On apples and ageing

Ignophi Hu

September 20, 2024

Objective: In reviewing the literature on various topics in the field of ’ageing’, similar issues kept resurfacing. To avoid redundancy, I decided to compile these recurring themes into a single discussion. The goal here is to examine the utility of the current concept of ’ageing’. In particular, this discussion considers how well this concept serves in addressing key objectives, such as measuring ’ageing’, evaluating the validity of ’ageing’ theories, assessing interventions, and examining the validity of experiments conducted in the field of ’ageing’.

Keywords: *Ageing, measurement, model, validity*

Contents

1. How to Measure ‘ageing’?

Given the critical role of measurement in any research field, this question has been central to numerous debates in the field of ’ageing’. It has resurfaced in various forms, with its most recent iterations being “what is a biomarker of ageing?” and “how to measure biological age?” [1, 2]. Before addressing this frequently asked question and its numerous incarnations, I will consider a simpler - maybe less popular - one: “How to measure an ’apple’?” Based on my pilot experiment asking this question, answers usually fall into two general categories: (1) Those who decide on a dimension of interest, answering directly for example “using a ruler” or “using a balance” and (2) Those who reply with a question “what do you want to know?”[1]

What is a measurement?

A measurement is a comparison [3]. It is the expression of one variable as a function of another. When one says something weighs 10 kg, this “something” (element 1) is “10” (relation) whatever the “kg” (element 2) is. A measurement can be viewed as a projection of one element onto another. It is a relative relationship based on the chosen projection space. In other words, the same element can be expressed in as many ways as there are references to project onto. Given a real set element (**Figure** 1), it will be associated with a set of projections (referred to here in as “a measurement set”). For example, consider an ’apple’. One can measure this element in many ways: with a balance (its weight), with a ruler (its width), by throwing it in still water and counting the waves, by the length of its shadow from a light source at a certain angle, by burning it and weighing the ashes, by throwing it against the wall and measuring the area of the splash or by noting its most dominant color (e.g., red). Consider the relationships among these measurements. Some measurements covary; for instance, a larger width correlates with a heavier weight, and in turn, a heavier weight might indicate a greater quantity of ashes after burning. However, the strength of this correlation varies; for example, weight expressed in kilograms

[1] There is also a third group which just refused to engage with the experiment

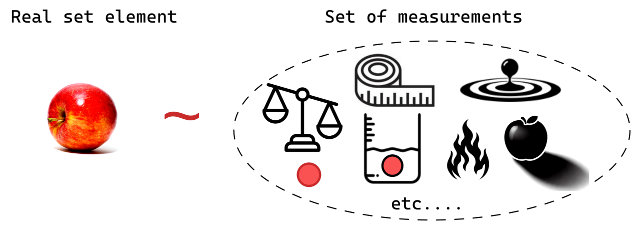

Figure 1. Conceptual Illustration of a measurement set. This diagram exemplifies how a real set element can be associated with a diverse set of projections.

closely predicts weight expressed in pounds but correlates less strongly with volume. Conversely, some measurements are more independent, exhibiting weak correlation. For instance, the weight of an object is not that informative about its dominant color or its saltiness.

Before proceeding further with this example, it is helpful to briefly introduce some concepts and terminologies from the fields of psychometrics and scale development.

- **Model**: A model is a set of elements and probabilistic relationships among them. As such, to measure something is to model it, as it builds a relationship between an element of interest and a reference of choice (projection space).

- **Validity**: The validity of a model is a measure of its utility in making a specific inference. For e.g., Consider the objective of determining which of these two objects is heavier—a book or a pencil. I can place each object on my palm and feel the pressure it exerts against it. In other words, I am projecting each object into the sensory space of pressure perceived by my palm. i.e. I am modeling each object, building a relationship between each object (element of interest) and the pressure sensation (projection space). Expressing these objects as palm pressure models establishes a relationship among them, allowing for comparison and thus aiding in determining which one is heavier. As such, the palm pressure model has some validity in achieving the objective of determining which of these two objects is heavier. Now, consider a different objective: determining which of the same two objects is 'blue'. The exact same palm pressure model would not be as helpful for this objective. Thus, the validity of the palm pressure model depends on the objective. Considering again the first objective (determine which is heavier, a book or a pencil), one can place the objects on a scale and compare the indicated values. This scale model of the objects also allows achieving the objective. Thus, multiple models can be valid for a given objective.[2] As such, a model's validity is a function of its intended use. For a given model there are as many validity measurements as there are objectives it is applied to achieve. As nicely expressed by Nunnally in 1970,

  > "Strictly speaking, one validates not a measurement instrument but rather some use to which the instrument is put" [4].

- **Reliability**: The reliability of a model refers to the generalizability (similarity) of a desired inference across a specific facet (variable) of interest [4]. For e.g, test-retest reliability refers to the generalizability of an inference across the facet of time (the time between two tests). i.e., how similar are the inferences one makes from a measurement done at time t1 and one made at time t2. Altitude reliability refers to the generalizability of an inference across the facet of altitude. i.e., how similar are the inferences one makes from a measurement done at sea level and one done at very high altitudes. As such, one cannot simply say "a model is reliable". For a given model there are as many reliability measurements as there are facets to generalize over. When one speaks of a model's reliability, it is in relation to a specific facet over which a specific inference is to be generalized[3-4].

To summarize, a model's validity and reliability are not intrinsic properties. These attributes depend on

[2] For simplicity of the example, I assume the 'true' inferences are known and thus can be used to evaluate the validity of the models in the example, though this is usually not the case and will be discussed further in this article and future works due to its significant implications.

[3] When aiming to generalize an inference across multiple facets, the technical term used is "generalizability" rather than "reliability".

[4] Note that "reliability" is encompassed within the broader term of "validity". One can simply say that "reliability" is the validity of a given model for the objective of "generalizing an inference over a given facet". Though it is a typical objective when developing a measurement scale, deserving its own term.

the specific inference(s) intended from the model's application and facet(s) over which the inference is to be generalized. Now, applying these concepts to the apple measurement question leads to asking, "for what objective?" Though one might ask, does one need to answer this question? is there no model among the 'apple' measurement set that would be the optimal choice[5] irrespective of objective? It would be difficult to empirically test this question as it would require first knowing all the elements of the 'measurement set' and then applying them to 'all possible objectives'. However, the No Free Lunch (NFL) theorem for optimization algorithms might be interesting to discuss in relation to this question [5]. An algorithm is a sequence of steps designed to perform a specific task or solve a particular problem. The NFL theorem states that there is no single algorithm that will perform optimally for all possible problems. Each algorithm's performance is contingent on the specific problem context, meaning that an algorithm that excels in one scenario may perform poorly in another. Across all algorithms and all possible problems, the average performance is essentially the same. A similar situation might apply to the question above. Across all measurement models and all possible objectives, the average validity might essentially be the same. However, this does not imply that a measurement model can be chosen arbitrarily for a given subset of objectives, as specific models may perform better for particular objectives due to their unique characteristics and the context of their application. This underscores once more the importance of the objective for the choice of the measurement model.

As such, to address the question of "how to measure 'ageing'", one needs to answer the question of "for what objective?". A typical answer to such a question goes as follows: "To study the processes of 'ageing' and to identify interventions to slow down or reverse 'ageing'" [2, 1]. To tackle this ill-defined objective, I will have to go on a linguistic tangent and address a different question first: What is 'ageing'?

2. What is 'ageing'?

There is no universally agreed-upon definition of the term 'ageing' with various definitions varying widely in scope and some even questioning whether "there is such a thing as 'ageing'" [6, 7, 8]. Definitions are attempts to increase the validity of a word[6] by trying to delineate its intended meaning to ensure greater consistency across users. This linguistic debate over definitions carries practical implications. Each definition of 'ageing' incorporates a set of assumptions critical to assessing experimental methods and interpreting their outcomes. To illustrate this, I will use another question from the 'ageing' field "what should a theory of 'ageing' explain". A theory is a model typically used to explain some set of observations and make a set of testable predictions. As such, its validity depends on how well it explains the desired observations and how well its predictions match reality. This point is straightforward and generally agreed upon. The challenge arises in enumerating and agreeing on the set of observations that is to be explained[7] as well as the validity of the tests of a theory's prediction. This issue manifests through disagreements in the field over which 'ageing theories' are considered "disproven", for e.g.

> "... Moreover, some of them are now obsolete in the light of current data on the biological basis for ageing. Nevertheless, the misconception that there are more than 300 valid theories of ageing or numerous valid theories [...] still persists among many authors and researchers. In addition, some of these outdated and discarded hypotheses, [...] still can be found in today's medical textbooks, scientific publications aimed at the general public, and scientific writing" [9].

To illustrate how the disagreement over the definition of the term 'ageing' contributes to this problem, I will

[5] in terms of validity / reliability

[6] Words, much like measurements, are models designed to serve specific objectives, primarily communication (Appendix I). As such, to evaluate the "validity" of a word, one must consider the inferences made by its user. The typical 'objective' of the user is the transfer of information. Put another way, the user is making an 'inference' about the concepts induced in the recipient as a result of using a specific symbol. Similarly, the recipient of a word is making an 'inference' about the concepts the user of the word wishes to induce in their mind. In what follows, to avoid repetitiveness, when I refer to "a word's validity", it will be assumed that this is the objective. Given this objective for using words, the 'validity' of a word hinges on both its user and its recipient, specifically on their respective models of a given symbol (Appendix I).

