Kolmogorov complexity as a smoking gun of the hard problem of consciousness

Igor Salom
Institute of Physics, Belgrade University, Pregrevica 118, Zemun, Serbia

The longstanding problem to understand if, why, and how objective functioning of the brain gives rise to a subjective perspective has been, in the last few decades, commonly known as the hard problem of consciousness. However, due to the strictly subjective and qualitative character of subjective experience, it is difficult to get a firm grip on the problem itself, which led some philosophers even to deny the very existence of the problem. In this paper, we point to a relation between the quantity of information (i.e. Kolmogorov complexity) and the phenomenon of subjective experience. In a thought experiment that we construct, the amount of information existing subjectively will be significantly higher than the amount of information existing objectively. We argue that such a quantifiable discrepancy clearly identifies one mathematically well-defined aspect of the hard problem which, in turn, makes it at least much harder to deny its existence. If we take a stronger stance, this aspect of the problem further undermines hopes that a satisfactory strictly physicalist explanation of the subjective experience could be ever given.

INTRODUCTION

In his 1995 paper, David Chalmers has famously coined the term "hard problem of consciousness" to address the mystery of "why should physical processing give rise to a rich inner life at all?" [1]. While it is the very existence of subjective experience that is the essence of this conundrum, this existence is always exemplified by pointing to qualitative manifestations of subjectivity: sensory experiences (qualia), internal dialogue, thinking, feeling, etc. Due to the strictly subjective and qualitative nature of these manifestations, they elude precise definitions and, in particular, any attempt to mathematical rigor. The aim of this paper is to point out that, under certain circumstances, a well-defined mathematical quantity that effectively measures the amount of information - the so-called Kolmogorov complexity [2, 3] - can be used as a quantitative indicator of the existence of the subjective perspective. These "certain circumstances" are peculiar in the sense that the objective amount of information existing in a system (a box, or the entire universe) is, in these cases, manifestly lower than the amount of information subjectively observed by an agent (who is a part of the same system).

If and when such a discrepancy occurs, it should clearly present a glaring issue: even those not familiar with the information theory are accustomed to the fact that the missing information must be then somewhere hidden. For example, it is intuitively clear that, no matter how good the compression method is, the entire Hollywood movie production in the last 100 years cannot be (losslessly) recorded on a single (700MB) compact disk - and if such an ordinary CD could be indeed used to play any Hollywood movie at our bidding, it would be clear that the most part of the information is actually hidden somewhere else, and not located on the CD. In our case, if an agent observes more complexity in the system than there objectively exists, then this difference must be contained somewhere - and since this part of the information is not objectively present, it then represents a well-defined quantity that exists only subjectively. Consequently, if we can pinpoint a mathematically well-defined entity that exists only subjectively (and not objectively), it becomes very hard to deny the existence of the subjective perspective or to surmise that the latter is simply produced by objective processes alone. In particular, in such circumstances, the difference between agents with subjective experience and philosophical zombies [4, 5] becomes even more apparent and clearly defined: in the latter case, there simply is not any system of information, i.e. any perspective in which there is more information than objectively existing. Conversely, once we acknowledge the existence of a perspective containing more information than objectively existing, we have switched from philosophical zombies to agents with internal experience, and we have introduced a subjective, that is, a non-objective realm.

The very fact that situations with more subjectively than objectively existing information can occur, may at first sound highly puzzling or even absurd. But, on closer inspection, the phenomenon is not that surprising nor unexpected: as we shall clarify, the missing information is actually encoded in the "self-location" information possessed by the agent.[1] However, once some naive intuitions and misconceptions are removed, it turns out that this recognition does not resolve the problem at all. On the contrary, it is rather that this particular informational aspect of the self-locating uncertainty was, to our knowledge, unjustly neglected or glossed over. It emphasizes that the very existence of the internal information system (one which is introspected), irrespective of any additional phenomenal states or qualia, is deeply myste-

[1] Notions of "self-locating beliefs" and "self-locating information" come from the decision theory (e.g. see [6]), while the notion of "self-locating uncertainty" is of great importance in many-worlds interpretation (MWI) of quantum mechanics [7]. Our usage will be closely related to that in MWI, as will become clear in the context.

rious. Namely, we will make it clear that, in principle, an arbitrary amount of information can be hidden in this self-location (e.g. equivalent to gigabytes of data) while, at the same time, this information is absent from the objective reality – a fact that cannot be easily ignored. Just as in the case of the hypothetical CD containing all of the movies, this information is real, indisputably observed, and thus must be accounted for.

