-discontinuation (when $T > 12$ weeks).

(66) (which implies initiation followed by non-discontinuation) is always achievable by G.

In conclusion, the three conditions (64,65,66), and the three associated core phenomena, are achievable by G alone, if and only if $K < 0$ if $a < 0$, and if and only if $b > c$ and $K < 0$ if $a \geq 0$ (since, in this case, the additional constraint $b - c > \frac{K}{k+l}$ is automatically verified). We already knew that K must be negative in the case of the interruption scenario at $t > 0$ (see Fig. 8B), so, if it is constant, it is necessarily negative. Note that this means that actually $E_\psi(t)$ is strictly decreasing in the no interruption scenario (so, in Fig. 8C, the case where it is strictly increasing or constant cannot occur). This give us the following result:

Proposition 2. *The three core phenomena are explained by simply assuming that the two psychological parameters $E_0 - E_{min}$, the basal state mood level above its minimal threshold, and P_0, the pleasure level of food rejection during the*

intervention period, are person-specific. More precisely, the intensities a, b, c, d, e of the five causal relationships exerted on mood E_ψ need not be variable from individual to individual but can be considered constant parameters, identical for all individuals, from the moment they satisfy either of the constraints $a \geq 0, b > c, K < 0$ or $a < 0, K < 0$, where $K = lb - (k+l)c - kd - e$. Note that, in this case, the mood is necessarily strictly decreasing in the no interruption scenario, as it in the interruption in $t > 0$ scenario.

It is interesting to note that $E_0 - E_{min}$ reflects the level of mood before the start of the diet, in a way the psychological capital before the intervention, and P_0 another psychological feature that is specially mobilized during the intervention, the two complementing each other. Now, the constraint that the causal intensities, if constant, must satisfy so that the mere inter-individual variability of these two psychological features is sufficient to explain the three core phenomena can be stated for each solution model as follows (note that the value of a, which does not appear in K, is not constrained).

Constraint on causal intensities for models mi (51)*:*

$$\begin{array}{llll} m1: & 1 < \dfrac{b}{c} < 1 + \dfrac{k}{l} & & \\ m2: & \dfrac{d}{b} > \dfrac{l}{k} & m3: & \dfrac{e}{b} > l \\ m6: & \dfrac{d}{c'} > 1 + \dfrac{l}{k} & m7: & \dfrac{e}{c'} > k + l \end{array} \tag{67}$$

here is no constraint for $m4$ and $m5$. We see that the constraint relates the ratio of the intensities of the two active causalities in each model to the values k, l of the parameters of the thermodynamic model.

Constraint on causal intensities for models Mi (63)*:*

$$\begin{array}{llll} M1: & 1 < \dfrac{b}{c} < 1 + \dfrac{k}{l} & & \\ M4: & \dfrac{d}{b} > \dfrac{l}{k} & M5: & \dfrac{e}{b} > l \\ M6: & \dfrac{d}{c'} > 1 + \dfrac{l}{k} & M7: & \dfrac{e}{c'} > k + l \\ M8: & kd > lb + e' & M9: & e > lb + kd' \\ M10: & b > c, \quad (k+l)c > lb + kd' & M11: & b > c, \quad (k+l)c > lb + e' \\ M12: & c' > b', \quad lb' + kd > (k+l)c' & M13: & c' > b', \quad lb' + e > (k+l)c' \\ M14: & e > (k+l)c' + kd' & M15: & kd > (k+l)c' + e' \\ M16: & e > kd' & M17: & kd > e' \end{array} \tag{68}$$

There is no constraint for $M2$ and $M3$. Consider, for example, the meaning of the constraint for the model $M1$ we have studied in detail, derived from statement 1. In this model, E_ψ is subjected to stimulation by EI of intensity

b, inhibition by dES of intensity c, and stimulation by P of intensity a. The constraint thus stipulates that the intensity of stimulation by EI must be greater than that of inhibition by dES, but not too great, their ratio having to be upper bounded by $1 + \frac{k}{l}$, where k and l are the the linear thermodynamic parameters related to metabolic and behavioral adaptation, respectively. Our study shows that if this constraint holds, the three core phenomena can be explained with only $E_0 - E_{min}$ and P_0 being person-specific quantities, and the intensities a, b, c being constant.