[7] assumptions accepted

use a specific example of a theory and an experiment that was done to test it. The somatic mutation theory of 'ageing' proposes that the accumulation of DNA mutations in somatic cells throughout an organism's life is central to the deterioration observed in 'ageing' [10]. In 1961, there was an interesting experiment that was performed to assess the validity of this theory [11]. The study ingeniously employed male haploid and diploid forms of the wasp Habrobracon, utilizing their inherent genetic differences—haploids having a single set of chromosomes and diploids having two. If the aforementioned theory holds some validity, the following predictions could be made:

- 1) Radiation exposure which induces DNA mutations, should impact the haploid wasp more than the diploid wasp.
- 2) Under no radiation exposure, the haploid wasp should age faster compared to the diploid wasp.

The study found that haploid males, which have a single set of chromosomes and therefore no redundant genetic information within a given cell, were more adversely affected by radiation than diploids (measured in terms of a shorter lifespan) (**Figure** 2). On the other hand, under non-irradiated conditions, both haploid and diploid male wasps exhibited "similar" lifespans and mortality rates (**Figure** 2). Interesting results! So now back to the somatic mutation theory, is it now invalid? The answer to that question depends significantly on the assumptions made by the observer evaluating this experiment. To demonstrate this, I will explore two examples of such assumptions and their impact on the answer.

- **Ageing is similar across species:** This assumption manifests in various forms, for example "is 'ageing' universal?", "do different species age at different rates?", and "is species X an adequate model to study human 'ageing'?". If one accepts this assumption, then given the experiment above the somatic mutation theory's validity is in question. On the other hand, if one does not accept it, this experiment's observations are irrelevant to the somatic mutation theory's validity for human 'ageing'. There are various arguments to dismiss its relevance for humans if desired. For example, it could be pointed out that the wasps have a maximum lifespan of 92 days, whereas humans often live over a century. This difference in lifespan could mean that while somatic mutations might not appear relevant in the short-lived wasps, they could be more impactful over the longer human lifespan.

  As such, If one is to accept this assumption, then a "theory of 'ageing'" should take into consideration if and why different species age differently and how [1]. Conversely, if one does not accept it then one does not need to consider it and can instead focus on explaining a subset of observations related to the 'ageing' in a particular species or group of species that are of interest. One might object that a "better" (i.e. more valid) theory of 'ageing' is one that makes the least assumptions and explains the most observations. Of course, everyone would welcome a more useful model. However, which one is assuming the least? the one assuming that whatever 'ageing' is, it is similar across species? or that whatever 'ageing' is does not have to be similar across species? Either one of these is an assumption and each side of the debate has certain set of observations and resolution level to choose from to push their claims.

- **Average lifespan is an adequate measure of 'ageing':** Alternatively, one could challenge the claim that "both haploid and diploid male wasps exhibited similar lifespans and mortality rates", and, more critically, its implications concerning 'ageing'. There are multiple ways to counter this claim. First, one could point to the tail section of the lifespan curve (**Figure** 2) and contend that diploids appear to have a slightly longer survival, implying a lack of sufficient power and the need for a greater number of wasps to detect smaller differences. Second, one might argue that maximal lifespan is a 'better' metric than average lifespan, and under this chosen projection, the claim might be less well-supported. Third, one might leverage the fallacy of composition by highlighting that lifespan represents just one aspect of 'ageing' and does not alone justify claims about the whole. Such criticisms would be usually accompanied with a call for more appropriate "characterization or phenotyping", incorporating additional measures such as mobility, "cellular" damage measurements, and so forth.

In this example, the results happened to be negative for the theory in question, but the same strategies (which are but 2 examples of a much broader repertoire) and arguments could be used if the results had supported the theory and someone wanted to dispute its validity. The aim was to demonstrate how disagreements

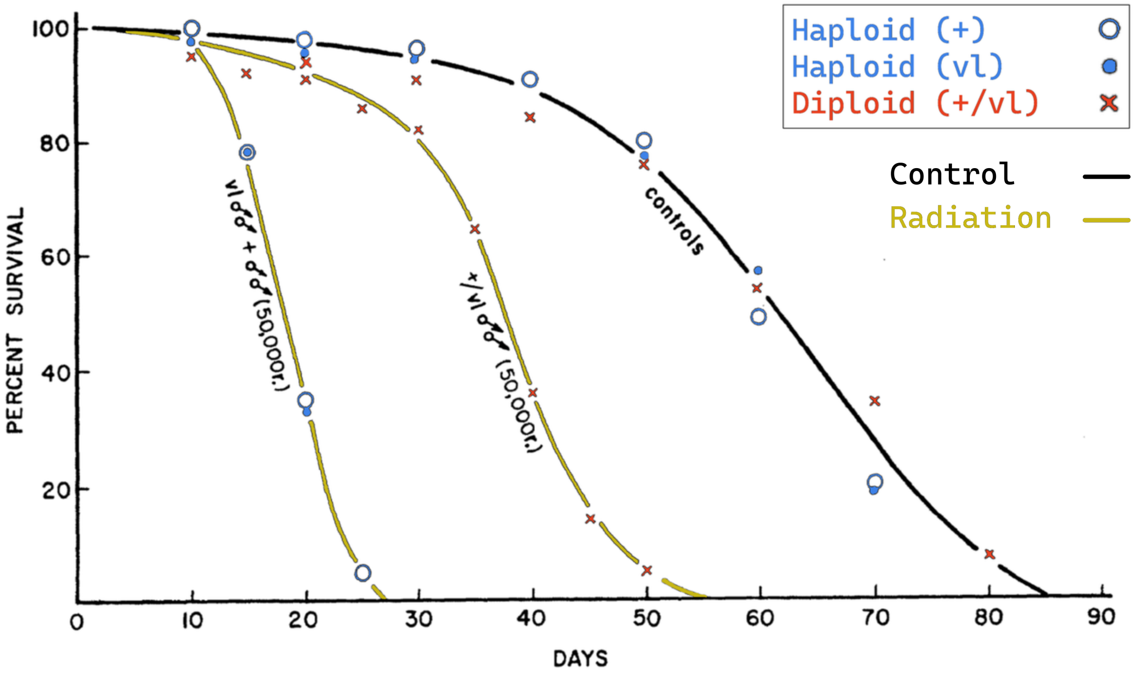

Figure 2. Adapted from [11] Clark & Rubin (1961), *Radiation Research*, 15(2), 244–253, https://doi.org/10.2307/3571256. Figure modified with added colors and legend for clarity. Life span of haploid (+ or vl) and diploid (+/vl) males of *Habrobracon* sp. after exposure as adults to 50,000 r. Haploid (+) males (blue open circles); haploid (vl) males (blue closed circles); diploid (+/vl) males (red crosses).

over fundamental assumptions about ’ageing’ lead to an inability in assessing hypotheses. This dispute over “disproven” theories is but merely one symptom of the broader disagreement over what the term ’ageing’ represents. Many of the other unending debates and pseudo-questions in the field can be derived from the current popular definition of the term. For example, lets consider a recent work with the stated goal of:

> “Here, we advance a framework for the terminology and characterization of biomarkers of aging, including classification and potential clinical use cases” [2].

The definition of ’ageing’ put forward by this paper is the following:

> “The process of accumulation of consequences of life, such as molecular and cellular damage, that leads to functional decline, chronic diseases, and ultimately mortality” [2].

What follows are examples of debates that arise from or exacerbated by such a definition, highlighting how it leads to unsolvable disputes because, at their core, these debates are rooted in a disagreement over fundamental assumptions implicitly carried by the definition, making meaningful evaluation unattainable.

2.1 When does ’ageing’ start?

To apply the definition above one has to measure the so-called “accumulation of consequences of life”. This “accumulation” implies that it had to start from a state where there was zero accumulation of these “consequences” or at least a minimal amount. And thus the natural question that follows is “when does ’ageing’ start?” One might argue that one does not require this state to measure ’ageing’ since one can examine changes at various stages rather than solely from the ’minimum stage’. Still, the above question has great appeal because it implies that the “first” change / accumulation might give a better hint on the “causative” consequences that are leading “ultimately to mortality”. A clear illustration of the “ultimate” result of using such a broad and ambiguous definition of ’ageing’ in practice is evident in one conclusion addressing this question: “We propose a model of ‘ground zero’, the mid-embryonic state characterized by the lowest biological age at which both organismal life and aging begin” [12]. This view is not singular; it is echoed in a publication appropriately named ’Ageing definitions, mechanisms and the magnitude of the problem’, which asserts that “’Ageing’ in contrast refers to any time related process and could be said to begin at conception” [13]. These claims have a few implications. On one hand they imply that one has to consider all the ’changes’ that happen from conception and then attempt to sort these out whether they are leading “ultimately to mortality”. On the other hand, It implies that a young adult at peak physical and reproductive performance already has accumulation of “damage”. I will let the reader evaluate how such a conclusion regarding the start of ’ageing’ is helpful. Such claims are not surprising if one considers the “example” part of the definition “such as molecular and cellular damage” - which is supposed to help clarify and specify what is meant by “consequences of life”. It’s a relief that this clarification doesn’t stretch into atomic damage—we have enough on our plate with just molecules and cells!

