Understanding Consciousness as Data Compression

Phil Maguire[1], Philippe Moser[1], and Rebecca Maguire[2]

[1]Department of Computer Science, National University of Ireland, Maynooth
[2]School of Business, National College of Ireland
{pmaguire, pmoser}@cs.nuim.ie, rebecca.maguire@ncirl.ie

In this article we explore the idea that consciousness is a language-complete phenomenon, that is, one which is as difficult to formalise as the foundations of language itself. We posit that the reason consciousness resists scientific description is because the language of science is too weak; its power to render phenomena objective is exhausted by the sophistication of the brain's architecture. However, this does not mean that there is nothing to say about consciousness. We propose that the phenomenon can be expressed in terms of data compression, a well-defined concept from theoretical computer science which acknowledges and formalises the limits of objective representation. Data compression focuses on the intersection between the uncomputable and the finite. It has a number of fundamental theoretical applications, giving rise, for example, to a universal definition of intelligence (Hutter, 2004), a universal theory of prior probability, as well as a universal theory of inductive inference (Solomonoff, 1964). Here we explore the merits of considering consciousness in such terms, showing how the data compression approach can provide new perspectives on intelligent behaviour, the combination problem, and the hard problem of subjective experience. In particular, we use the tools of algorithmic information theory to prove that integrated experience cannot be achieved by a computable process.

Keywords: *Consciousness, hard problem, data compression, artificial intelligence, integrated information, combination problem, qualia, scientific standards, algorithmic information theory*

***Journal of Cognitive Science* 17-1: 63-94, 2016**
Date submitted: 12/05/15 Date reviewed: 01/28/16
Date confirmed for publication: 03/31/16

1. Introduction

According to Koch (2004), the ultimate goals in studying consciousness are first, to understand the neural mechanisms that underlie personal experiences, and, second, to develop techniques for quantifying this process objectively in humans and animals. In recent years there have been numerous attempts to address consciousness following the "formalise and quantify" approach advocated by Koch (e.g. Baars, 2007; Dennett, 1991; Hameroff, 1998). Progress is promised by new techniques in neuroscience which allow precise neural correlates of cognition to be isolated (Aru, Bachmann, Singer & Melloni, 2012; Koubeissi et al., 2014). And yet, no matter how closely neural activity is scrutinized, science seems to get no closer to unravelling the mystery of how such activity is transformed into first person experiences. Strawson (2011), for example, describes Dennett's attempts to objectify the subjective element of consciousness as "the craziest claim that has ever been made in the history of human thought".

Some philosophers, sensing the depth of the problem, have suggested that the explanatory gap between subjective consciousness and the objective world is insurmountable (e.g. Chalmers, 1996; McGinn, 1991), which leaves little to say that could be informative. However, this resignation towards naturalistic dualism is not one we support. In this article we make the link with algorithmic information theory (AIT), a field of mathematics which focuses specifically on the limits of representation. We argue that discoveries in AIT can provide a structure which allows a deeper perspective on the problem of consciousness.

2. Algorithmic Information Theory

In 1936 Church and Turing, building on Gödel's work, clarified the concept of universal computation. In brief, the idea is that recursive symbol manipulation is so powerful that any effective method can be expressed in terms of an algorithm specified for a single universal machine: instructions and data can be stored in the same format, thus enabling the concept of a stored-program computer. This idea also implies that it is possible for humans to appreciate, indeed precisely define, the existence of problems

whose solutions cannot be reached by any effective method.

In practice, because these results concern Platonic ideals (i.e. involving infinite computational resources of space and time), we do not witness direct implications of uncomputability in everyday life. However, a field called algorithmic information theory (AIT), pioneered by Solomonoff, Kolmogorov and Chaitin in the 1960s, focuses on the interaction between the finite world of objective description and Church and Turing's idea of uncomputable problems. According to Chaitin, AIT is "the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously" (Calude, 2013). Whereas information theory can be regarded as a theory of communication, AIT might be regarded as a theory of 'representability', addressing the relationship between information and computation that underpins systems of representation.

A remarkable result in AIT is the identification of a universal lower semi-computable semi-measure. A semi-measure is a function that provides a distribution for the set of all strings, such that the sum comes to less than 1. A semi-computable function is one that can only be computed from a single direction by successive increasing approximations (see Li & Vitányi, 2013). For example, as we continue to compute the output, our upper bound gets lower and lower, bringing us closer and closer to the right answer. And yet, we can never complete the computation: even if we have not managed to bring the upper bound down any further for a very long time, we can never rule out the possibility that further computation will cause it to drop again. Thus, a function which is only semi-computable reflects a process that is inexhaustibly difficult: we continue to get closer and closer without ever getting there for sure.

The existence of a *universal* lower semi-computable semi-measure entails that there is a function for assigning probabilities to strings that is universal, insofar as it predicts just as well as any other such semi-measure, up to a multiplicative constant. In other words, there exists an ultimate notion of 'difficulty', one that dominates all other possible definitions.

As it happens, there are multiple different ways in which this same mathematical object arises, suggesting that "the captured notion has an inherent relevance that transcends the realm of pure mathematical

abstraction" (Li & Vitányi, 2013). Perhaps the most intuitive of these is the idea of *data compression*. When we reduce the size of a file by pulling out patterns, the file gets smaller and smaller. However, we can never terminate the process and say "this is as small as it gets", because we can never successfully run all of the programs that might pull out another pattern. Some of the programs we're checking will be ones that never halt, but we have no way of separating them from the set of programs which will successfully halt after a very long time. Data compression is as difficult a process as we can define. At the limit it is uncomputable, and the closer we get to that limit, the more difficult it becomes to identify any further redundancy (Chaitin, 2006).

2.1 Applications of data compression

Due to its universality, data compression lends itself to a range of theoretical applications, such as providing the grounding for the idealized theory of inductive inference developed by Solomonoff in the 1960s (see Li & Vitányi, 2013; Rathmanner & Hutter, 2011). The fundamental premise of this theory concerns the connection between likelihood and parsimony (Chater & Vitányi, 2003). Solomonoff (1964) assigns universal a priori probabilities to hypotheses based on the length of their shortest descriptions, and then updates these weights following Bayes' theorem to make optimal predictions. This idea can be interpreted as a formalisation of Occam's razor, the idea that, all being equal, simple theories should be preferred because they are more likely (Chater & Brown, 2008). Kirchherr, Li and Vitányi (1997) argue that it successfully ties up "a couple of millennia of philosophy".

According to Rathmanner and Hutter (2011), "Solomonoff created a completely general theory of inductive inference. Subsequent developments have shown that his system solves many of the philosophical and statistical problems that plague other approaches to induction. In the same/similar sense as classical logic solves the problem of how to reason deductively, Solomonoff solved the problem of how to reason inductively."