6. Discussion

We refer to the psychology-based companion preprint (Irigoin-Guichandut et al., 2022) and the submitted article (Irigoin-Guichandut et al., 2023) for a detailed discussion of the psychological issues related to this work and just outline the main conclusions.

A primary goal of psychological science is to establish robust and generalizable theories of human behavior (Rozin, 2009). Because most psychological theories are verbal, hence would be often impossible to test and if necessary reject, some scholars have declared that psychology is facing a theory crisis (Muthukrishna and Henrich, 2019) and have recommended their formalization (e.g., Bjork, 1973). These formal theories are frequently conflated with data models, that are statistical models obtained by fitting experimental data. It has been argued that formal theories need to be distinguished from data models (Fried, 2020a; Haslbeck et al., 2022). Both theories and data models require an assumption of structural homogeneity, i.e., they should apply to all individuals of a population, while structural heterogeneity would require person-specific theory or data model for each and every individual (Lykken, 1991). This assumption is mostly tacit and has rarely been stated or questioned (Lykken, 1991, 2004; Borsboom et al., 2009; Richters, 2021). Furthermore, theories and data models face a challenge, their poor ability to explain inter-individual variability (Borsboom et al., 2009), initially identified by Cronbach (1957). In an attempt to "solve this problem by integrating four perspectives: historical research, philosophical analysis, mathematical modeling, and substantive psychological theory", Borsboom (2008) asked the following question: Are there data "models which may be used to connect intra-individual dynamics and interindividual differences"? This approach has met with some success and suggested a potentially negative answer (Borsboom et al., 2009).

We felt that this question deserved to be reexamined focusing on theories instead of data models, and our approach fits in with the Theory Construction Methodology (TCM) for the formalization and testing of theories that has recently been proposed with the same first author (Borsboom et al., 2021). When formalizing verbal theories, parameters are used, the values of which can be estimated using experimental data. But this is done in general in the form of aggregates that treat inter-individual variability as noise, explaining the poor ability of theories to explain such variability. Here, we avoided aggregates by

taking care mathematically, in the framework of a qualitative, not statistical, approach, of the infinity of person-specific values of parameters and asked positively to the question: Is there a theory that could explain inter-individual variability? which corresponds to the question asked by Borsboom (2008) but applied to theories instead of data models. The TCM is thus used to formalize a theory and test its ability to explain inter-individual variability. It adopts the distinction between theory, empirical phenomena and empirical data, and the view, common in psychology, that theory relates to empirical data indirectly, via empirical phenomena (Haig, 2005). It proposes to test theories against empirical phenomena and defines a theory as "a set of theoretical principles that putatively explain ... [empirical] phenomena", and empirical phenomena, often called 'effects' in psychology, as "robust generalizations of patterns in empirical data" (Borsboom et al., 2021, pp. 756 and 758), e.g., relationships between two variables.

Our formal theory can appear outrageously simple when compared to behavior-change theories (Michie et al., 2014). Actually, its specificity to a timescale could explain why it is devoid of socio-environmental and cognitive factors. This timescale applies to E_ψ fluctuations and inter-related variables. With an aggregate interval of one week and an observation interval of twelve weeks, our theory does not consider the shorter (e.g., E_ψ decrease throughout working days with recovery during the weekend, with an existence interval of one week) and the longer (e.g., E_ψ decrease throughout Fall with recovery in Spring, with an existence interval of one year) cyclical E_ψ fluctuations. The same reasoning would apply to E_ψ fluctuations due to socio-environmental and cognitive factors whose existence intervals are shorter than one week, or a few times longer than twelve weeks. In addition, the severely hypocaloric dietary intervention that we have considered induces a phenomenon of elevated mood (Fond et al., 2013) allowing the theory to be conceived based on mood disorder symptoms. It seems that the high value of E_ψ prevents a discontinuation due to the potential decreasing effect of socio-environmental or cognitive factors operating at shorter timescales, whose effect becomes unapparent, and makes its slow decrease visible, both of these making the theory conceivable.