2.2 Is ’X’ accelerated ’ageing’?

The second family of pseudo-questions that result from such a definition comes in many flavors for e.g., is disease ’X’ accelerated ’ageing’? is intervention ’X’ a model for accelerated ’ageing’? is ’ageing’ the cause for disease ’X’? This is also related to the long-standing debate about “confounding ’ageing’ and age-associated diseases”. Proponents of the distinction argue that the accompanying diseases are ’distinct’ from ’ageing’ and are specific pathological conditions that require separate treatment. Opponents argue that these diseases are ’intrinsic’ to ’ageing’, seeing them as its inevitable manifestations. The crux of this debate ultimately revolves around which strategy is more suited to identify more effective interventions[8]. The definition itself seemingly predisposes all chronic ’diseases’ to be viewed under the umbrella of ’ageing’. Is there a chronic ’disease’ that does not involve some ’molecular and cellular damage’ that ’ultimately’ contributes to ’mortality’? Under this definition, is it a surprise that factors (to name a few) such as smoking, alcohol consumption, obesity, radiation, excessive sun exposure, poor diet, sleep deprivation, organ transplantation, Down’s syndrome, HIV, Huntington’s disease, Werner syndrome and Sotos syndrome have been linked to ’ageing’? Sure, why not, they all boil down to ’molecular and cellular damage’ that ’ultimately’ lead to ’mortality’. If the definition wasn’t limited to chronic ’diseases’, the term ’ageing’ might contend with the term ’medicine’ due to its all-encompassing scope. These claims warrant discussion as they are directly linked to our initial objectives: evaluating a ’theory’ of ’ageing’ and, more importantly, developing and ’validating’ measures of ’ageing’. For instance, if an observer posits that smoking or obesity accelerate ’ageing’, then a ’valid’ theory would need to predict and explain the impact of such factors. In a similar fashion, this extends to validating ’measures’ of ’ageing’. For example, if an intervention like smoking is assumed to accelerate ’ageing’, a valid measure of ’ageing’ should be able to identify the presence of such a deleterious factor. Yet, the validity of such measures can be contested based on the initial assumptions about what constitutes an ’accelerating’ factor, underscoring the circular reasoning that often pervades this field.

This debate over ’accelerated’ models can be illustrated through a simple example: Consider a laptop, which is designed to operate within specific environmental limits and to perform specific functions. It is far easier to damage a laptop than to enhance its performance, as there are myriad ways to degrade its functionality compared to the more limited improvements. This uneven distribution of negative versus positive scenarios is characteristic of most designed systems. The question then becomes, if the goal is to increase its lifetime while maintaining the same performance, which methods of breaking it provide insights into how it ’normally’ breaks down (i.e., when used within its intended environmental limits)? And thus which methods of breaking it are informative on how to improve it? Placing a laptop in an oven at high temperatures demonstrates the importance of keeping within specific temperature limits but does not necessarily indicate the role or extent of high temperatures in its ’normal’ breakdown. Similarly, using an incorrect power supply highlights the significance of proper power input but does not confirm its impact or the degree of its impact under regular usage conditions. As such, as will be discussed later, it comes down to a question of aetiological similarity between the ’accelerated’ models and the

[8] (and allocated grant money)

'normal' state of interest. Now, as with the argument regarding 'ageing' across species, one might argue that these 'accelerated' models (which are within a species) are a welcome but not necessary requirement for a 'theory' and measure of 'ageing' to meet. In the end, all these linguistic & philosophical debates can be simply settled by examining 'normal ageing' in a single species. After all, a theory of 'ageing' "should" explain at least that right?

2.3 What is 'normal ageing'?

As mentioned earlier, to measure something is to compare it. One cannot measure without the use of a reference and knowing the reference is essential for interpreting any measurement. The use of the term "accelerated" 'ageing' implies a comparison to a standard rate of 'ageing'. This implicit standard is commonly referred to as 'normal' or 'natural' 'ageing'. Given the pivotal role of evolutionary history in molding organisms, the concept of the 'evolutionary optimum state' will frequently emerge in discussions of 'normal ageing'. This idea posits that there exists a set of environmental, dietary, and behavioral conditions that align most closely with our evolutionary history. These conditions are thought to define the 'optimal' state under which "true 'ageing'" would be at a minimum and should therefore serve as the reference point for what constitutes 'normal ageing'. This argument is vividly demonstrated in the discussions of the 'beneficial' effects of regular exercise and dietary behaviors in alleviating many contemporary health issues. From our contemporary reference these interventions appear to 'slow-down' 'ageing', but from the 'optimum state' reference these were abnormal damaged states. This argument can be illustrated through a simple hypothetical: Imagine a world where smoking became universal[9]. In such a world, the 'normal'—or most common—form of 'ageing' would invariably include darkened lungs and a higher incidence of lung cancer at a certain age. However, compared to our reference population, this represents an accelerated form of 'ageing'[10]. Furthermore, we know that removing smoking as a factor would not stop 'ageing'.

This issue extends beyond theoretical debates and directly impacts the evaluation of empirical studies. One of the debates in which this is reflected is the argument over what constitutes a "benign vs detrimental" environment in terms of experimental laboratory conditions, for e.g.:

> "But the classification of an environment as "benign" or "detrimental" depends on the evolutionary history of the population: e.g., can new environments, even if "stress free" be considered benign? Can environments where the population has been long adapting, even if "stressful", be considered "detrimental"? This is not superficial rhetoric, since we have seen recurrent arguments in the literature that defend contrasting expectations for the evolution of aging, relative to the general theory of aging, as a function of the environment/history of the populations" [14].

Another hotly debated example is 'calorie restriction' as a means to slow-down 'ageing'. Critics of these studies argue that the control groups, which are subjected to ad-libitum feeding, represent an 'unnatural' state. They contend that the effects observed in these studies do not truly reflect 'ageing' but rather the consequences of abnormal overeating. Yet, proponents of this study design argue that ad-libitum fed mice offer a 'better' representation of the current human population, and thus the current 'normal ageing'. Ultimately, as will be discussed in the next section, the debate becomes one of etiological similarity; specifically, whether the causes of "accelerated 'ageing'" observed under ad-libitum feeding are informative on the causes of "normal 'ageing'" experienced under the 'optimum state' condition?[11] Ultimately, the conclusion hinges on which underlying assumptions one is willing to accept.

These are but two examples of a widespread debate over what constitutes an adequate control. Although the argument for the 'optimum state' is compelling, it faces significant challenges due to the difficulty of reaching consensus on historical information. This difficulty could be seen through the unending discussions over which diet is "best" given our evolutionary history. Some individuals might advocate for a pragmatic

[9] assuming it is a relatively recent change and not part of the evolutionary history

[10] under the broad and ill-defined definition above

[11] As expected, proponents of the ad-libitum experimental design advocate a "yes", arguing that it is applicable, while those supporting the "optimum state" argue for a "no", claiming it is not representative.

response to these exchanges, arguing that in practice it is unhelpful. For either view point, one has to start somewhere. So why not start with the current “common” ’ageing’ pattern as the reference, explicitly accepting that it might diverge from the ’optimum state’ and that some “slowing” interventions that might be identified are simply a return to that state. Starting with this reference, we can iteratively refine our understanding empirically towards better ’states’, one randomized trial at a time. After all, whether a factor is ’extrinsic’ or ’intrinsic’ to human ’ageing’ doesn’t really matter as long as ’it improves the current situation’. This position does not however put an end to this murky question. Although it clarifies the original question, it still leaves us with a very similar one “What exactly is this reference ’ageing’?”. Answering this involves providing a “definition”. A definition’s function is to attempt to delineate (draw a line/boundary) what qualifies for a given symbol. This is typically achieved by (1) enumerating the objects it applies to, (2) enumerating what it does not apply to, and/or (3) providing a set of requirements an object must meet to qualify. After all to “de-scribe (write about) any thing is to select amongst an infinity of possible features: it is inevitably to circum-scribe (draw a line round) what is salient for the purpose” [15]. This need to define criteria for what constitutes ’normal ageing’ circles back to our initial question of “How to measure ’ageing’”. Many researchers are trying to define ’ageing’ by first seeking a way to measure it. Yet, the validity of any proposed measurement of ’ageing’ remains unassessable because there is no agreed-upon range of phenomena to which the label should apply. This creates a circular problem: researchers are looking for a way to measure ’ageing’ in order to define it, but without a clear definition, they can’t agree regarding the measurement’s validity. This circularity leads to inconsistency, where the same label is applied in contradictory ways. The best example of this is the previously mentioned debate about conflating ’ageing’ with age-associated diseases. Proponents of a strict distinction argue for a version of ’normal ageing’ that excludes these ’age-associated diseases’, viewing them as separate from the ’normal ageing’ phenomena itself. In contrast, opponents of this view consider ’age-associated diseases’ as an integral part of ’normal ageing’. This disagreement underscores a fundamental conflict over how to define the reference point of ’normal ageing’ and what characteristics it should include. The crux of this debate is not merely semantic or philosophical but represents a profound disagreement on the validity of empirical ’ageing’ studies. Proponents advocating for a clear distinction between ’ageing’ and ’age-associated diseases’ require the selection of ’normal ageing’ individuals who have not developed these diseases, in order not to confound results pertaining to ’normal ageing’. Conversely, opponents of this distinction might choose a sample from the general population, irrespective of ’age-associated’ disease status. This is but one example of a broader debate over inclusion criteria for the label of ’normal ageing’. It is a crucial point, particularly given the challenges of detecting small effect sizes in such studies where even minor inconsistencies in the studied population can have significant impacts. Opponents of distinguishing between age-associated diseases and ’normal ageing’ might argue for including individuals with such diseases, as it increases the statistical power and feasibility by amplifying measurable differences. However, proponents of the distinction would then point out that the results are for a disease-specific population and might not be valid to generalize to ’normal ageing’ in disease-free individuals. In much the same way that debates surrounding ’ageing’ and ’accelerated models’ struggle with the lack of clear criteria of what qualifies for these labels, the concept of ’normal ageing’ is as contentious as the etiological validity of the ’accelerated’ models themselves.