While Solomonoff induction serves as the gold standard to aim for, its semi-computability means that it can only be approximated in practice (see

Hutter, 2007). Minimum description length (MDL) modelling (Rissanen, 1978) provides a more practical means of reaching an inference, restricting the set of allowed codes, and identifying the best hypothesis as the one that provides the greatest compression of the data. Vitányi and Li (2000) show that, in general, "data compression is almost always the best strategy, both in model selection and prediction...the better a theory compresses the data concerning some phenomenon under investigation, the better we have learned, generalized, and the better the theory predicts unknown data". All successful predictive systems, including plants and animals, are approximations of an ideal data compressor. As an expression of the ultimately difficult process, data compression allows us to formalise the foundations of prediction, explanation, communication, and language itself. And thus, we argue, it must be able to tell us something about consciousness.

2.2 Data Compression as understanding

At first blush the concept 'data compression' seems esoteric, a niche idea related to information storage in computer science. However, we can see that the concept runs much deeper. Data compression occurs when information is bound together through the identification of shared patterns. For example the sequence 4, 6, 8, 12, 14, 18, 20, 24... can be simplified as the description "odd prime numbers +1". The latter representation is shorter, hence we can say that it has been 'compressed'. Following Occam's razor, the elegance and concision of this representation suggests that it is the true underlying pattern which governs the sequence. Someone who manages to identify this pattern might claim to have 'understood' the sequence, because, with high probability, they can predict the numbers that follow.

Data compression is not just something that happens when files are reduced in size on a computer. Because of its connection to induction and prediction, compression can be viewed as providing reliable proof of understanding. According to Chaitin (2006), "A useful theory is a compression of the data; compression is comprehension". The higher the level of compression that is achieved, the better a system's predictions will be, and the greater the extent to which it can be said to 'understand' the

data.

Despite the universality and utility of this mathematical concept, many remain unaware of it. According to Rathmanner and Hutter (2011): "Although it achieves excellent theoretical results and is based on solid philosophical foundations, the requisite technical knowledge necessary for understanding this framework has caused it to remain largely unknown and unappreciated in the wider scientific community..."

In the following sections we explore the idea that it may prove fruitful to regard consciousness in terms of a system's capacity to compress data (i.e. to comprehend; to predict; to do what is ultimately difficult). We begin by looking at the question of intelligence, before turning to the combination problem and, finally, the hard problem of subjective experience.

3. Intelligence as compression

Intuitively, there appears to be a relationship between intelligence and consciousness. Animals that pass the mirror self-recognition test, such as elephants, chimps, bonobos, orang-utans, dolphins and killer whales, are among those we consider the most intelligent. Quantifying intelligence might thus be viewed as the first hurdle that any account of consciousness should clear. As we will see, this is something that data compression achieves very convincingly.

Based on the principle of data compression, an enhanced version of the Turing test for machine intelligence (Turing, 1950) has been established for which the challenge is to compress, to the greatest extent possible, 100 megabytes of textual information drawn from Wikipedia (see Hutter prize; Legg & Hutter, 2007). Hutter (2009) clarifies: "This compression contest is motivated by the fact that being able to compress well is closely related to acting intelligently, thus reducing the slippery concept of intelligence to hard file size numbers. In order to compress data, one has to find regularities in them, which is intrinsically difficult. So compressors beating the current "dumb" compressors need to be smart(er)."

Hutter (2009) also explains why data compression should be regarded as a universal measure of intelligence: "Intelligence has many faces, like creativity, solving problems, pattern recognition, classification, learning,

induction, deduction, building analogies, optimization, surviving in an environment, language processing, knowledge, and many more. A formal definition incorporating all or at least most aspect of intelligence is difficult but not impossible. Informally, intelligence is an agent's ability to achieve goals in a wide range of environments...One can prove that the better you can compress, the better you can predict; and being able to predict [the environment] well is key for being able to act well."

A potential flaw of the Turing test is that a program might pass the test simply by exploiting weaknesses in human psychology. If a given system passes the test, we cannot be sure if it was because of the quality of the responses or the gullibility of the judge (French, 2012). Hutter's compression test, by contrast, is more reliable because, at the limit, data compression is uncomputable, and the closer you get, the harder it becomes.

A 100MB Wikipedia file will contain many complex patterns that represent a broad spectrum of human thinking. Seeking to compress Wikipedia is equivalent to designing an algorithm that can write new Wikipedia pages, ones which fit seamlessly with the rest of encyclopaedia (see Li & Vitányi, 2013). These arguments reinforce the suitability of the Hutter prize as a reliable measure for quantifying intelligence, and its putative status as an AI-complete challenge.

3.1 Arguments that data compression is not intelligent

There are several intuitive criticisms that spring to mind regarding the compression-based approach to modelling human cognition. One is that the human brain does not compress information in the same manner that a computer program archives data. For example, people are forgetful, they 'chunk' information together and only remember those details perceived as important. Unlike off-the-shelf compressors, humans do very poorly at reproducing text verbatim, yet are much better at recalling general ideas, which can be conveyed in different words (e.g. Schacter, Guerin & Jacques, 2011). In other words, an important aspect of human intelligence is knowing what to forget and when to forget it.

As it turns out, such arguments are not relevant. A file that is compressed in a lossy fashion can be viewed as a string that has been separated into

two components, noise and signal, or alternatively, data and program. The process of separating these two, and identifying what information can be safely forgotten, requires lossless data compression. As stated by Hutter (2009): " lossless compression is still the right way to go ...noise does not at all harm the strong relation between compression and understanding / intelligence / predictability".

Another common objection to the compression-based approach is that "while humans are better than computers at speech recognition, language translation, reading, answering questions, etc., humans cannot compress text efficiently" (Mahoney, 2009). Granted, machines appear better suited to pulling patterns out of data than humans. Off-the-shelf compressors, such as Lempel-Ziv, can quickly take gigabytes of data and compress them down to a fraction of their original size. However, in this case computers are using weak modelling and very fast deterministic computation (i.e. brute force) to pull ahead of humans.

Machines are faster, they are cheaper, more reliable, more durable, they can hold greater memory. These attributes give them a short-term edge. However, in the longer run, humans have the ability to discover alternative mechanisms for performing the same task even more quickly. Any machine that implements a hard-coded data compression program, no matter how fast, is going to eventually be defeated by humans, since the latter can always innovate a superior compression algorithm.

Ultimately, compressing text is not just about brute force speed. Computer programs struggle to identify the underlying *meaning* of the text and hence require longer descriptions to encode it, while people are very good at identifying patterns that link words and sentences together. As a result, humans can do better than machines at the Hutter compression test. Shannon's (1951) estimate for the entropy of English suggests that the 100MB of data used for the Hutter Prize could be compressed down to 12 MB by a human, if they spent enough time at the task. In contrast, the highest level of compression that has been achieved to date by any program, with a €50,000 prize on offer from Hutter, is 15.9 MB (Hutter, 2009).