The example we will consider below will not presuppose any specific details about the physical universe: in particular, as a thought experiment, it can be put either in the context of classical physics, in the context of quantum physics, or, in principle, in some other framework. However, the significance of our conclusions is much higher if the laws of physics are strictly deterministic in a way that leads to the conservation of the quantity of information (since all future states are already encoded in the initial state). This class includes the case of classical physics[2], as well as certain interpretations of quantum mechanics. Of the latter, it is worth mentioning that this problem of discrepancy between the Kolmogorov complexity of objective versus subjective reality is strikingly acute and instructive in the context of many-worlds interpretation (MWI) of quantum mechanics [9, 10], and bears many implications both for this interpretation and for the hard problem itself. Namely, this interpretation tacitly allows that the objective state of the universe is of exceedingly low complexity (since the interpretation is deterministic and the initial state may be very simple), while we subjectively observe immense amounts of information (so that essentially all information about the observable universe - from human DNA to distant stars - exists only subjectively, hidden in the self-location knowledge). While we will add a few more remarks on MWI in what follows, the missing information problem in MWI context deserves to be analyzed in more detail and in its own right, and thus we will treat this issue elsewhere.

In the following section we develop our thought experiment. After that, we will discuss its implications, as well as some potential objections to our reasoning (in particular, the relevance of the notion of "relative information"). The final section is a brief summary.

THE LIBRARY OF BABEL THOUGHT EXPERIMENT

As announced, the subject of our investigation in the context of the hard problem will be the quantity of information, as expressed by Kolmogorov complexity [3]. In most simple terms, Kolmogorov complexity represents the (least) amount of information (e.g. expressed in the number of bits) required to fully specify an object. It is formally defined for sequences of symbols (bits) as the length of the shortest computer program that reproduces that precise sequence. Strictly speaking, it is defined up to an additive constant (related to the particular choice of the programming language or, more formally, to the choice of the particular universal Turing machine), but for our purposes, these and further mathematical details will not be relevant.

In this context, it will be sufficient that Kolmogorov complexity is a well-defined mathematical notion that grasps our intuitive idea of how difficult (i.e. lengthy) it is to fully specify an object. In our case, we will discuss the complexity of physical objects, by considering the Kolmogorov complexity of their full description. In order to make the connection more concrete, we may imagine being in possession of an ideal 3D printer, capable of printing any physical object with arbitrary (or even ideal) precision. (In the quantum-mechanical case, the printer should be able to ideally prepare given allowed quantum state.) The complexity of an object (i.e. the information contained in that object) would then correspond to a maximally compressed set of instructions for this printer, required to create the object.

So defined complexity of an object, or more generally of a physical system, is of a particular significance when the laws of physics are deterministic.[3] Namely, in a deterministic case, the state of a given isolated system at any moment in future is completely determined by any of its past states. In turn, this means that if we possess the complete description of the initial state of the system, and have full knowledge of the dynamical laws, we are able to, in principle with arbitrary precision, compute the state of the system at any given future moment. This further means that Kolmogorov complexity of the system state at any moment in the future cannot be greater than the complexity of the initial state plus the length of code that implements computation of the system's temporal evolution (and plus the length required to specify that time instant with given precision). If, in addition, the laws of physics are symmetric under time inversion (as is the case both with Newtonian physics and with Schrodinger's equation), then also the past states of the system can be computed as a function of any given future state. In such cases, Kolmogorov complexity of an isolated system, defined as a sum of the state description plus the algorithm for time evolution (plus the length of time parameter), becomes a conserved quantity. (We stress that computational time and memory resources required to compute with arbitrary given precision the total description of the system state at a given moment are irrelevant for the definition of the quantity of information.) Conserved quantities are generally important properties

[2] Excluding the zero-measure set of initial conditions that results in Norton's dome [8] type of situations.

[3] Strictly speaking, they should be also "effectively computable" [11], but this is automatically satisfied for all physical theories of interest.

of any system. We will thus concentrate on these cases when Kolmogorov complexity is a constant of motion and thus a fixed property of the system.