2.4 Is ’X’ similar to ’normal ageing’?

All the questions so far, such as when ’ageing’ begins, whether it is ’similar’ across species, or whether ’accelerated’ models resemble ’normal ageing’, are fundamentally questions of ’aetiology’. Aetiology is the study of the causes or origins of a condition. A practical definition of a cause is the ability to predict the outcomes of an intervention [15]. In other words, the aim of using other species or ’accelerated’ models to study ’ageing’ is to make predictions about the effects of interventions on ’normal’ ageing in humans. Given the inherent difficulties of studying ’normal ageing’ in humans, there is great appeal in attempting to establish etiological similarities between ’normal ageing’ and ’accelerated’ models and/or other species. To illustrate the challenges of such etiological comparisons, I will consider the following claim “Apples are similar to strawberries”. This claim’s validity depends on the observer’s choices in mapping these symbols to concepts and back to reality and the choice of projection to test it (section 4). This claim can be valid, if one considers that both are eadable. On the other hand, its validity is in question if one was to choose their volume. One might argue that additional details will help avoid this silly problem for example “Apples are similar to strawberries in terms of their color”. However, even for the same choice of concept and projection family (i.e. co-varying subset of measurements), a choice on the level of resolution of the measurement (manifested in the specific projection) would still need to be agreed upon. For example, if we measure ’dominant color’, both apples and strawberries might be classified as ’red’. However, this classification can be contested if one defines ’color’ at a finer scale for e.g. by a specific wavelengths or if one was to argue regarding the “degree” of difference that would be eligible for it to be considered “similar”. Furthermore, and of even larger importance, what can be contested is the choice of projection itself.

Most people—I hope—would agree on the self-evident and obvious nature of this plain and silly example. And yet, many of the persistent aetiological ’debates’ in the field of ’ageing’ mirror this question.

- **Is ’ageing’ similar across species?**
  This question - which was briefly mentioned in the wasp example - manifests in various forms, for example “Is ageing universal?”, “Do different species age at different rates?”, and “Which species do not age?”. As with the silly example above, the answer to these questions depends on the choice of projection and resolution. If one measures ’ageing’ as the age-dependent increase in mortality at a population level, then it becomes straightforward to compare this measure across species and argue for the universality of ’ageing’. Another commonly used projection is ’mobility’ which allows for e.g. to compare worm ’ageing’ and human ’ageing’ as both exhibit a decrease in ’mobility’ with age. However, as the resolution of these projections is increased, this general claim becomes more contested. One can mirror this with a silly example comparing the deterioration of a building and a laptop. Over time, both experience an increase in ’problems’, suggesting a similarity in their deterioration. Yet the question is whether studying the deterioration of the laptop can help us deal with the deterioration of the building. This effect of resolution and choice of projections is clearly evident in the stark contrast between consensus on the definition of ’ageing’ at the population level versus the individual level:

  > “The definition of aging at the actuarial or population level is reasonably well agreed (see Comfort, 1979; Kirkwood, 1985; Finch, 1990; Patridge and Barton, 1996). Aging is defined in terms of its negative effects on age-specific survival and fecundity. However, the more difficult problem is to define aging in terms of its physiological effects in individuals. Many aspects of the phenotype alter with aging. Furthermore, there is great variability in many of these changes among individuals within the population” [16].

- **Is ’ageing’ similar within species?**
  This question manifests in debates about the similarity or divergence in aetiology between ’normal ageing’ and ’accelerated ageing’ models[12]. Proponents of similarity argue that accelerated ’ageing’ models, such as progeria or Werner syndrome, offer a condensed version of the broader, slower processes we observe in typical ageing scenarios. These models, they suggest, mirror the ’fundamental biological mechanisms’ at play but at ’a hastened pace’. Conversely, critics of this viewpoint contend that accelerated ’ageing’ models might represent distinct pathological states rather than accelerated versions of ’normal ageing’. They point out that the specific damage seen in these conditions often involves specific genetic mutations or environmental stresses that do not universally occur in ’normal ageing’. Given the degrees of freedom afforded to both proponents and opponents in selecting observations, phenotypes, and the granularity of comparison, each camp justifies their respective arguments. The debate could be conclusively settled by showcasing an intervention that is efficacious in both ’accelerated’ and ’normal ageing’ contexts (i.e. the ability to predict the outcome of the intervention in both populations). Yet, due to the significant financial and temporal investments needed, the discourse frequently drifts into philosophical explorations, with contributors freely expounding their varied opinions. And if a trial is executed and the results are not as expected, discourse typically pivots to critiques concerning the appropriateness of chosen endpoints

[12] as both qualify to the label of ’ageing’ under the common broad ill-definition of the term

or to the study design itself—illustrating the concept of experimenter's regress, which will be discussed later.

- **Is 'ageing' deterministic or stochastic?**
  It is perhaps one of the most enduring debates in the field, manifesting in various forms such as the argument over longevity versus 'ageing' genes, and the notion of whether ageing is genetically predetermined. In this debate, as in others, the extensive leeway both sides have in defining terms (with 'ageing', deterministic, and stochastic definitions being notably broad), choosing observations, interpreting data, and selecting the level of resolution for comparisons turns the discussion into a contest of assumption preference. As succinctly expressed by (Rose, 1991) "In general, these ambiguities of definition have probably been critical in keeping alive a metatheoretical concept that is either wrong, when defined precisely, or trivial, when defined broadly [17]"
- **Do different tissues 'age' at different rates?**
  This question vividly underscores the influence of projection selection on the interpretation of these broad claims. If 'replicative capacity' was adopted as a criterion for the rate of ageing, both neuronal and muscle cells would seemingly fail this assessment early in life. On the other hand, selecting 'mitochondrial activity' as the metric would result in red blood cells and the eye lens not meeting the criteria. Given the vast differences in shape, function, environment, and behavior among various cell types, any attempt to compare their 'ageing' processes necessitates a selective focus on specific criteria. This selection inevitably influences the outcomes of such comparisons. This issue has driven the popularity of more general measures such as DNA mutations or telomere shortening. Yet, even these broadly applicable metrics are fraught with interpretation challenges. The diverse environments and internal architectures that different cells are exposed to necessitate that any attribution of 'cause' in ageing will inevitably be contentious, confounded by the cell's identity. A clear manifestation of this problem are the claims regarding the 'immortality' and absence of 'ageing' of cancer cells, due to their perceived ability to replicate indefinitely. However, are these cells not characterized by a rapidly changing genetic identity i.e. 'genomic instability'? Under this alternative projection, is it possible to contend that these cells are experiencing 'ageing'? [18]

This notion of measurement resolution can be articulated through the contrast of abstraction versus concreteness. Abstraction and concreteness represent opposite ends of a continuous scale of resolution. As abstraction increases, the resolution decreases, shifting the focus to broader patterns. Conversely, concreteness enhances the resolution, revealing more specific and detailed patterns. This contrast is also mirrored in model building under the labels of underfitting (abstraction) vs overfitting (concreteness) [15]. Echoing previous discussions on model validity and reliability, assessing whether a model underfits or overfits hinges on having a defined objective. Such an objective enables the evaluation of whether a model's resolution appropriately matches its intended purpose.

We began with the query of how to measure 'ageing'. This raised the issue of the underlying objective of the measurement, as without it one would not be able to evaluate the validity of the measurement model. The standard response is the grandiose goal: "To study the processes of 'ageing' and to identify interventions to slow down or reverse 'ageing'". This is akin to answering the question, "Why do you want to measure an apple?" with a vacuous response like, "To study the apple and identify things that affect it". Affect what, exactly? Its color? Its size? Its taste? How many can fit in a given box? The response is devoid of detail that it could mean anything. Yet, the severity of this vagueness is even more pronounced when one considers what the term 'ageing' represents. The current trendy definition is so all-encompassing that it permits its users to adopt incompatible assumptions and diverging evaluation criteria. The symptoms of these foundational conflicts manifest in the lack of consensus over which theories have been 'disproven' and in the perpetual debates over ill-defined questions. A more significant issue is that this all-encompassing definition exacerbates challenges inherent in the study of complex organisms.

3. How to study a complex system?

The study of complex systems presents a number of inherent challenges, which are exacerbated in the field of 'ageing' due to the timeframes involved as well as the ambiguous and inconsistent terminology. In what follows, I will discuss some of these challenges and explore their implications.

3.1 Causation attribution & theory building

A practical definition of a cause is the ability to predict the outcomes of an intervention. This means, given a simple causal structure of three variables — A -> B -> C (**Figure** 3-A) — we can construct a causal model that enables predictions about various interventions and their effects. Specifically, such a model would allow us to predict:

- the effect an intervention on A would have on B
- the effect an intervention on A would have on C, through the intermediary B
- the effect an intervention on B would have on C
- an intervention on B would have no effect on A
- that intervening on B can block the effect of A on C

These model predictions can then be experimentally tested to verify the degree of validity of the assumed structure among these variables. The conditional independencies, such as the observation that an intervention on B has no effect on A, are as crucial as the dependencies, like the effects of interventions on A or B on C. These aspects allow us to test the directionality of effects among the variables. Furthermore, confirming these independencies and dependencies can guide effective decision-making in scenarios where interventions are feasible.

Next, let us consider how this maps to a typical diagram of 'elements' of interest in theories of 'ageing'. Given the structure of biological systems, especially at the cellular level - which is the focus of a large number of 'modern' theories of 'ageing' - it is characterized by interconnectedness among its components, recursiveness, and feedback loops. Such a diagram is illustrated in Figure 2-B. In this representation, there is a bi-directional connection between each of the 'elements'. This signifies that each element can influence all others, either 'directly' and/or 'indirectly' through other 'elements'. Crucially, this suggests that intervening at any single point in the network would propagate effects throughout, impacting all other elements to varying extents. This interconnected structure of biological systems allows for a vast degree of freedom in theory building and the selection of observations, making it difficult to dispute any particular theory. For example, one can hold this diagram from the node "mitochondria" and argue for the "mitochondrial theory of 'ageing'" whose supporters would argue has great explanatory power because 'damage' to the mitochondria can lead to damage to all the other components and can give rise to the 'characteristic features' of 'ageing' [19]. Another would hold the diagram by the node "metabolism" and argue that selection for slower metabolism through various mechanisms can then lead to changes in all the other nodes and thus is giving rise to the 'features' of 'ageing' [20]. Others might focus on the 'epigenome', and argue for the 'information theory of ageing', claiming that disruptions in the epigenome trigger all other forms of cellular damage, positioning it as the central driver of 'ageing' [21].