3.2 Compression and cognition: empirical evidence

In order to thrive in an uncertain environment, organisms must be able to anticipate future events. AIT tells us that the more efficiently an organism can compress its experiences, the more accurate its predictions will be (Vitányi & Li, 2000). As a result, organisms have evolved brains which are prodigious compressors of information: compressing sensory information provides them with an understanding of their environment, allowing them to optimize their decision making.

Research in the area of artificial intelligence and cognitive science is increasingly identifying data compression as a key organisational principle. Schmidhuber (1992) pioneered the idea of using predictive coding to allow recurrent neural networks to compress observation streams, an early precursor of what has since evolved into the idea of very deep learning machines (see Schmidhuber, 2013). Wolff (1993) identified a link between cognition and AIT, pointing out that the storage and processing of information in computers and brains, from the recognition of objects to the use of natural language, can be understood in terms of information compression.

Chater and Vitányi (2003) have argued that data compression should be considered as a unifying principle in cognitive science. They suggest that much of perception, learning and high-level cognition involves finding 'sensible' patterns in data. It is the simplicity, or parsimony, of these patterns which supports their predictive power.

Schmidhuber (2006, 2009) proposes data compression as the simple principle which explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music and jokes. He argues that data becomes temporarily interesting once an observer learns to predict (i.e. compress) it in a better way, making it subjectively simpler and more 'beautiful'. From this perspective, curiosity can be viewed as the desire to create and discover patterns that allow for compression progress, with the level of interestingness being related to the effort required. According to Schmidhuber, this drive for compression motivates exploring infants, mathematicians, composers, artists, dancers and comedians, as well as artificial systems.

In a similar vein, both Maguire et al.'s (2013) theory of subjective information and Dessalles' (2011) simplicity theory view data compression as a key explanative construct in the phenomenon of surprise. When people experience a stimulus which is expected to be random, yet turns out to be compressible, a surprise response is triggered by the brain. Unanticipated compressibility suggests the existence of a previously undetected pattern, resulting in an urgent representational updating process. Maguire et al. (2013) suggest that people often rely on compressibility rather than probability to judge likelihood and make decisions in real world situations.

Adopting the perspective of the mind as a compressor, Gauvrit, Zenil and Tegnér (2015) connect AIT to experimental observations in the areas of working memory, probabilistic reasoning and linguistic structures. They argue that the concepts of data compression and algorithmic complexity provide an important normative tool which can shed light on a broad range of cognitive processes, from language use to the interpretation of EEG and fMRI data. Zenil, Marshall and Tegnér (2015) also show that algorithmic complexity can be used to validate results from the behavioural analysis of animals, including foraging communication by ants, flight patterns of fruit flies, and tactical deception and competition strategies in rodents.

Casali et al. (2013) suggest that a signature of conscious processing lies in the complexity of brain activity patterns distributed among interacting cortical areas. They perturbed the cortex of various subjects with transcranial magnetic stimulation to engage distributed interactions in the brain, and then compressed the spatiotemporal patterns of the electrocortical responses using Lempel-Ziv to quantify their algorithmic complexity. The compression measure was found to reliably discriminate levels of consciousness during wakefulness, sleep, anaesthesia and minimally conscious states.

In sum, there is growing empirical evidence that intelligent behaviour and cognition can be modelled in terms of data compression carried out by the brain. But what of first-person experiential consciousness? In the following section we propose that, like intelligence, unitary behaviour can be viewed in terms of data compression. Specifically, we propose that a system's behaviour is interpreted as unified when it performs sophisticated data compression that is difficult to reverse.

4. The Combination Problem as Compression

Subjective experience seems to carry the qualitative characteristic of being unified and singular, a property which is at odds with the reducibility of the physical world. How can the brain, whose processing is clearly distributed, give rise to a consistently integrated perspective? As articulated by James (1890): "Take a sentence of a dozen words, and take twelve men and tell to each one word. Then stand the men in a row or jam them in a bunch, and let each think of his word as intently as he will; nowhere will there be a consciousness of the whole sentence." We propose that combination is manifested, not by any local physical convergence or supernatural process, but through the process of data compression.

Research in neuroscience has shown that information in the brain is encoded in a highly compressed state (Rolls & Treves, 2011). When the brain compresses information in this way, it is binding data together through the identification of shared patterns that were originally dispersed in space and time. The process of grouping and encoding these patterns yields computational results which reflect a 'coming together' of information. We will now explore the idea that the interpretation of a system as being unified is a characterisation of its processing as being computationally difficult to reverse and disintegrate.

4.1 United through cooperation

The reproductive success of an organism is dependent on cooperation between all of its constituent components, leading to a form of compression which unites data distributed across space and time. For instance, it does not make sense for an organism's legs to maintain independent agendas. Because the interests of both legs are intimately bound, it is more productive for them to cooperate with each other in achieving a single set of objectives (e.g. walking). Accordingly, the brain sources sensory information from all over the body and unites it through compression, thereby optimising predictive accuracy for the organism as a whole. Tactile information from every limb is compressed alongside visual information from the eyes and aural information from the ears, giving rise to a form of understanding that

is centralised and representative of the organism's experiences as a singular unit. The resulting decisions of the organism also appear centralised: to the external observer it seems as if the organism's body is being 'controlled' by a single entity.

The success of an organism also depends on cooperation through time. Accordingly, the response it exhibits to a sensory stimulus depends not just on its immediate processing, but also on its memories. Patterns in a current stimulus are matched against patterns distilled from historical stimuli, leading to a form of understanding that combines not only distributed sensory organs but also an organism's past and present states (see Maguire & Maguire, 2010).

If the behaviour of a system is too complex to be broken down, then the only way to predict its actions is to treat it as a unified system. Thus, when data compression is sufficiently sophisticated so as to be irreversible in practice, it forces external observers to adopt the intentional stance (see Dennett, 1991), and treat the system as if it was enjoying unitary experiences (see Bringsjord & Zenzen, 1997).

The adoption of the intentional stance clearly does not apply to off-the-shelf compressors. These compression schemes, such as Lempel-Ziv, are trivially reversible. Similarly, a robot that optimizes its behaviour using a unsophisticated compression algorithm can have its various information sources 'unbound' and analysed separately. The most accurate predictions of a robot's behaviour are achieved when it is viewed as a mechanical automaton without any singular perspective. Thus, we do not regard robots as conscious.

Leibniz expressed a similar sentiment back in 1686, with his famous image of the mill: Consciousness, he said, "cannot be explained on mechanical principles, that is, by shapes and movements. Imagine that there is a machine whose structure makes it think, sense and have perception. Then we can conceive it enlarged, so that we can go inside it, as into a mill. Suppose that we do: then if we inspect the interior we shall find there nothing but parts which push one another, and never anything which could explain a conscious experience" (Strawson, 2011).