A distinctive characteristic of our theory is that it combines a psychological model (with its two psychological variables and one relationship) with person-specific parameter values, and a thermodynamic model (with its four energetic variables and three relationships) with a unique value of parameters, the latter exerting a causal influence on the first by means of two relationships (extended to four in our search of all possible alternative models) with person-specific parameter values. Relying on thermodynamics (basically, on the principle of the conservation of energy applied to a living organism) and its coupling with the psychological model allows us to establish this model on firmer grounds. Relying more generally on well established theories in the field of physics or biology and their causal relationships with psychological variables is an approach which deserves to be more largely explored, aiming at universally accepted theories in psychology.

Thanks to thermodynamics, the energetic variables and their causal relation-

ships accurately summarize the energy flow through living organisms resulting from thousands of different molecular processes, well known in bioenergetics, a part of biochemistry and molecular biology, and lead to the transformation of energy so that biological work can be performed. Consequently, they can provide an analogy to conceptualize psychological variables and their underlying molecular processes, which is in line with the NIH Science of Behavior Change program and its different levels of analysis of mechanisms including psychological and neurobiological levels (Nielsen et al., 2018), an approach inspired by the NIMH Research Domain Criteria initiative in psychopathology (Kozak and Cuthbert, 2016). In addition, all molecular processes studied in bioenergetics are under the influence of a control system that maintains energy balance in certain circumstances and that can be perturbed in the case, for example, of starvation or hypocaloric diet. At the thermodynamic level of the organism, the combined model proposes a model of bioenergetics and its control system which goes beyond physiological adaptations, that only require energetic variables, with the addition of variables E_ψ and P and their causal relationships. Hence, the combined model proposes at the thermodynamic level a model of this control system which is studied at the neurobiological level (Morton et al., 2014).

The conditions of the three dynamic phenomena of dietary behavior change that witness inter-individual variability are expressed in terms of the parameters of the combined theory (hidden in the $\overline{X}(t)$ variables occurring in (8) or (10); explicit in (34,35,37)). In so doing, they also propose a theoretical and specific solution to the schism between experimental and differential psychology (Cronbach, 1957), connecting psychological theories to relatively stable psychological traits people differ on that explain inter-individual variability. That is how, adopting an individual differences perspective, a psychological trait labeled 'ability to restrict food intake' could be conceived and match any of the three dynamic phenomena, depending on its value, low, medium, or high.

A certain form of generalization might be offered by the strong program of construct validity (Kane, 2001). Theories that are commonly said to "help explain, predict, and control phenomena" (Fried, 2020b, pp. 271), may also help implicitly define their constructs by their network of interrelationships. Cronbach and Meehl (1955) ideally requested this type of definition as part of this program, constructs being embedded in a 'nomological network' made of 'laws' relating constructs to each other and to observable variables. Even though validity theorists agree that this request is unrealistic in psychology "given the dearth of highly developed formal theories" (Kane, 2001, pp. 326), our combined theory, which is made of constructs, observable variables provided by thermodynamics, and precise interrelationships, could constitute a nomological network and implicitly define the construct E_ψ.

Finally, our evaluation of the combined theory and of potential alternative theories (considering all possible causal relationships from energetic variables to E_ψ) overcomes a potential challenge (Eronen and Bringmann, 2021), i.e., phenomena would be unable to impose a constraint strong enough to reject theories. Actually, we performed the exclusion of theories unable to explain all

three single-timescale dynamic phenomena by using Bounded Model Checking (Biere et al., 1999) after over-approximation and discretization, which reflects the strength of the constraint imposed by these three phenomena. Our recourse to a qualitative approach encompassing all possible values of individual parameters, and to an integer-based discretization resulting in a finite model that can be exhaustively checked, is novel in psychology, where the formalization attempts rely in general on statistical data models unable to formally exclude theories.

7. Conclusion

We have developed a theory of severe hypocaloric dietary behavior change that explains the dynamic eating behavior change phenomena of non-initiation, discontinuation or non-discontinuation. This theory is firstly expressed as a causal network where the fluctuations of the key psychological variable representing mood, in response to the causal effects exerted on it by the other variables (psychological and energetic, the latter being governed by thermodynamics principles), determine one or the other phenomenon occurrence. A qualitative (i.e., over-approximated) analysis of the theory, using experimental evolution profiles of the causal input variables, is performed, after discretization, in the form of a finite integer-based model which can be exhaustively queried. All minimal models (in terms of active causalities exerting on mood) are found and proved valid thanks to a direct analytical solving of the system of ODEs obtained as the under-approximation of the theory by linearization. Whereas such a bounded model checking technique has already been used in systems biology, it is, to the best of our knowledge, the first time it has been used in psychology.