If one is familiar with these as well as other cellular 'theories', the pattern is evident, where the argument typically follows a common structure:

- 'X' is essential for life because damage to 'X' is lethal to the organism.
- Problems with 'X' can cause problems with other 'elements'.
- Problems with 'X' result in the 'characteristic' features of 'ageing'.
- Intervention 'Y' in model organism and/or cell line 'Z' that enhances the function of 'X' slows down or reverses certain features of ageing.
- *Optional: 'X' is conserved across species!*

One can easily replace 'X' with any number of genes, processes, or abstract terms and push a 'theory' of 'ageing', for e.g., DNA, protein homeostasis, stress response, protein folding, mitochondria, lysosome, metabolism, immunity, telomere, transposable elements, information, free radicals and so on. The interconnected structure of biological systems allows supporters of one theory to argue that effects observed by other theories' favored interventions are merely confounded through their impact on the particular element that their theory prioritizes. This is especially exacerbated at the cellular level as blocking different elements is difficult and where establishing proximal and ultimate causes is challenging. For instance, proponents of the epigenome theory might argue that the beneficial effects of an intervention targeting the mitochondrial are, in fact, due to downstream effects on 'epigenome stability'. Similarly, advocates of the metabolic theory might claim that improvements observed through mitochondrial interventions are actually mediated by shifts in metabolic processes. And since the system is highly interconnected, supporters of any given theory can almost always find some evidence to bolster their claims. The interplay between these large degrees of freedom—ranging from the definition of terms, selection of assumptions, and choice of measurements, to the inherent complexity and recursive nature of biological systems—creates fertile ground for endless theorizing. This environment enables individuals to selectively interpret data, weaving narratives that reinforce their theoretical frameworks while shielding them from experimental refutation.

3.2 Characterization & the fallacy of composition

As explored earlier, any real element possesses a range of measurable features or "projections", each informing on different aspects of it. These projections, however, are not isolated from one another; they exhibit varying degrees of correlation. In some cases, knowing one projection can provide insight into another. As systems grow more complex, the number of projections increases, and the relationships among them can become more variable. In complex systems, the risk of the fallacy of composition becomes pronounced. This fallacy occurs when one incorrectly assumes that properties of individual parts can be extrapolated to the whole system. A clear example was mentioned in the wasp example. It was observed that haploid and diploid wasps had similar average lifespans, leading to the claim that both groups aged similarly. However, proponents of the somatic mutation hypothesis would argue that this conclusion commits the fallacy of composition: it assumes that because their lifespans (a part of 'ageing') are similar, the entire 'ageing' process behaves similarly in both groups. They would point out that lifespan is just one projection of 'ageing', and other aspects — such as genetic stability, mobility etc. — may differ significantly.

This issue ties into the wider problem of determining how to properly characterize interventions, especially in relation to 'normal ageing' clinical trials. As the saying goes, "there are no solutions, only trade-offs" [22]. Any intervention comes at a price. This is especially the case in complex systems with interdependent parts designed to function within specific parameters, improving one feature often necessitates sacrificing another. The critical task is to identify this cost and and evaluate whether the benefits justify it. The problem is that these trade-offs are not easily predictable due to the system's complexity. Unintended consequences can arise in numerous, often unforeseen, ways. To identify these trade-offs, one must observe the system for a "sufficient" duration and with "enough detail"; otherwise, crucial effects might be missed. While more information is generally useful, collecting it always comes at a price. Every additional variable measured comes at a cost to the limited resources available—from funding to time. Especially in long-term studies, the more data we attempt to gather, the less logistically feasible it becomes. As such, one has to prioritize a set of measurements to focus on otherwise it is unfeasible. This leads us to an essential question: "What is the minimal set of features that we should measure?" The answer to this question is dependent on the objective and the specific inferences one intends to make. If one wants to make inferences about 'ageing', we are once again confronted with the initial challenge of selecting the appropriate set of measurements. One can imagine two extremes. At one extreme, if all features are perfectly correlated, measuring just one would suffice. Knowing this single feature would reveal the values of the others, making the characterization process much simpler. The complexity is reduced to a single dimension, allowing us to generalize across all aspects of 'ageing'. At the other extreme, none of the features are correlated—they are entirely independent. In this case, knowing one feature gives no insight into the others, and we would need to measure each one individually to better understand the system. This scenario presents a harder characterization problem. To evaluate where a suggested 'ageing' measurement set lies on this continuum — and thus assess the validity of claims regarding 'ageing' — one must first have a clear definition of what this label actually represents. This leads us back to the issue of the lack of consensus regarding the "whole" of 'ageing', which allows different researchers to operate under divergent assumptions. Each researcher can rely on their own interpretation of 'ageing', and when faced with contradictory findings, they may fall back on the fallacy of composition by debating whether the results are 'valid' for the undefined and inconsistent concept of the 'whole'. As we will discuss later, the best solution to this problem is to avoid overgeneralizing from a specific set of measurements to the broad, vague concept of 'ageing'. However, this comes at a cost — it limits the ability to make sweeping

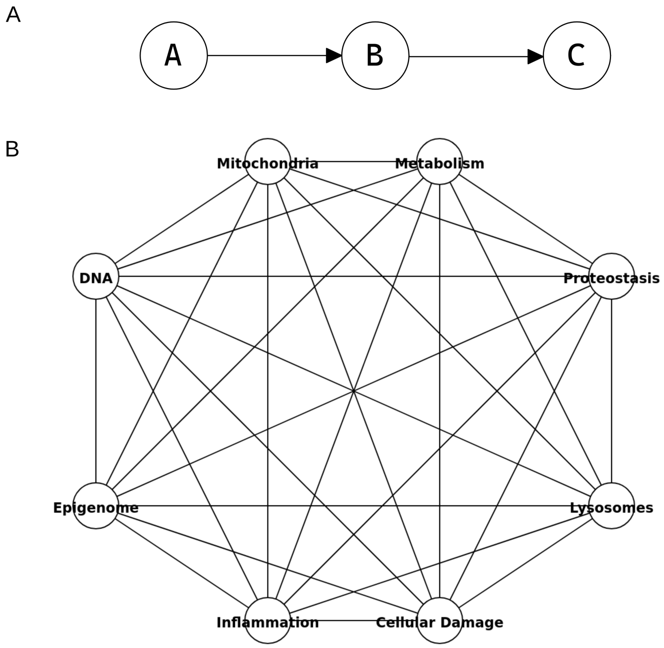

Figure 3. (**A**) A simple causal structure. (**B**) A diagram illustrating the interconnected elements typically employed in cellular theories of ageing.

claims about "slowing" or "accelerating" the 'ageing' process, which often attracts attention and funding.

The most glaring symptoms of these problems manifest in the persistent rebranding and the emergence of new terms designed to "better capture" what 'ageing' is supposed to represent. One of these fashionable terms in recent years was "healthspan", defined as "the period of life spent in good health, free from the chronic diseases and disabilities of ageing" [23]. Yet, the term offers only a modest improvement over 'ageing' by explicitly defining it as "a period of time", pushing the argument to what constitutes "good health" or the "disabilities of ageing", while leaving all other key assumptions unaddressed. As such individuals, armed with their own agendas, readily mold "healthspan" to fit their preferred models, picking and choosing the elements that align with their hypotheses. And, as with the term 'ageing', the "... imprecise definitions of healthspan have likely contributed to controversies related to rapamycin in mice (Johnson et al. 2013; Neff et al. 2013; Richardson 2013) and mutation of the insulin-like receptor DAF-2 in C. elegans (Bansal et al. 2015; Ewald et al. 2018; Hahm et al. 2015)" [23]. Leading Matt Kaeberlein to note that, "[u]ntil such time as a comprehensive healthspan metric is adopted, it would seem prudent to refrain from using the term healthspan in the scientific literature, except as a conceptual construct" [23].

Yet "healthspan" is but one example of a broader strategy of branding and presentation. A standout example is the ever-evolving array of names for 'measurement of ageing', which has cycled through 'biomarker of ageing', 'functional age', 'measurement of senescence' and now shows up as 'biological age' or 'ageing clock'. Another prominent case is the adoption of the "hallmarks of ageing", a 'framework' that has gained significant traction in the past decade. It was presented with the aim of bringing clarity to a field riddled with ambiguity. Yet, as it was eloquently stated

> "... the aging hallmarks account takes a somewhat arbitrary set of popular ideas from the aging field and, seemingly, dresses them up as a paradigm, even though a genuine paradigm as present in the hallmarks of cancer account does not exist in aging. This resembles an exercise in mimicry: as the hoverfly mimics the wasp to fool predators into believing that it has a sting, the hallmarks of aging puts on a resemblance to the hallmarks of cancer, to give the impression of a paradigm where one does not exist" [24].