Leibniz's thought experiment centres around disintegration: the idea of reducing a supposedly conscious system into separate components, and then

inspecting those components individually. In contrast, our proposal is that conscious systems are precisely those whose sophisticated data compression precludes such a deconstruction, preventing us from seeing them in terms of separate machine-like components.

According to our perspective, the dividing line between conscious and automatic behaviour is one of complexity. The binding of information is not something that takes place in any absolute sense, it is instead *interpreted* as having taken place relative to the difficulty of predicting a system's behaviour. In the case of humans, and other intelligent creatures, the sophistication of the data compression carried out is too complex to reverse in practice, hence we are *forced* to adopt the intentional stance. We regard a system as conscious only when it becomes too complex, and thus unprofitable, to regard it as an automaton.

4.2 Integrated information theory

The above account of binding through data compression bears close resemblance to Tononi's integrated information theory, which also construes consciousness in terms of sophisticated information processing (Tononi, Sporns & Edelman, 1994; Tononi, Edelman & Sporns, 1998; Tononi, 2004, 2008, 2012, 2015). Tononi proposes that consciousness can be quantified in terms of the complexity of a system's organisational structure, specifically its capacity to 'integrate' information.

According to Tononi (2008), what we mean when we say that the human brain produces consciousness is that it integrates information deeply, thus producing behaviour that is hard to reduce back into its original informational constituents. Tononi (2008) explains the foundations of his theory through two thought experiments, which we adapt below. The first thought experiment establishes the requirement for a conscious observation to generate information. The second establishes the requirement for a conscious observation to be incorporated with previous memories, hence generating *integrated* information.

4.2.1 Requirement 1: Generating information

Imagine that a factory producing scented candles invests in an artificial smell detector. The electronic nose is used for sampling the aroma of the candles passing on the conveyor belt below and directing them to the appropriate boxes. Let's suppose that the factory is currently producing two flavours of scented candle: lemon and lavender. In this case the detector only needs to distinguish between two possible smells.

A batch of lemon scented candles is passed underneath and the sensor flashes *lemon*. Can we say that the detector has actually experienced the smell of lemon? Clearly it has managed to distinguish lemon from lavender, but this does not guarantee that it has experienced the full aroma that humans appreciate. For example, it may be the case that the electronic nose is latching onto a single molecule that separates the two scents, ignoring all other aspects. The distinction between lemon and lavender is a binary one, and can thus be encoded by a single bit. In contrast, humans can distinguish more than 10,000 different smells detected by specialized olfactory receptor neurons lining the nose. When humans identify a smell as *lemon*, they are generating a response which distinguishes between 10,000 possible states, yielding $\log_2 10,000 = 13.3$ bits of information.

Tononi's (2008) first thought experiment highlights the idea that the quality of an experience must be expressed relative to a range of alternative possibilities. For instance, if the whole world was coloured the same shade of red, the act of labelling an object as 'red' would be uninformative. Descriptions of experiences must be situated within a context where they discriminate among many alternatives (i.e. they must generate information).

4.2.2 Requirement 2: Generating integrated information

Tononi's (2008) second thought experiment establishes that information alone is not sufficient for conscious experience. Information must also be integrated.

Imagine that the scented candle factory enhances the electronic nose so that now it can distinguish between 1 million different smells, beyond even human ability. Can we now say that the detector is truly smelling lemon when it outputs *lemon*, given that it is producing more information than a human? What is the difference between the detector's experience and

human experience?

Like the human nose, the electronic nose uses specialized olfactory receptors to diagnose the signature of the scent, and then looks up this signature in a database to identify the appropriate response. However, each smell is responded to in isolation of every other. The exact same response to a lemon scent would occur even if the representations of the other 999,999 smells were deleted from the database. The factory might as well have purchased a million independent smell detectors and placed them together in the same room, with each unit independently recording and responding to the candles on the conveyor belt. An unintegrated set of responses does not yield a subjective experience. To bind the repertoire, the system must generate *integrated* information. Somehow, the response to the smell of lemon must be encoded in terms of its relationship with all other possible experiences.

4.3 Consciousness as integrated information

Inside the human nose there are different receptors which are specialized to respond to particular smells. This process of detection is not itself integrated. For example, with selective damage to certain olfactory receptors, a person could conceivably lose their ability to smell lemon while retaining their ability to smell lavender. It is only when a smell signal affects cognition that it becomes integrated.

According to Tononi's (2008) theory, when we perceive lemon the effect that it has on our brain is integrated across many aspects of our memory, in a way that external observers find hard to disentangle. Let's consider, for example, a subject who has just experienced the smell of lemon. According to the integrated information theory, the changes caused by her olfactory experience are not localized to any one part of her brain, but are instead widely dispersed and inextricably intertwined with all the rest of her memories, making them extremely difficult to reverse. As a result, it would prove very difficult to operate on this subject's brain and eliminate her recent memory without affecting anything else.

In contrast, deleting the same experience from the memory of an artificial smell detector would be easy. Somewhere inside the system is

a database with discrete variables used to maintain the detection history. These variables can simply be edited to erase a particular memory. The information generated by the electronic nose is not integrated. It does not influence the subsequent information that is generated. It lies isolated, detached and dormant.

4.4 Quantifying integration using edit difficulty

Like our theory of binding through data compression, Tononi's (2008) integrated information theory implies that people attribute consciousness to systems whose processing is complex and hard to reverse. Schmidhuber (2014) is critical of the theory's lack of parsimonious elegance. He suggests that a simpler and more general view would be to express consciousness as a feature that emerges naturally from data compression during problem solving. However, Maguire et al. (2014) show that, for information lossless processes, Tononi's quantification of integration is *equivalent* to data compression. Both of these theories are, in essence, saying the same thing. They view consciousness, not as an objective phenomenon which can be analysed independently of the world, but as a quantification of the sophistication of an information processor relative to its environment.

Building on the convergence between integration and data compression, we propose the formal notion of 'edit difficulty' to quantify consciousness. Edit difficulty expresses how deeply a given stimulus has been integrated within a set of memories through data compression. For example, it expresses the difficulty a neurosurgeon would face in seeking to edit a given memory within someone's brain. The more compressed the memory, the more sophisticated the predictions it supports, and the more difficult it is to edit.

Consider, for example, the case of an uncompressed data file featuring complex patterns, such as the 100MB Wikipedia file used in the Hutter test. In its original state, every bit has independent significance: if we destroy 10% of the data, we lose at most 10% of the Wikipedia file. In contrast, when the same file is compressed to the limit, each bit in the final representation is fully dependent on every other bit for its significance. Wiping out the first bit would corrupt everything, leaving only a 50%

chance of getting all the bits right and a 50% chance of getting them all wrong: there would be no way of guessing whether the first bit was a 1 or a 0, because both guesses would make just as much sense. In this instance, the significance of the first bit has been totally integrated with all of the other bits through the process of data compression; there is no remaining redundancy.