The characteristics of our theory, that certainly contributed to make it successful, are the following: it couples through causal relationships both psychological and biological or physical (in our case, energetic) variables, the latter governed by universally accepted principles (in our case, thermodynamics when applied to the human organism); it operates at a specific timescale, fitting the dynamics of the phenomena to handle and allowing to discard the processes that operate at shorter or longer timescales (such as socio-environmental and cognitive factors in our case) as non-pertinent with regards to the explanation of these phenomena; it is person-specifically parameterized (in our case, these parameters represent the basal value of mood and the constant value of pleasure during the intervention, i.e., the two psychological variables and, possibly, the intensities of the causalities exerted on mood, but we have seen that the latter can be constant if they satisfy a given constraint) but, avoiding aggregating parameter values and considering inter-individual variability as noise that comes under statistical analysis, takes care mathematically of the infinity of person-specific parameter values by processing them qualitatively as over-approximated intervals, providing an explanation of inter-individual variability; it may constitute a nomological network that implicitly defines the construct corresponding

to the key psychological variable mood. This paves the way toward a universal mood-focused thermodynamic-grounded timescale-specific theory of [eating] behavior change with explainable inter-individual variability.

More generally, we think that our approach is a first step towards helping solve the theory crisis in psychology and it would be worth attempting to apply and test it (with necessary generalizations) to other psychological issues. Indeed, for a number of scholars there is a dearth of psychological theories achieving universal acceptance and thus a need to formalize present theories, which are most of the time verbal. Actually, our model checking technique after formalization allows the safe reject of model candidates. In addition, two longstanding key problems are considered as possible root causes of this theory crisis in psychology, namely that most theories are unable to explain inter-individual variability, and to implicitly define their constructs by their network of interrelationships. In fact, our theory proposes an explanation of inter-individual variability of dietary behavior change (providing thus a negative answer to the question of knowing if the tacit assumption of structural homogeneity in psychological theories is a potential cause of their poor ability to explain inter-individual variability), and an implicit definition of a construct derived from mood disorder symptoms.

8. Notations

a, b, c, d, e: algebraic intensities of linear causalities from P, EI, $-dES$, EE, ES respectively to E_ψ
dES: energy store variation
ΔE: basal mood value above its minimal threshold
$\Delta E^-, \Delta E^+$: minimal and maximal values of ΔE
E_{min}: minimum mood threshold
E_ψ: mood or psychological energy
EE: energy expenditure
EI: energy intake
ES: energy store
I: intervention
I_0: positive value of I during intervention period (imposed restriction of energy intake)
k, l: positive intensities of linear stimulation from ES to EE (metabolic adaptation) and of inhibition from ES to EI (behavioral adaptation), respectively
m: size of the discrete quantity space of the variables EI, dES, EE, ES, P, given as $[-m, m]$
P: pleasure of food rejection
P_0: constant positive value of P during the intervention period
R: restriction
s_X: sign of causality from X to E_ψ
X_0: basal value of variable $X = EI, EE, ES, W, E_\psi$
$\overline{X}$: causal (positive) effect on mood of causal input variable X
W: work

9. Statements and Declarations

The authors have no competing interests to declare that are relevant to the content of this article.

Authors' contributions

Conceptualization: M. I-G.
Methodology and causal analysis: P. D., M. I-G., L. P.
Qualitative formalism and ASP coding: P. D., L. P.
Analytical resolution: P. D.
Discussion, psychological issues, and references: M. I-G., L. M.
Writing: P. D., M. I-G.

References

Abraham, C., Michie, S., 2008. A taxonomy of behavior change techniques used in interventions. Health psychology 27, 379–387. doi:10.1037/0278-6133.27.3.379.