Another important property of the Kolmogorov complexity of an object is that it quantifies the total information that can be derived from the object, limiting thus the (explanatory) abilities of the object. For example, if we can hear Beethoven's Symphony No. 9 but cannot identify the source, it is not reasonable to seek the explanation for the music among objects whose complexity is lower than the complexity of the music piece (unless we expect some additional source of information to exist - e.g. radio waves broadcasting the music). In general, it should be fairly obvious that no more information can be drawn from the physical object than from its total description. If the laws are deterministic, this also holds for any future state of the system, which cannot contain more information than the description of the initial state. For example, if the total description of the object describes a simple ideal crystal lattice of a cubic shape (with Kolmogorov complexity of the order of tens of kilobytes), on close inspection of the object itself we certainly cannot expect to find a hidden inscription of Shakespeare's Hamlet. This remains so even if wait a bit, or manipulate the object in some sense: e.g. if we cut the object, unless we do it in a way to introduce the missing information (the text of Hamlet in this case) by our external action. Hardly surprising conclusion, but very important in what follows, is this: *there are no more Shakespeare's verses in the actual object than there are in its total description.* What is less intuitively clear is that this holds not only for any meaningful information but also for any random information - the object itself cannot help us generate a random number any more than its description can. This can be most easily seen as follows: whatever info we might gain from any interaction with the physical object, we could equally well gain the same info by simulating the same interaction on the computer, starting from the object description. Needless to say that all these conclusions remain valid also for a composite system consisting of a few spatially separated objects.

Now it is time to specify the physical system of interest here. We will consider an immensely long array of equally spaced books which are identical in every minute detail, apart from the printed content of the book. And the content of the books will be the following. The first book will be entirely filled by letter "A" and no other character: 410 pages, each page of forty lines, each line of eighty letters.[4] The second book will be precisely the same, of the same size, except for the very last letter that will be "B". The next book will have the letter "C" at the end, and so on - in short, this arrangement will feature each possible sequence of 29 basic characters (26 letters of the English alphabet, the period, the comma, and space) covering these 410 pages, sorted in the strict lexicographical order. Altogether, $29^{410\cdot 40\cdot 80}$ books, that is, approximately, one-followed-by-two-million-zeros of books. In order to have an isolated system, we will take that the books are enclosed in a stupendously long box, and let them be in a perfect vacuum (alternatively, we may also imagine that these books are the only objects in the universe of this thought experiment). The fact that such an array could not even remotely fit into our observable universe is, of course, of absolutely no relevance for our thought experiment (actually, we could have chosen even realistically smaller values, only at the cost of making our main point less intuitively clear).

Next, we will consider the amount of information contained in this system, i.e. its Kolmogorov complexity. To reduce the complexity, we will take that each page is built as a perfect crystal sheet of A4 size (120 x 297 x 0.1 millimeters), made of some element that can form perfect cubic crystal lattices (e.g. iron), and that the pages are merely laid one over the other, with no cover pages. The distance between adjacent books will always be the same, e.g. exactly one meter, and all books will be identically oriented. (In principle, we should also require that the entire system is at zero temperature, but this and similar technical details are unlikely to influence our conclusions.) Instead of using ink for printing, we may engrave letters in the crystalline pages. In this way, we have assured that the required instructions for 3D printing of all these books become surprisingly simple, i.e. short: the physical substance of each of the books can be now very concisely specified in full detail, as well as their spatial location. As for the content of the books, it is of even lesser complexity: apart of some code needed to specify the font of the letters, the program for printing the book contents nearly reduces to a single "for loop" of the form "for i = 1 to $29^{410*40*80}$ print(i)". Because, obviously, what is written in each book is just its ordinary number, written in the numeral system with base 29, where basic characters take the role of digits. (Therefore, in a programming pseudo-language, the algorithm for printing the books is simply "move from first to the last book and in each one print its ordinary number in the base 29 numeral system".) It is likely that full instructions for 3D printing the entire immense array of these books can be compressed into no more than a few tens of kilobytes.

The important question to be addressed is this: does this box, full of books, contain "The Tragedy of Hamlet"?

Well, we have already concluded above that there is as much Hamlet in the object as there is in its total description! Since the "for loop" that populates the contents of the books contains no mention of the Danish prince or anything alike, so the 3D printed object also cannot contain anything like the text of Hamlet.