Imagine now seeking to edit the first word in a Wikipedia page that has been compressed like this. Where is this word encoded in the compressed file? There is no easily delineated set of bits which corresponds to the first word and nothing else. Instead, the whole set of data has been integrated, with every bit from the original file depending on all the others. To discern the impact that the first word has had on the compressed encoding, the compression scheme would have to be completely reversed. There are no shortcuts. Hence, we can see that the more that data is compressed, the more difficult it becomes to edit.

To formalise this idea we consider a stimulus, first in its raw unintegrated state, and second, encoded in its integrated state. The level of integration is equivalent to the difficulty of identifying the raw information and editing it within its integrated state.

In the following definition z and $f(z)$ are the raw stimulus and the encoded stimulus respectively. We consider the difficulty of editing z into z', for example, editing the smell of lemon to turn it into the smell of lavender. If this operation is performed on a raw, unintegrated dataset, then the task is straight-forward: the bits that differ are simply altered. Consider, however, the challenge for the neurosurgeon operating on a subject's brain. If the stimulus has not been widely integrated, then the neurosurgeon can concentrate on a single localised area of the brain. The encoding will be overt, reflecting the original unintegrated format in which the information was originally transmitted. However, if the stimulus has been successfully integrated then its encoding will be widely distributed, with effects on all kinds of other memories, making it effectively impossible to isolate and edit. The edit difficulty is so great that the subject's current brain state is largely useless for identifying a target edited brain state.

We quantify the integration of an encoding process operating on a stimulus as the minimum informational distance between the current state

and any possible edited state. If every edited state is completely unrelated to the current one, then the integration is 1; if there exists an edited state which is only trivially removed from the current one, the integration is 0.

Formally, the edit difficulty of f for stimulus z is a number between 0 and 1 that measures the level of integration of $f(z)$. It is measured by looking at all strings z' similar to z, and finding the one that minimizes the ratio of length of the shortest description of $f(z)$ given $f(z')$ to the length of shortest description of $f(z)$. The smallest ratio obtained is the edit difficulty. Since the numerator is always positive and less or equal to the denominator, the edit difficulty is between 0 and 1.

In the following definition, $C(x)$ is the length of the shortest program x^* such that a universal Turing machine U on input x^* outputs x. Thus, $C(x)$ is the amount of algorithmic information contained in x. For two strings x, y the conditional Kolmogorov complexity $C(x \mid y)$ of x given y is the size of the shortest program p such that U on input p, and provided y as an extra input, outputs x.

Let $f : \{0,1\}^* \rightarrow \{0,1\}^*$ be a 1-1 function. The edit difficulty of f for stimulus z is the smallest number of bits needed to produce $f(z)$ given the description of $f(z')$, where z' is a stimulus similar to z, which has been edited in some way. The value is normalised by dividing it by the length of the shortest description of $f(z)$, i.e.

$$ed_f(z) = \min\{\frac{C(f(z) \mid f(z'))}{C(f(z))} : z \neq z', C(z \mid z') \leq \log \mid z \mid\}$$

4.5 On the computability of integration

The above formalisation of edit difficulty captures the essence of Tononi's integrated information theory in an intuitive and parsimonious manner. Tononi (2008) views information as integrated when "it is not decomposable into a collection of causally independent parts", which is exactly what edit difficulty measures. Furthermore, when edit difficulty is high, the whole necessarily contains more information than any of its minimal parts, since none of those parts considered in isolation is sufficient for editing any memories. This matches Tononi's (2008) idea that integrated information is

"the amount of information generated by a complex of elements, above and beyond the information generated by its parts".

What kind of systems can integrate information in this way? We now prove an interesting result, namely that there is no computable function that can integrate information to even the slightest degree: the process of irreversible information binding is not one that can be objectively described.

In the following proof a function is said to be integrating so long as its edit difficulty is bounded away from zero for all its inputs.

Definition 1 *A 1-1 function f is integrating if there exists $\epsilon > 0$ such that for all inputs z, $ed_f(z) \geq \epsilon$.*

The following result shows that no integrating function can be computed. In fact, even partial computable functions (that may not halt on all inputs) fail to compute an integrating function.

Theorem 1 *Let f be an integrating function and let φ be a partial computable function with infinite domain. Then φ fails to compute f on its domain, i.e. there exists a string x such that $\varphi(x)$ halts but $\varphi(x) \neq f(x)$.*

Proof. By contradiction, let f, φ be as above and suppose φ computes f on its domain. Let $\epsilon > 0$ be such that the edit difficulty of f exceeds ϵ on every input. Since φ is partial computable, its domain $\mathrm{dom}(\varphi)$ is a computably enumerable set. Without loss of generality we can assume we have a computable enumeration of it, where exactly one string enters the enumeration at each stage s. Consider the following partial computable function $\psi(x)$ which searches for the (unique if it exists) u and minimal s such that $\varphi(u)$ halts after s steps of computation and $\varphi(u) = x$. If u is found, get v, that is, the string enumerated in $\mathrm{dom}(\varphi)$ at stage s (if $v = u$ get the next string in the enumeration), and output $\varphi(v)$. This ends the description of ψ.

If x is in the range of φ, i.e. $x = \varphi(u) = f(u)$ for some unique u, then $\psi(x)$ halts and outputs $\varphi(v) = f(v)$ for some v, and we have $C(f(v) \mid f(u)) \leq a$, where a is a constant independent of u, v.

Similarly, using a simpler version of ψ one can show that $C(v \mid u) \leq a$. Thus

$$ed_f(v) \leq \frac{C(f(v) \mid f(u))}{C(f(v))} \leq \frac{a}{C(f(v))}$$

because $C(v \mid u) \leq a \leq log\ |v|$ if x is chosen such that the corresponding $f(v)$ is large enough, and thus $a/C(f(v)) < \epsilon$ which contradicts the assumption that the edit difficulty of f exceeds ϵ.

This result establishes that efforts to formalise and quantify the process of integration are misguided. As soon as we understand the mechanisms by which information is combined together, we gain the ability to reverse that process, thus breaking the spell of integration. Returning to Leibniz's example, as soon as we can explain a process in terms of a mechanical model, we can enter it as into a mill, and see that there is nothing integrated about it. The systems that can integrate information are strictly those whose nature precludes us from identifying any formal model.

Given our definition of integration, we are left with two options. We must abandon either 1) the idea that people enjoy integrated consciousness, or 2) that our language is strong enough to express consciousness objectively.