What this 'framework' effectively achieves is an act of diplomacy — giving a nod to every player in the field, ensuring no one's pet theory, favorite ’biomarker’, or preferred animal and/or accelerated model is left out in the cold. This framework is actually a nice illustration of the causation attribution problem previously discussed. Each hallmark—for e.g. mitochondrial dysfunction, genomic instability, or epigenetic alterations—can be framed as both a cause and a consequence of the others, creating a circular narrative flexible enough to support any story. This flexibility stems from the structure of cellular systems, but is significantly amplified by the broadness of the terms themselves. After all, what is meant by ’genomic instability’ exactly? DNA mutation? Aneuploidy? DNA tangling? Chromosomal rearrangements? Transposable element activation? Mitochondrial DNA depletion? Microsatellite instability? Double-strand breaks? Oxidative DNA damage? The term is broad by design, allowing for e.g. both somatic mutation theorists and transposable element theorists to align their hypotheses with the framework. And can one not consider “telomere attrition”, presented as its own hallmark, merely another form of genomic instability? One can come up with numerous terms that would fulfill the same function for e.g. “cytoplasmic disruptions”, “organelle instabilities”, “nuclear alterations” or “extracellular matrix alterations”. It's the intellectual equivalent of saying, “things go wrong with the stuff inside the cell”. It's an attempt to repackage the familiar — “consequences of life, like cellular and molecular damage” — using different terminology, with all the problems carried over. For instance, this framework suffers from the same issue over aetiology of what is an accelerated model of ’ageing’. Its broad and encompassing terms allow virtually any intervention to be linked to ’ageing’. After all, what intervention doesn't impact at least one of these ’elements’ - “genome stability”, “epigenome state”, “mitochondrial function”, “intracellular communication”, “proteostasis”, or “nutrient sensing” - in at least some projection? If the change of the intervention is positive, then you have found an intervention that slows a ’feature’ of ’ageing’, if it is negative, then you have found an “accelerated ’ageing’ model” that purportedly sheds light on ’normal ageing’! This is where the hallmarks shine, not in their explanatory power, but in their sheer adaptability. Whether the results are positive or negative, ’ageing’ is invoked either way, turning any experiment into a contribution to the ’ageing’ field. All roads, it seems, lead to ’ageing’.

3.3 Experimenter's regress

This concept is best illustrated with a thought experiment. Consider these two claims: (1) Partial reprogramming can reverse ’ageing’, and (2) a methylation clock measures ’ageing’. Now, consider the following experiment: A control group and a treated group, where the treatment involves “partial reprogramming” for a specific duration and doze optimized by prior studies. The read-out is methylation “age” - established by previous studies - measured before and after the treatment for both groups. If the result shows no significant change in the methylation “clock” for the treated group compared to the control, what can be inferred? The interpretation will depend on the observer's assumptions regarding the two starting claims:

- If the observer believes in the rejuvenating effect of partial reprogramming but is skeptical of the methylation clock, they can argue that the experiment shows the methylation clock does not measure ’ageing’ accurately and that more adequate ’ageing’ measures are needed.
- If the observer believes in the clock measuring ’ageing’ but is skeptical of partial reprogramming, they can argue that partial reprogramming does not affect ’ageing’.
- If the observer believes in both the rejuvenating effect of partial reprogramming and in the clock measuring ’ageing’, they can argue against the study design, for e.g., the treatment was not performed “as it should” (longer time, different doze, etc.) or that the methylation clock was measured in the wrong tissue, etc.

This cycle of interpretation and reinterpretation based on prior beliefs is at the heart of the concept known as “Experimenter's regress”. The term was coined by sociologist of science Harry Collins and refers to the idea that the validity of an experiment's outcome is often contingent on the assumptions of the observer interpreting the results [25]. Essentially, if the experiment doesn't produce the expected results, one can always question the experimental design, the choice of measurements or the underlying assumptions. In the case of our thought experiment, each observer's interpretation of the experiment's results reflects their prior commitments to the validity of the partial reprogramming technique in reversing ’ageing’, the validity of the methylation clock in measuring ’ageing’, or both. However, this scenario is not unique to partial reprogramming or the methylation clock. The same pattern of interpretation could emerge with any other intervention or measurement technique. Whether the intervention in question is caloric restriction, pharmacological treatment, or genetic modification, and whether the measurement tool is telomere length, protein aggregation, or some other ’biomarker’, the underlying problem remains the same.

While the concept of experimenter’s regress is not unique to biology, the nebulous and all-encompassing definition of ’ageing’ in this field significantly amplifies the problem. Without a consensus on what constitutes ’ageing’, any measurement technique or experimental design is prone to endless reinterpretation. This increase in degrees of freedom manifests at multiple levels of the research process. Researchers may select from a wide array of measures, physiological functions, or lifespan data, each of which could be argued to represent ’ageing’ depending on the underlying assumptions. This subjectivity extends to the resolution of these measurements — whether the focus is on molecular changes, cellular processes, or whole-organism outcomes — as well as the choice of organism for the experiments. This flexibility in defining ’ageing’ not only perpetuates the cycle of experimenter’s regress but also makes it difficult to reach a consensus on what constitutes a ’proper’ experiment in the field.

4. So, how to measure ’ageing’?

In his book on probability theory, The statistician Bruno de Finetti declared: “PROBABILITY DOES NOT EXIST”. His argument was that probability should be understood not as an objective reality, but as a conceptual tool - a model - used to describe uncertainty from the viewpoint of an observer who lacks complete information [15]. Many concepts and models might be said not to ’exist’ in any concrete sense. After all, by design, they are inherently abstractions, distilled from numerous instances of reality. This abstraction is what makes them useful across a range of instances. Just as a map is not the territory it represents, these models are not the reality they describe but rather conceptual tools. As such, models should be assessed based on their utility over a set of given objectives, rather than whether they ’exist’ in a literal sense.

The goal of this article was to examine the utility of the current concept of ’ageing’ in addressing a range of objectives, particularly the measurement of ’ageing’, the evaluation of the validity of ’ageing’ theories, the evaluation of interventions, and the assessment of the validity of experiments conducted in the ’ageing’ field. For these objectives, the current concept of ’ageing’ proves to be more of a hindrance than a help. It attempts to encompasses all species, from single-celled to multi-celled organisms, plants to animals, and any harmful or ’beneficial’ interventions within them. When a concept tries to do too much, it ends up achieving very little. This overextended concept muddies both theory and experiment. To evaluate a theory, there must be consensus on the observations it aims to explain and the predictions it is meant to test. However, in the case of ’ageing’, the set of observations is ill-defined and largely left to the discretion of individuals. This lack of clarity allows the target to be endlessly reshaped, letting the phenomena a theory aims to explain be adjusted to fit any need. Each ’theorist’ has a wide range of observations spanning species to pick and choose from. If a candidate theory shows positive results in one species, the claim that ’ageing is universal’ is often invoked to reinforce its relevance. On the other hand, if challenged by observations in another species, it’s often dismissed by claiming that ’ageing’ varies across species, leaving the possibility that the theory still holds for humans. Similarly, each theorist can draw from a wide array of ’accelerated’ interventions in humans and model organisms, selectively embracing those that support their theory while dismissing others as not ’true’ examples of ’accelerated ageing’. And if, even after sifting through the vast array of species and accelerated models, a theory still comes up short, it can always retreat to the claim that it at least explains ’normal human ageing’ — a concept that remains itself ill-defined and heavily debated. Given the long timespans involved in ’normal human ageing’, the vast variability across individuals, and the numerous physiological changes observable at different resolution levels, this provides a broad set of observations from which each theorist can selectively focus on what suits their narrative. This overextended concept also undermines the evaluation of experimental results. Disagreements over implicit assumptions and the choice of metrics give researchers a wide array of options to selectively interpret or dismiss experimental findings, depending on whether they align with their favored theories. When a result challenges a theory, proponents can easily argue that the wrong metrics were used — after all, since the 'whole' of 'ageing' remains undefined, no set of metrics is ever sufficient. This flexibility means that any unfavorable outcome can be explained away by claiming the experiment didn't capture the 'true' or 'complete' aspect of 'ageing'. These excessive degrees of freedom in the selection of measurements, observations, and assumptions exacerbate challenges inherent to the study of complex systems from the causation attribution problem to the experimenter's regress.

If the current 'concept' of 'ageing' is so detrimental, then why has it persisted for so long? Surely, a concept that garners such popularity and generates so much literature must be serving some function. As discussed in the case of the hallmarks framework, the real allure lies in its intellectual malleability: whether the results are positive or negative, something can always be tied back to 'ageing'. It's a catch-all that thrives on ambiguity, offering just enough structure to sound scientific while remaining vague enough to avoid scrutiny. Furthermore, by the time any 'ground truth' about human ageing emerges, most of these claims will be safely beyond scrutiny—shielded by the sheer timescales involved in human studies. The temporal, financial, and ethical challenges of long-term human experiments provide a convenient buffer, allowing 'theorists' to maintain their claims with little fear of direct contradiction. Even if such studies are eventually done, the same well-honed strategies — word definitions, selective metrics, flexible assumptions and inadequate study design — will be ready to reinterpret any inconvenient results.

The problem of ill-defined, all-encompassing concepts is neither unique to 'ageing' research nor particularly new. History provides numerous examples where similarly vague terms dominated various fields, often slowing progress. One prominent example is the concept of 'nature', as critiqued by Robert Boyle in 'A Free Enquiry Into the Vulgarly Received Notion of Nature' [26]. Boyle critiqued the anthropomorphizing of 'nature', where it was invoked not merely as a descriptive term but as an active force or agent governing phenomena. Nature was used as a label associated with phenomena spanning various fields—from physical concepts like the resistance to vacuums and the motion of celestial bodies, to biological processes such as the growth of plants and the healing of wounds. Boyle's warning of 'nature' resonates here: This tendency to overuse and overextend a concept not only prevents meaningful scrutiny but also creates a false sense of understanding — one that hinders, rather than advances, the scientific pursuit of truth [26]. This underscores the subtle yet crucial role language plays in shaping scientific understanding. Although not strictly essential, the scientific method relies heavily on language to describe observations, develop hypotheses, design and document experiments, and report findings. Word choice and the way a question is framed carry implicit assumptions that can alter the approach taken and the conclusions drawn. A poorly defined or overly broad question introduces ambiguity at every stage of the scientific process—from selecting metrics to interpreting results. This is why the adage "asking the right question is half the solution" holds particularly true in scientific research - a well-posed question sets clear boundaries for what is being investigated and establishes concrete criteria for success or failure. For example, instead of asking, "Does partial reprogramming reverse 'ageing'?" and/or "Does a methylation clock measure 'ageing'?", one could ask a more 'valid' question: "Can a methylation clock distinguish between a partial reprogramming-treated group and a control group?" This question has a clear objective—determining whether the methylation clock can differentiate between treated and untreated groups—based on a concrete ground truth: group labels. Framed this way, the experiment's results can clarify if, and to what extent, the methylation clock is a valid model for the explicit objective. By avoiding the vague term 'ageing' and focusing on the clear ground truth of group labels, it also mitigates the experimenter's regress by limiting interpretive flexibility. While the regress may still exist, its influence is significantly reduced. This approach further avoids the fallacy of composition, sidestepping the ill-defined and all-encompassing concept of 'ageing'. The obvious downside of this phrasing is that it forfeits the allure of broader claims that might seem more significant but rest on unsubstantiated ground — such as grand conclusions about 'slowing the ageing process', which often attract attention and funding. Yet, since the model's validity and reliability are objective-dependent, such claims would not be supported by the current experiment in any case. With such re-framing, the criteria for measurement become tied to more explicit goals, rather than trying to capture an ill-defined and constantly shifting notion of ’ageing’. Furthermore, this avoids the circularity of using a measurement to define a concept and then using that same concept to validate the measurement.