American Psychiatric Association, DSM-5 Task Force, 2013. Diagnostic and statistical manual of mental disorders: DSM-5™ (5th ed.). American Psychiatric Publishing, Inc., Washington. doi:10.1176/appi.books.9780890425596.

Baral, C., 2003. Knowledge Representation, Reasoning and Declarative Problem Solving. Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511543357.

Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F., 1996. Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. Journal of Personality Assessment 67, 588–597. doi:10.1207/s15327752jpa6703_13.

Biere, A., Cimatti, A., Clarke, E.M., Zhu, Y., 1999. Symbolic Model Checking without BDDs, in: Cleaveland, W.R. (Ed.), Tools and Algorithms for Construction and Analysis of Systems. TACAS 1999, Springer Berlin Heidelberg, Berlin. pp. 193–207. doi:10.1007/3-540-49059-0_14.

Bjork, R.A., 1973. Why mathematical models? American Psychologist 28, 426–433. doi:10.1037/h0034623.

Borsboom, D., 2008. A science in search of its subject: The relation between psychological processes and individual differences. Technical Report. Psychologische Methodenleer. URL: https://www.nwo.nl/en/projects/451-03-068.

Borsboom, D., Kievit, R.A., Cervone, D., Hood, S.B., 2009. The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis, in: Valsiner, J., Molenaar, P.C.M., Lyra, M.C., Chaudhary, N. (Eds.), Dynamic Process Methodology in the Social and Developmental Sciences. Springer, Berlin, pp. 67–97. doi:10.1007/978-0-387-95922-1_4.

Borsboom, D., van der Maas, H.L.J., Dalege, J., Kievit, R.A., Haig, B.D., 2021. Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. Perspectives on Psychological Science 16, 756–766. doi:10.1177/1745691620969647.

Cronbach, L.J., 1957. The two disciplines of scientific psychology. American Psychologist 12, 671–684. doi:10.1037/h0043943.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281–302. doi:10.1037/h0040957.

Eronen, M.I., Bringmann, L.F., 2021. The Theory Crisis in Psychology: How to Move Forward. Perspectives on Psychological Science 16, 779–788. doi:10.1177/1745691620970586.

Fishbein, M., Hennessy, M., Yzer, M., Douglas, J., 2003. Can we explain why some people do and some people do not act on their intentions? Psychol Health Med. 8, 3–18. doi:10.1080/1354850021000059223.

Fond, G., Macgregor, A., Leboyer, M., Michalsen, A., 2013. Fasting in mood disorders: neurobiology and effectiveness. A review of the literature. Psychiatry Res. 209, 253–258. doi:10.1016/j.psychres.2012.12.018.

Foster, G.D., Wadden, T.A., 1995. The social and psychological effects of weight loss, in: Fairburn, C.G., Brownell, K.D. (Eds.), Eating Disorders and Obesity: A Comprehensive Handbook. Guilford Press, New York, pp. 500–506.

Fried, E.I., 2020a. Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. Psychological Inquiry 31, 271–288. doi:10.1080/1047840X.2020.1853461.

Fried, E.I., 2020b. Theories and Models: What They Are, What They Are for, and What They Are About. Psychological Inquiry 31, 336–344. doi:10.1080/1047840X.2020.1854011.

Gebser, M., Kaminski, R., Kaufmann, B., Schaub, T., 2012. Answer Set Solving in Practice. Synthesis Lectures on Artificial Intelligence and Machine Learning, Springer, Berlin. doi:10.1007/978-3-031-01561-8.

Gebser, M., Kaminski, R., Kaufmann, B., Schaub, T., 2018. Multi-shot ASP solving with clingo. doi:10.48550/arXiv.1705.09811.

Glanz, K., Rimer, B.K., U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, 1995. Theory at a Glance: A Guide for Health Promotion Practice. NCI, U.S.

Gourlan, M., Bernard, P., Bortolon, C., Romain, A.J., Lareyre, O., Carayol, M., Ninot, G., Boiché, J., 2016. Efficacy of theory-based interventions to promote physical activity. A meta-analysis of randomised controlled trials. Health Psychol Rev. 10, 50—66. doi:10.1080/17437199.2014.981777.