4.6 Evidence from neuroscience

Recent results in neuroscience seem to suggest that memories are indeed open to editing. For example, Ramirez et al. (2013) successfully created a false memory by optogenetically manipulating memory engram-bearing cells in the hippocampus, leading mice to show increased freezing in a context where a foot shock was never delivered. This result suggests that it is possible to generate an internally represented and behaviourally expressed fear memory via artificial means. On the other hand, Gräff et al. (2014) showed the reverse, that it is possible to attenuate remote memories of fear in mice by using a HDAC2-targeting inhibitor during the reconsolidation process that is initiated upon memory recall.

The key question when it comes to integrated information is the specificity with which human memories could potentially be edited. For example, it is always possible to 'edit' someone's knowledge by hitting them over the head with a hammer. What is not clear is whether one specific memory can be successfully disentangled or separated from all

other memories and edited in isolation. Can you be made to believe that your toothbrush is green instead of blue, without affecting anything else you know?

If it turns out that people's memories can be torn apart and manipulated at will to any level of specificity, then the human mind would really be no different to an artificial smell detector. In such a case, not only would the motivation for adopting the intentional stance be vitiated, it would also imply that we could no longer trust the reliability of our own memories, thus eliminating any possible grounding for objectivity.

4.7 Scramble-in, scramble-out

Our tentative suggestion that the brain carries out irreversible data compression raises some intriguing questions regarding how and where this feat might be achieved. Somewhere, brain processes must feature intractable complexity, which has the effect of binding information together.

When stimuli are picked up by the brain they enter at disintegrated locations. For example, visual stimuli enter through the optic nerve and are processed initially by the primary visual cortex. When a visual stimulus is encoded in the occipital lobe it clearly has not yet been integrated with the rest of cognition. Stanley, Li and Dan (1999), for instance, analysed an array of electrodes embedded in the thalamus lateral geniculate nucleus area of a cat and were able to decode the signals to generate watchable movies of what the cat was observing.

Similarly, the initiation of action must be localised in particular areas of the motor cortex which control the relevant muscles. Because this readiness potential must detach from the rest of cognition, it is no longer integrated. For example, following up on Libet's original experiments, Soon et al. (2008) demonstrated that, by monitoring activity in the frontopolar prefrontal cortex, they could predict a participant's decision to move their right or left hand several seconds before the participant became aware of it.

However, assuming people's behaviour is irreversibly integrated, then somewhere between the stimulus entering the brain and a decision to act leaving the brain, there must be a point where the information cannot be fully disentangled from the rest of cognition. At some point between

perception and action, the contents of cognition are effectively entangled into a unified, complex whole and cannot be separated, thereby forcing the adoption of the intentional stance. We label this idea 'scramble-in, scramble-out' to reflect the irreversible integration and disintegration that must occur.

The aspects of cognition that have been clarified by neuroscience so far tend to involve processing before scramble-in or after scramble-out. For example, it is well established that the occipital lobe is involved in visual processing or that the prefrontal cortex encodes future actions before they are performed. These components are modular in that they have specialized, encapsulated, evolutionarily developed functions. However, somewhere between input and output there must also be a binding process of integration that no modelling can disentangle.

Fodor (2001) summarizes as follows: "Local mental processes appear to accommodate pretty well to Turing's theory that thinking is computation; they appear to be largely modular...By contrast, what we've found out about global cognition is mainly that it is different from the local kind...we deeply do not understand it".

5. Subjective Experience as Compression

In summary, we have proposed that the behaviour of a system appears unitary when the data compression it carries out is so sophisticated that it forces the external observer to adopt the intentional stance. And yet this theory does not seem to offer a complete account of consciousness. Consciousness is not merely something we attribute as external observers. Intuitively, it is not simply a matter of behaviour. Instead, we *feel* consciousness personally. The aspects of consciousness we have addressed so far, namely intelligence and information binding, neglect the crucial *subjective* aspect of consciousness.

It seems possible to conceive of a sophisticated artificial compressor that compresses large amounts of current and historical data in parallel, though without experiencing the same form of awareness that humans are familiar with. In this section we explore the idea that the compression carried out by the brain is likely to have an additional ingredient which sets it apart from insentient compression systems, namely that of socially motivated self-

modelling.

5.1 Self-compression

According to Dunbar and Shultz (2007), intelligence was selected for, not by the need for technical competence, but by the computational demands of living in large, complex societies. When we watch other individuals, we realise that their behaviour reflects deep and complex patterns (a property known in AIT as logical depth, see Bennett, 1995). Rather than simply cataloguing and memorising every action they perform, we can instead posit the more succinct hypothesis of a concise 'self' which motivates these actions (the intentional stance). By maintaining this theory of selfhood we can compress the behaviour of others and thus make accurate predictions as to how they will behave in different contexts.

But the behaviour of other humans also has another component, namely that they react to you, the observer. In order to best predict and manipulate the behaviour of others, it pays to maintain a model of one's own self (see Friston & Frith, 2015). Indeed, this was exactly the approach adopted by Schmidhuber (1990) for developing a very deep learning machine. His system consisted of two recurrent neural networks, one for modelling the history of actions and perceptions, and the other a reinforcement learner that used the compressed self-model to plan and maximize success, "thus showing a rudimentary form of self-introspective behaviour" (Schmidhuber 1991; see also Schmidhuber, 2015, for a review of how mirror neurons can be explained as by-products of history compression). According to Halligan and Oakley (2015), people have a psychological predisposition to anthropomorphise their own behaviour, the evolutionary advantage being that it "enables the development of adaptive strategies such as predicting the behaviour of others, which could be beneficial to species survival".

Consider the following thought experiment: you are stranded on a desert island with one other person. Is it possible to conceive of her as a philosophical zombie without subjective experience? When you observe her behaviour you seek to compress it, to model her and anticipate what she will do next. You immediately appreciate that her behaviour is logically deep: there are patterns in it, but it proves infeasible to reduce those patterns

down to a set of independent rules. Furthermore, you realise that she is responding to your attempts at modelling. She is observing you observing her observing you. To be able to cope with this recursion you must maintain a model of your own actions, compressing your own behaviour to anticipate her responses. In effect, modelling this other is so difficult that it forces you to become aware of yourself as an independent entity. And still you cannot manage to disintegrate her behaviour. The optimal strategy you can adopt is to treat her as a unified, integrated whole and yourself as well. In other words, her behaviour exhibits a level of data compression so profound, it forces you to adopt the intentional stance and model both you and her as conscious entities.

We propose that this 'understanding of the self' is a requirement for accurately modelling and predicting the complex behaviour of others (see Hesslow, 2002; Humphrey, 2006; Metzinger, 2004). If an individual lived in complete isolation within a simple environment, there would be no motivation for maintaining such a model. It is only when people are embedded in a complex competitive social environment that the goal of interacting with others requires them to anthropomorphise their own actions. This recursive modelling gives rise to an understanding of selfhood, an appreciation of the first-person experiential self.