One should take heed of the lessons from the past decades. Numerous attempts to refine the definition of ’ageing’, have ironically led to an even looser and more relaxed version of it, in a desperate attempt to reach some form of consensus. Standardizing a term requires sacrificing some of its original uses [27], and the broad range of applications for ’ageing’ has made it impossible to create a single definition that satisfies everyone. In its current state, this symbol is at best a label for a field of study covering countless biological phenomena, much like the word ’physics’ serves as an umbrella term for everything from quantum mechanics to cosmology. If we take the term as such, an answer to the question of how to measure ’ageing’ might be: the number of publications per year. After all, that’s one metric that has seen continuous growth, thanks in large part to the vagueness of the term itself. But I suspect this isn’t quite what those in the field have in mind when they ponder the question of ’measuring ageing’. Ultimately, knowing to identify and avoid poorly framed questions is as critical as finding answers.

Correspondence

If you find any mistakes or have certain comments, I would appreciate it if you can let me know at Ignophi_mail@pm.me.

References

[1] Arshag D. Mooradian. Biomarkers of aging: Do we know what to look for? *Journal of Gerontology*, 45(6):B183–B186, 1990. doi: 10.1093/geronj/45.6.B183.

[2] Mina Moqri, Christopher Herzog, John R. Poganik, Biomarkers of Aging Consortium, Julie Justice, Daniel W. Belsky, Andrew Higgins-Chen, Alexey Moskalev, Georg Fuellen, Alan A. Cohen, Ivan Bautmans, Martin Widschwendter, Jian Ding, Andrea Fleming, Joan Mannick, Jing-Dong J. Han, Alex Zhavoronkov, Nir Barzilai, Matt Kaeberlein, and Vadim N. Gladyshev. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, 2023. doi: 10.1016/j.cell.2023.08.003.

[3] Paul Lockhart. *Measurement*. Harvard University Press, Cambridge, MA, 2012.

[4] David L. Streiner, Geoffrey R. Norman, and John Cairney. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press, USA, 5th edition, 2015.

[5] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

[6] Richard Peto and Richard Doll. There is no such thing as aging: Old age is associated with disease, but does not cause it. *BMJ*, 315(7115):1030–1032, 1997.

[7] Alan A. Cohen, Veronique Legault, and Tamas Fülöp. What if there’s no such thing as “aging”? *Mechanisms of Ageing and Development*, 192:111344, 2020. doi: 10.1016/j.mad.2020.111344.

[8] José Viña, Consuelo Borrás, and Josep Miquel. Theories of ageing. *IUBMB Life*, 59(4–5):249–254, 2007. doi: 10.1080/15216540601178067.

[9] Piotr Chmielewski. Rethinking modern theories of ageing and their classification: The proximate mechanisms and the ultimate explanations. *Anthropological Review*, 80(3):Article 3, 2017. doi: 10.1515/anre-2017-0021.

[10] Björn Schumacher, Joris Pothof, Jan Vijg, and Jan H. J. Hoeijmakers. The central role of dna damage in the ageing process. *Nature*, 592(7856):695–703, 2021. doi: 10.1038/s41586-021-03307-7.

[11] A. M. Clark and M. A. Rubin. The modification by x-irradiation of the life span of haploids and diploids of the wasp, habrobracon sp. *Radiation Research*, 15(2):244–253, 1961.

[12] Vadim N. Gladyshev. The ground zero of organismal life and aging. *Trends in Molecular Medicine*, 27(1):11–19, 2021. doi: 10.1016/j.molmed.2020.10.002.

[13] Nigel R. Balcombe and Alan Sinclair. Ageing: definitions, mechanisms and the magnitude of the problem. *Best Practice & Research Clinical Gastroenterology*, 15(6):835–849, 2001. doi: 10.1053/bega.2001.0244.

[14] Margarida Matos. A question never comes alone: Comments on 'what is aging?'. *Frontiers in Genetics*, 3:Article 150, 2012. doi: 10.3389/fgene.2012.00150.

[15] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, FL, 2018.

[16] Thomas B. L. Kirkwood. Alex comfort and the measure of aging. *Experimental Gerontology*, 33 (1–2):135–140, 1998. doi: 10.1016/s0531-5565(97)00114-9.

[17] Michael R. Rose. *Evolutionary Biology of Aging*. Oxford University Press, New York, NY, 1991.

[18] Leonard Hayflick. The illusion of cell immortality. *British Journal of Cancer*, 83(7):841–846, 2000. doi: 10.1054/bjoc.2000.1296.

[19] Lee Know. *Mitochondria and the Future of Medicine: The Key to Understanding Disease, Chronic Illness, Aging, and Life Itself*. Chelsea Green Publishing, White River Junction, VT, 2018.

[20] Jack Wordsworth, Poul Yde Nielsen, Esther Fielder, Sundar Chandrasegaran, and Darren Shanley. Metabolic slowdown as the proximal cause of ageing and death. 2023. doi: 10.1101/2023.08.01.551537. URL `https://doi.org/10.1101/2023.08.01.551537`. Preprint.

[21] Jing-Hui Yang, Mami Hayano, Patrick T. Griffin, Joao A. Amorim, Michael S. Bonkowski, John K. Apostolides, Evan L. Salfati, Mathieu Blanchette, Elizabeth M. Munding, Manda Bhakta, Yee C. Chew, Wendy Guo, Xiaoping Yang, Samuel Maybury-Lewis, Xiaoyu Tian, Jason M. Ross, Giuseppe Coppotelli, Michal V. Meer, Riley Rogers-Hammond, and David A. Sinclair. Loss of epigenetic information as a cause of mammalian aging. *Cell*, 186(2):305–326.e27, 2023. doi: 10.1016/j.cell.2022.12.027.

[22] Thomas Sowell. *The Vision of the Anointed: Self-Congratulation as a Basis for Social Policy*. Basic Books, New York, NY, 1995.

[23] Matt Kaeberlein. How healthy is the healthspan concept? *GeroScience*, 40(4):361–364, 2018. doi: 10.1007/s11357-018-0036-9.

[24] David Gems and João Pedro De Magalhães. The hoverfly and the wasp: A critique of the hallmarks of aging as a paradigm. *Ageing Research Reviews*, 70:101407, 2021. doi: 10.1016/j.arr.2021.101407.

[25] Harry M. Collins and Trevor Pinch. *The Golem: What You Should Know About Science*. Cambridge University Press, Cambridge, UK, 1998.

[26] Robert Boyle. *Robert Boyle: A Free Enquiry into the Vulgarly Received Notion of Nature*. Cambridge University Press, Cambridge, UK, 1996.

[27] Sahotra Sarkar. *Genetics and Reductionism*. Cambridge University Press, Cambridge, UK, 1998.

[28] Hilary Putnam. The meaning of "meaning". In *Mind, Language, and Reality: Philosophical Papers, Volume 2*, pages 215–271. Cambridge University Press, Cambridge, UK, 1975.

Appendix - On word models

In this section, I will apply the modeling concepts introduced earlier to words. As mentioned earlier, a model is a set of elements and probabilistic relationships among them. Under this definition, a word is a model, as it can be viewed as the set encompassing the relationship between a symbol and a probability distribution of concepts induced in the recipient's mind (**Figure** 4-A). This distribution represents the frequency of a concept's induction after exposure to a specific symbol. For example, it can be at the level of a specific individual (i.e. exposure of a given individual to the same symbol multiple times), or across a population of individuals (i.e. exposure of different individuals in a population to the same symbol).

This metaphoric representation of a word has practical applications beyond mere linguistic and aesthetic indulgence. On one hand, it would allow one to apply modeling concepts such as validity and reliability to words. On the other hand, this representation would allow to understand many of the properties and limitations of words.

To evaluate the 'validity' of a word, one must consider the inferences made by its user. The typical 'objective' of the user is the transfer of information. Put another way, the user is making an 'inference' about the concepts induced in the recipient as a result of using a specific symbol. (In what follows, to avoid repetitiveness, when I refer to "a word's validity", it will be assumed that this is the objective). Similarly, the recipient of a word is making an 'inference' about the concepts the user of the word wishes to induce in their mind. Given this objective for using words, the 'validity' of a word hinges on both its user and its recipient.