Hagger, M.S., Luszczynska, A., 2014. Implementation intention and action planning interventions in health contexts: state of the research and proposals for the way forward. Appl Psychol Health Well Being 6, 1—47. doi:10.1111/aphw.12017.

Haig, B.D., 2005. An Abductive Theory of Scientific Method. Psychological Methods 10, 371–388. doi:10.1037/1082-989X.10.4.371.

Hall, K.D., 2012. Modeling Metabolic Adaptations and Energy Regulation in Humans. Annual Review of Nutrition 32, 35–54. doi:10.1146/annurev-nutr-071811-150705.

Haslbeck, J.M.B., Ryan, O., Robinaugh, D.J., Waldorp, L.J., Borsboom, D., 2022. Modeling Psychopathology: From Data Models to Formal Theories. Psychological Methods 27, 930–957. doi:10.1037/met0000303.

Irigoin-Guichandut, M., Muller, L., Paulevé, L., Dague, P., 2022. In search of a way out of the theory crisis in psychology: Targeting two longstanding key problems. URL: https://europepmc.org/article/PPR/PPR526716.

Irigoin-Guichandut, M., Muller, L., Paulevé, L., Dague, P., 2023. From a specificity of very-low-calorie diets to a formal single-timescale theory of severe hypocaloric dietary behavior change. Psychological Review, submitted .

Jensen, M.D., Ryan, D.H., Donato, K.A., Apovian, C.M., Ard, J.D., Comuzzie, A.G., Hu, F.B., Hubbard, V.S., Jakicic, J.M., Kushner, R.F., Loria, C.M., Millen, B.E., Nonas, C.A., Pi-Sunyer, F.X., Stevens, J., Stevens, V.J., Wadden, T.A., Wolfe, B.M., Yanovski, S.Z., 2014. Executive summary: Guidelines (2013) for the management of overweight and obesity in adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Obesity 22, S5–S39. doi:10.1002/oby.20821.

Kane, M.T., 2001. Current Concerns in Validity Theory. Journal of Educational Measurement 38, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x.

Katok, A., Hasselblatt, B., 1995. Introduction to the Modern Theory of Dynamical Systems. Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge. doi:10.1017/CBO9780511809187.

Kozak, M.J., Cuthbert, B.N., 2016. The NIMH Research Domain Criteria Initiative: Background, Issues, and Pragmatics. Psychophysiology 53, 286–97. doi:10.1111/psyp.12518.

Kuipers, B., 1994. Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge. MIT Press, Cambridge, MA.

Langeveld, M., DeVries, J.H., 2015. The long-term effect of energy restricted diets for treating obesity. Obesity 23, 1529–1538. doi:10.1002/oby.21146.

Lifschitz, V., 2019. Answer Set Programming. Springer, Berlin. doi:10.1007/978-3-030-24658-7.

Lykken, D.T., 1991. What's wrong with Psychology, anyway?, in: Cicchetti, D., Grove, W.M. (Eds.), Thinking Clearly about Psychology Volume 1: Matters of Public Interest, University of Minnesota Press, Minneapolis. pp. 3–39.

Lykken, D.T., 2004. Paul Everett Meehl. Applied and Preventive Psychology 11, 53–56. doi:10.1016/j.appsy.2004.02.017.

MacLean, P.S., Wing, R.R., Davidson, T., Epstein, L., Goodpaster, B., Hall, K.D., Levin, B.E., Perri, M.G., Rolls, B.J., Rosenbaum, M., Rothman, A.J., Ryan, D., 2015. NIH working group report: Innovative research to improve maintenance of weight loss. Obesity 23, 7–15. doi:10.1002/oby.20967.

Michie, S., van Stralen, M.M., West, R., 2011. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implementation Science 6, 42. doi:10.1186/1748-5908-6-42.

Michie, S., West, R., Campbell, R., Brown, J., Gainforth, H.L., 2014. ABC of Behaviour Change Theories. Silverback Publishing, UK.

Morton, G.J., Meek, T.H., Schwartz, M.W., 2014. Neurobiology of food intake in health and disease. Nature Reviews Neuroscience 15, 367–378. doi:10.1038/nrn3745.

Muthukrishna, M., Henrich, J., 2019. A problem in theory. Nature Human Behaviour 3, 221–229. doi:10.1038/s41562-018-0522-1.