5.2 Limitations of scientific objectivity

We are still left with a difficult question: why should science struggle to account for this subjective, personal aspect of experience? Examples of subjective qualia include the pain of a toothache, the taste of sweetness, or the perceived redness of an apple. Despite their personal vividness, such experiences seem to defy objective, reducible description. How could subjective qualia be useful for predicting the behaviour of others, when they completely evade scientific description?

In seeking to answer such a question, we are pushing up against the limits of our existing framework for objectivity, so the reader should be warned that what follows is unavoidably speculative. We propose that, rather than reflecting a mystery of consciousness, the difficulty of providing a formal account of qualia instead reflects a limitation in the descriptive power of

scientific language. Science is a tool, designed by people, which facilitates communication in the absence of personal interaction. The language of science allows scholars to engage with each other's work, despite being widely distributed through both space and time (hence Newton's quip of "standing on the shoulders of giants" - without having had to meet them all in person).

A universal language of this type depends on abstraction away from any idiosyncratic, personal theories, which might depend on local understandings within a specific social community. The discipline is interested in identifying what stands as the case for all humans, no matter who they are, where they are, or what they have previously experienced. Accordingly, it describes objects in terms of measurement standards which are engineered to ensure agreement between humans. These standards are linked to globally accessible natural phenomena which are as incompressible (i.e. unpredictable, inexplicable, random) as possible (e.g. atomic decay). Disagreements and misinterpretations arise when different observers identify different patterns in the same object. The more irreducible scientific standards are, the less scope there is for deviations in interpretation, and the more stable a platform they provide for measurement.

The strength of the scientific standards maintained by the Bureau International des Poids et Mesures (BIPM) is sufficient for disintegrating many environmental phenomena. However, like their predecessors, such the Krypton-86 metre standard, the current definitions of length of time will eventually be replaced by superior ones, which are even less resistant to compression (see Tal, 2011). The fact that BIPM standards can be recognized as having been superseded implies that there must exist a more fundamental standard, an even harder compression problem, relative to which the limitations of current standards can be exposed.

For instance, how do we know for sure that the speed of light in a vacuum is indeed 'constant'? What evidence do we have for its stability? Simply asserting the claim by fiat is of no use. Stability arises instead from the knowledge that many people have tried to find predictable variance in the speed of light and, so far, failed. In other words, the test that measurement standards have passed, allowing them to be consecrated at the heart of the objective scientific perspective, is that of resisting compression by a large

group of well-motivated *people*. The scientific perspective ultimately relies on a social standard.

We propose that compressing human behaviour is the ultimately hard problem on which language, measurement and representation are founded. Anticipating and manipulating the behaviour of others is the hardest possible challenge humans face. The social prediction game provides the most inexhaustible source of complexity, and, as such, provides the most stable, incorruptible, immutable grounding for expressing what is and what is not the case.

5.3 On the predictive value of qualia

According to Dennett (1991), qualia are commonly regarded as ineffable (cannot be communicated), intrinsic (do not relate to anything) and private; positing their existence serves no purpose. And yet, if qualia were truly beyond representation in any language, how would we remember them? If memories of qualia could not be recorded in some way, then experiencing a feeling would always seem novel and arbitrary, as if it was being experienced for the very first time. In contrast, we remember what seeing red feels like. We know intuitively when a current subjective experience of red matches a previous subjective experience of red.

Dennett's view is based on the idea that science is capable of expressing everything that can be objectively demonstrated. Yet even the BIPM does not hold such a radical view. For example, the BIPM accepts that existing standards will eventually be objectively superseded, even though the manner in which this will be demonstrated is beyond the expressive power of our current scientific framework (see Tal, 2013, 2014). In other words, the BIPM implicitly accepts that there are objective phenomena which are currently beyond science, and that perhaps may always remain beyond science.

From the scientific perspective, it appears as if a philosophical zombie, without any subjective experiences, could act exactly the same as a human (see Chalmers, 1995). However, we suggest that qualia are not genuinely inefficacious. Instead, their efficacy is too sophisticated to be identified by science.

We propose that qualia are idiosyncratic community-maintained theories that facilitate the compression of human behaviour. Although they are easily expressed in the local language of the social prediction game, they cannot be expressed in the weaker global language of science. For instance, the *scientific* description of red is the element of the experience that can be communicated to all humans regardless of their personal experiences (see Jackson, 1986). In contrast, the *qualitative* description of red is the element of the experience that can only be described relative to immersion within a particular social context.

Science can only describe the features of a stimulus that humans could, in principle, agree on without ever meeting each other face to face. In contrast, a tightly-knit community of humans enjoy a much more nuanced understanding of agreement, grounded in their everyday interactions. Thus, although science is an excellent tool for communicating over distances, it does not necessarily represent the final word in objectivity.

The so-called explanatory gap (Levine, 1983) between scientific description and qualia arises because science seeks to abstract away from the local contexts which define human life, towards a more neutral 'objective' perspective, which, in exchange for greater limpidity, sheds its ability to make accurate predictions about human behaviour. Anticipating and manipulating the behaviour of others requires immersion in a community, embodied social interactions, as well as sharing human fears and aspirations. In short, it requires the maintenance of qualia.

Although these ideas are undeniably speculative, they do make a prediction which is open to empirical verification. If our view is valid, and the compression of human behaviour provides the deeper fundamental grounding for the shallower scientific perspective, then science should not be able match human intuition at predicting the behaviour of others.

6. Conclusion

Conscious experience does not seem anything like data compression. When we think of data compression we think of simple programs that are used to reduce the size of computer files. It does not seem that programs of this sort could experience anything. Indeed, it is hard to accept that consciousness

could be expressed in terms of information processing at all. However, this superficial understanding of data compression is misrepresentative.

Data compression does not merely quantify processor speed or memory capacity. Instead, it addresses something far more fundamental, namely the ultimate limits of formal systems and objective representation. The more powerful and deeper the language of a system, the greater the compression that it can achieve. At the limit, compressing a piece of data down to its shortest program coincides with defining one's own language. It is this ultimate and universal hardness which underlies data compression's suitability as a framework for expressing the hard problem of consciousness.

Unlike other natural phenomena which can be addressed from an 'objective' perspective, we have proposed that consciousness is a language-complete problem, that is, one equivalent in difficulty to formalising the foundations of language itself. Rather than getting closer to unravelling the mystery, we are instead led to question the structure of our representations and the reliability of our objective knowledge. Thus, consciousness becomes a question of information, computation and complexity, not a question of physics and chemistry.