To evaluate the 'reliability' of a word, one must consider the facet(s) over which the inference is generalized. If that facet is the 'recipient', the question is, is a word inducing similar concepts across these recipients? If that facet is 'time', the question is, is a word inducing similar concepts across time in the same recipient? For example, a user may revisit their own writings in the future. Are the induced concepts similar? If it is across multiple facets, such as 'recipient' and 'time' (i.e. generalizability), the question is, is a word inducing similar concepts across time and recipients? For example, consider two different individuals, one in 1800s and one in the 2000s, reading a book written by a third individual in 1800s. Are the induced concepts similar? The 'reliability' of a word hinges on the facet(s) over which its user is generalizing the inference.

For a given word, comparing its distribution of induced concepts between two individuals is informative of that word's validity between them. The larger the overlap between the 2 distributions the more valid the word is (**Figure** 4-B). i.e., more effective communication. Both the users of the word and the recipient, can infer what is being induced in the mind of the other by using that given word. The inference's strength is a function of the overlap among their distributions. To the other extreme, if there is no overlap, inference regarding the induced concepts is not possible and so is communication (**Figure** 4-B). From this, one might also expect higher validity when communicating with oneself compared to others, as the distribution of induced concepts might generally be more similar for the same individual, assuming not too much time has passed.

In a similar fashion, for a given population of individuals, considering the distribution of induced concepts for two different words can be informative about their relative validity across that population and on the words' relationships (**Figure** 4-C). Intuitively, the narrower a word's distribution, the higher its validity on average across all combinations of individuals within this population. For example, consider the following two words: "1" and "beautiful". "1" is likely to have a narrower distribution as it consistently represents a specific concept, whereas "beautiful" may have a wider distribution due to its varying interpretations and associations across different individuals. This higher validity means that when the word "1" is used, there is greater certainty that the user will infer the intended concept, whereas the word "beautiful" might lead to a broader range of interpretations and thus lower inferential power. Furthermore, if a word's induced distribution is a subset of another word, then it can be inferred that within the considered population, the first word is a hyponym of the second (**Figure** 4-C).

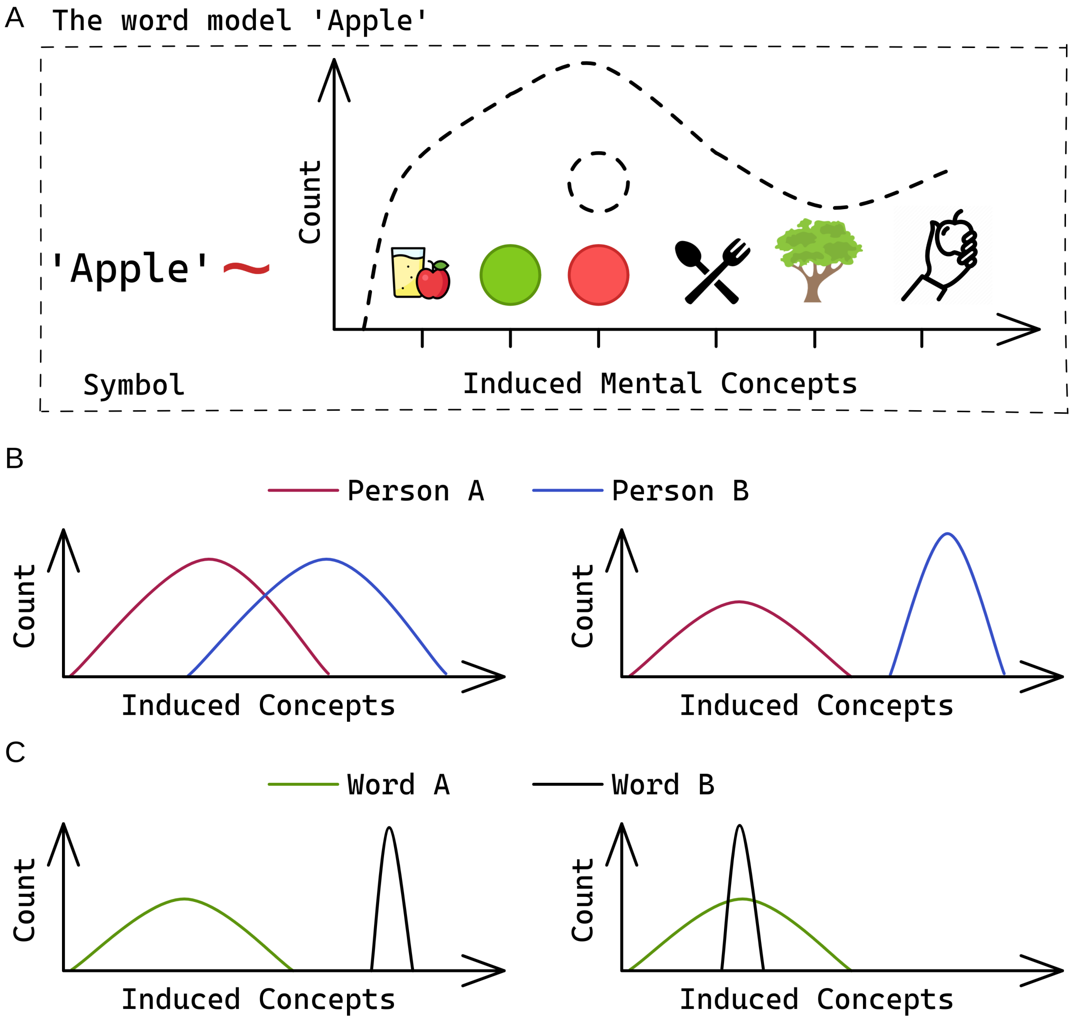

Figure 4. Word model metaphor. (**A**) A model of a word as a relationship between a symbol and a probability distribution of concepts induced in the recipient. (**B**) Illustration of the validity of a given word model. For a given word, comparing its distribution of induced concepts between two individuals is informative of that word's validity between them. The larger the overlap between the 2 distributions the more valid the word is, i.e., more effective communication. Both the users of the word and the recipient, can infer what is being induced in the mind of the other by using that given word. The inference's strength is a function of the overlap among their distributions. To the other extreme, if there is no overlap, inference regarding the induced concepts is not possible and so is communication. (**C**) Comparison of word models. for a given population of individuals, considering the distribution of induced concepts for two different words can be informative about their relative validity across that population and on the words' relationships. Intuitively, the narrower a word's distribution, the higher its validity on average across all combinations of individuals within this population.

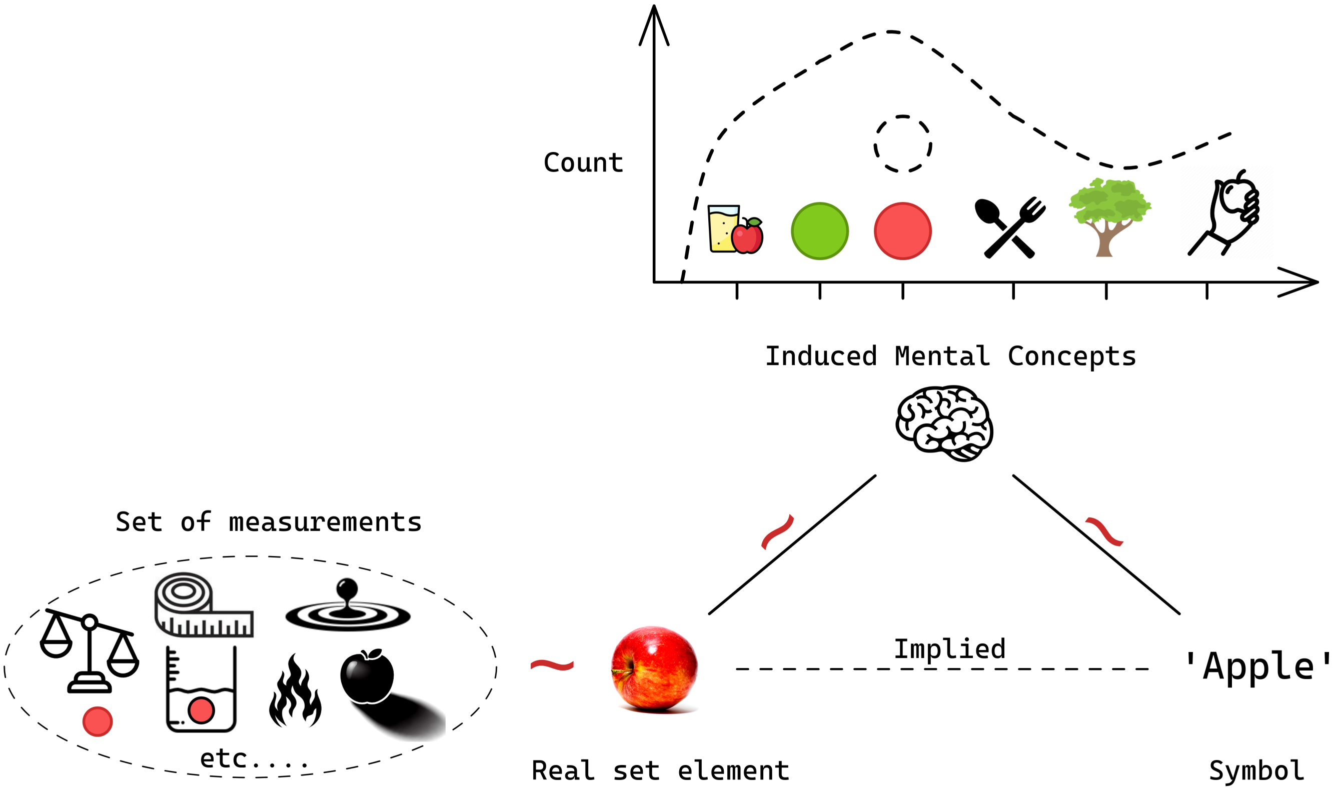

Figure 5. Illustration of the degrees of freedom available when mapping between word models and the set of measurements associated with a real set element. Inspired by an illustration from [28].