Nielsen, L., Riddle, M., King, J.W., NIH Science of Behavior Change Implementation Team, 2018. The NIH Science of Behavior Change Program: Transforming the science through a focus on mechanisms of change. Behav Res Ther. 101, 3–11. doi:10.1016/j.brat.2017.07.002.

Pearl, J., 2009. Causality. 2nd ed., Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511803161.

Potassco, 2010. Potassco, the Potsdam Answer Set Solving Collection. URL: https://potassco.org/.

Prestwich, A., Webb, T.L., Conner, M., 2015. Using theory to develop and test interventions to promote changes in health behaviour: evidence, issues, and recommendations. Current Opinion in Psychology 5, 1–5. doi:10.1016/j.copsyc.2015.02.011.

Richters, J.E., 2021. Incredible Utility: The Lost Causes and Causal Debris of Psychological Science. Basic and Applied Social Psychology 43, 366–405. doi:10.1080/01973533.2021.1979003.

Rosenbaum, M., Hirsch, J., Gallagher, D.A., Leibel, R.L., 2008. Long-term persistence of adaptive thermogenesis in subjects who have maintained a reduced body weight. The American Journal of Clinical Nutrition 88, 906–912. doi:10.1093/ajcn/88.4.906.

Rozin, P., 2009. What Kind of Empirical Research Should We Publish, Fund, and Reward?: A Different Perspective. Perspectives on Psychological Science 4, 435–439. doi:10.1111/j.1745-6924.2009.01151.x.

Stevens, S.S., 1946. On the Theory of Scales of Measurement. Science 103, 677–680. doi:10.1126/science.103.2684.677.

Strasser, B., Berger, K., Fuchs, D., 2015. Effects of a caloric restriction weight loss diet on tryptophan metabolism and inflammatory biomarkers in overweight adults. European Journal of Nutrition 54, 101–107. doi:10.1007/s00394-014-0690-3.

Sumner, J.A., Carey, R.N., Michie, S., Johnston, M., Edmondson, D., Davidson, K.W., 2018. Using Rigorous Methods to Advance Behaviour Change Science. Nature Human Behaviour 2, 797—799. doi:10.1038/s41562-018-0471-8.

Vallacher, R.R., Van Geert, P., Nowak, A., 2015. The Intrinsic Dynamics of Psychological Process. Current Directions in Psychological Science 24, 58–64. doi:10.1177/0963721414551571.

Vrijens, B., De Geest, S., Hughes, D.A., Przemyslaw, K., Demonceau, J., Ruppar, T., Dobbels, F., Fargher, E., Morrison, V., Lewek, P., Matyjaszczyk, M., Mshelia, C., Clyne, W., Aronson, J.K., Urquhart, J., for the ABC Project Team, 2012. A new taxonomy for describing and defining adherence to medications. British Journal of Clinical Pharmacology 73, 691–705. doi:10.1111/j.1365-2125.2012.04167.x.

Wadden, T.A., Stunkard, A.J., Brownell, K.D., 1983. Very low calorie diets: their efficacy, safety, and future. Ann Intern Med. 99, 675–684. doi:10.7326/0003-4819-99-5-675.

West, R., Godinho, C.A., Bohlen, L.C., Carey, R.N., Hastings, J., Lefevre, C.E., Michie, S., 2019. Development of a formal system for representing behaviour-change theories. Nature Human Behaviour 3, 526–536. doi:10.1038/s41562-019-0561-2.

Williamson, D.A., 2017. Fifty Years of Behavioral/Lifestyle Interventions for Overweight and Obesity: Where Have We Been and Where Are We Going? Obesity 25, 1867–1875. doi:10.1002/oby.21914.

Zaheer, S., Albert, S., Zaheer, A., 1999. Time Scales and Organizational Theory. The Academy of Management Review 24, 725–741. doi:10.2307/259351.

Zhang, Y., Liu, C., Zhao, Y., Zhang, X., Li, B., Cui, R., 2015. The Effects of Calorie Restriction in Depression and Potential Mechanisms. Current Neuropharmacology 13, 536–542. doi:10.2174/1570159X13666150326003852.