It may even be a mistake to regard consciousness as a *problem*. A problem is a situation that can be investigated, that might feasibly yield a solution. In contrast, we have suggested that exhaustively predicting human behaviour is so very hard that it represents, not a question to be resolved, but the reliable foundation of understanding relative to which all other objective standards are calibrated. Hence, rather than untangling the complexity of the mind, the exploration of consciousness may merely lead to the identification of stronger standards for describing the external world.

References

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, 36(2), 737–746.

Baars, B. J. (2007). *The global workspace theory of consciousness*. The Blackwell companion to consciousness, 236–246.

Bennett, C. H. (1995). Logical depth and physical complexity. *The Universal Turing Machine A Half-Century Survey*, 207–235.

Bringsjord, S., & Zenzen, M. (1997). Cognition is not computation: The argument from irreversibility. *Synthese*, 113(2), 285–320.

Calude, C. S. (2013). *Information and randomness: an algorithmic perspective.* Springer Science & Business Media.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., . . . (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine,* 5(198), 198ra105–198ra105.

Chaitin, G. (2006). The limits of reason. *Scientific American,* 294(3), 74–81.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chater, N., & Brown, G. D. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, 32(1), 36–67.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1), 19–22.

Dennett, D. C. (1991). *Consciousness explained.* Little, Brown.

Dessalles, J.-L. (2011). Coincidences and the encounter problem: A formal account. *arXiv preprint arXiv:1106.3932.*

Dunbar, R. I., & Shultz, S. (2007). Evolution in the social brain. *Science,* 317(5843), 1344–1347.

Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology.* MIT press.

French, R. M. (2012). Moving beyond the turing test. *Communications of the ACM* 55(12), 74–77.

Friston, K., & Frith, C. (2015). A duet for one. *Consciousness and Cognition*, 36, 390–405.

Gauvrit, N., Zenil, H., & Tegnér, J. (2015). The information-theoretic and algorithmic approach to human, animal and artifcial cognition. In G. Dodig-Crnkovic & R. Giovagnoli (Eds.), *Representation and reality : Humans, animals and machines.* Springer.

Gräff, J., Joseph, N. F., Horn, M. E., Samiei, A., Meng, J., Seo, J., ... (2014). Epigenetic priming of memory updating during reconsolidation to attenuate remote fear memories. *Cell*, 156(1), 261–276.

Halligan, P., & Oakley, D. (2015). Consciousness isn't all about you, you know. *New Scientist,* August 15.

Hameroff, S. (1998). Quantum computation in brain microtubules? The Penrose-Hameroff 'Orch OR' model of consciousness. *Philosophical Transactions-Royal Society of London Series A Mathematical Physical and Engineering*

Sciences, 1869-1895.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences,* 6 (6), 242-247.

Humphrey, N. (2006). *Seeing red: A study in consciousness.* Harvard University Press.

Hutter, M. (2007). On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1), 33–48.

Hutter, M. (2009). Human knowledge compression contest. *http://prize.hutter1.net/.*

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, 291–295.

James, W. (1890). The consciousness of self. *The Principles of Psychology,* 8.

Kirchherr, W., Li, M., & Vitányi, P. (1997). The miraculous universal distribution. *The Mathematical Intelligencer,* 19(4), 7–15.

Koch, C. (2004). *The quest for consciousness.* New York.

Koubeissi, M. Z., Bartolomei, F., Beltagy, A., & Picard, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy & Behavior,* 37, 32–35.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications,* 157, 17.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Auarterly,* 64(4), 354–361.

Li, M., & Vitányi, P. (2013). *An introduction to Kolmogorov complexity and its applications.* Springer Science & Business Media.

Maguire, P., & Maguire, R. (2010). Consciousness is data compression. In *Proceedings of the thirty-second conference of the Cognitive Science Society* (pp. 748–753).

Maguire, P., Moser, P., Maguire, R., & Griffith, V. (2014). Is consciousness computable? Quantifying integrated information using algorithmic information theory. *In Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

Maguire, P., Moser, P., Maguire, R., & Keane, M. (2013). A computational theory of subjective probability [Featuring a proof that the conjunction effect is not a fallacy]. In *Proceedings of the thirty-fifth annual conference of the Cognitive Science Society* (pp. 960–965).

Mahoney, M. (2009). Data compression programs. *http://mattmahoney.net/dc/.*

McGinn, C. (1991). *The problem of consciousness: Essays toward a resolution.* Blackwell Publishers, UK.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. MIT Press.

Ramirez, S., Liu, X., Lin, P.-A., Suh, J., Pignatelli, M., Redondo, R. L., . . . (2013). Creating a false memory in the hippocampus. *Science,* 341(6144), 387–391.

Rathmanner, S., & Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy,* 13(6), 1076–1136.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.

Rolls, E. T., & Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in Neurobiology,* 95(3), 448–490.

Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: an adaptive perspective. *Trends in Cognitive Sciences*, 15(10), 467–474.

Schmidhuber, J. (1990). An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 253–258).

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in modelbuilding neural controllers. In *From animals to animats: Proceedings of the first International Conference on Simulation of Adaptive Behavior* (sab90).

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2), 234–242.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2), 173–187.

Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems* (pp. 48–76). Springer.

Schmidhuber, J. (2013). My first deep learning system of 1991 + deep learning timeline 1960-2013. *http://people.idsia.ch/ juergen/firstdeeplearner.html.*

Schmidhuber, J. (2014). I am Jürgen Schmidhuber, AMA! *https://www.reddit.com/r/MachineLearning/comments/2xcyrl/* .

Schmidhuber, J. (2015). On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249* .

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control,* 7 (1), 1–22.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543–545.

Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience,* 19(18), 8036–8042.

Strawson, G. (2011). Soul dust. *The Observer,* January 8.

Tal, E. (2011). How accurate is the standard second? *Philosophy of Science,* 78(5), 1082–1096.

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass,* 8(12), 1159–1173.

Tal, E. (2014). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science*, axu037.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience,* 5(1), 42.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3), 216–242.

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch Ital Biol,* 150(2-3), 56–90.

Tononi, G. (2015). Integrated information theory. *Scholarpedia,* 10(1), 4164.

Tononi, G., Edelman, G. M., & Sporns, O. (1998). Complexity and coherency: integrating information in the brain. *Trends in Cognitive Sciences,* 2(12), 474–484.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences,* 91(11), 5033–5037.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind,* 433–460.

Vitányi, P., & Li, M. (2000). Minimum description length induction, bayesianism, and kolmogorov complexity. *Information Theory, IEEE Transactions on*, 46(2), 446–464.

Wolff, J. G. (1993). Computing, cognition and information compression. *AI Communications*, 6(2), 107–127.

Zenil, H., Marshall, J. A., & Tegnér, J. (2015). Approximations of algorithmic and structural complexity validate cognitive-behavioural experimental results. *arXiv preprint arXiv:1509.06338.*