Yet, this is where our conclusions collide with our intuition. First of all, somewhere in that vast row of books is obviously an exact copy of The Tragedy of Hamlet,

[4] Here we allude to Jorge Luis Borges's short story "Library of Babel", and the book size is taken to be the same as there. However, unlike in Borges's story, in our case the books will be perfectly arranged.

without a single typo or an extra space, and billions and billions more with a few typos, extra spaces, or some variations - all of them telling the drama of the Prince of Denmark. But they are lost in the immense, meticulously ordered junk. Not just lost in the ordinary sense (of their location being forgotten/unknown), but lost in the sense of informational, and thus in some sense literal, vanishing. There is absolutely nothing we can learn about Hamlet if given these books: e.g. there is no way to check what was the precise wording of the famous "There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy", or no way even to check the spelling of "Horatio". Absolutely nothing we can gain about Shakespeare's drama that we had not already known before. For, if there was anything to find out from this line of books, then we could also get that information by inspecting the printer instructions alone (or maybe by simulating our actions on the computer). But the instructions describing all the text in the books contain only a single, utterly Hamletless, for-loop. To find Hamlet in that mass of books, one would have to be Shakespeare, and also to Shakespeare, all these books could not have helped a bit to write this play, even if he had time to rummage through them.

Nevertheless, our intuition objects: there is that perfect copy of Hamlet somewhere there, and we can clearly imagine us standing there, in front of it, reading it, and finding out the story of Hamlet. Is not it, in that sense, very much real and existing?! Yet, for this to happen, we would have to stand in front of it - namely, we would have to move some number of meters from the beginning of the box. This number of meters that we have chosen, if written down in the basis 29 numerals, is exactly the text of Hamlet - so, if we were asked in advance to choose the position where we want to stand, we would have to specify it by writing down the precise copy of Hamlet. And if we specify it in advance, then the Hamlet is there only because we have introduced it inside the box, encoded in our location.

Overall, this is a curious situation that the matter (a hardcopy of the Hamlet) somehow exists without information it should contain (the information about Hamlet). Modern physics – from quantum mechanics, over relativity, black-hole physics and cosmology – teaches us that information is an essential notion, maybe even more sound and basic than matter. Thus, whether it still makes any sense to say that the copy of Hamlet exists, and in which exactly sense, is rather dubious in our view, but this is not the subject of our primary interest here. What suffices for our purposes here is to conclude that there is definitely no sense in which *information* about Hamlet exists in these books. We are not an inch closer to the story of Hamlet after printing all these books than we were before.

Of course, this is so only as long as all the books are totally identical in every other respect apart from the content. If we put a single tick mark, or just a dot, on any page of the "perfect Hamlet" copy, the Hamlet suddenly pops fully into existence, in every common sense. Namely, if we consider the Kolmogorov complexity of the entire box now, there must also be the information about the position of that tick mark. And this information on the whereabouts of the tick mark cannot be supplied without identifying the book where it is marked, i.e without telling (encoding) the whole story of Hamlet. We cannot do any better by attempts to specify only the position of the tick mark, e.g. its distance from the box beginning - since to spell this precise distance we must again write down the text of Hamlet and interpret it as the number of meters. In any case, to 3D print this tick mark (or a single notch) along with the books, we must enlarge our instruction code at least for the length of (compressed) text of Hamlet. And if we do so, the information becomes really there: just as much in the printing instructions as in the row of the actual books.

In general, singling out, in any sense, any of the books, takes an amount of information pretty much equivalent to the information contained in that book, and thus it adds the same amount to the printing instructions. We emphasize that Hamlet is in no sense special here. We, of course, do not think special only in comparison to, for example, Macbeth, but also in comparison to any other, non-sensical or completely randomly filled book (a genuinely random text would be totally non-compressible and thus would carry an even higher Kolmogorov complexity increase than any proper English text). As already stated, this array of printed books cannot help us even to generate a random sequence of characters, let alone to read works of art. The former might not be so obvious, as someone may be tempted to hastily think: "I can pick any of the books and there will be 410 pages long random sequence". But, of course, all the randomness we gain from that book was already contained in our "random" choice of the book, hence this random information we already had to "invest" - and nothing to gain.

To proceed further with our thought experiment, we note that the information about Hamlet (as well as the information from every other book) will still be absent from the box even if we add some stuff to accompany each of the books, as long as the uniformity is not broken. For example, we might add in front of each book an identical camera that will just stand and take photos of pages. We can also add a mechanism to turn pages and activate the camera so that each page gets photographed. Granted, these new elements will make the box contents certainly more complex - with complexity increased for the description of the camera and the mechanism - but all the contents of the books will be still reproduced by the same trivial "for loop" that has no mention of Hamlet. Nothing in this sense is changed even if we add an identical computer to each camera, that will use some artificial intelligence (AI) to do the optical character recognition of the book in front of it and to digitally store recorded information in text form. Neither the fact that now we also have a digital copy of each book in the computer has any

bearings to our previous conclusions. The entire information present in this box will be the sum of complexities of the book fabric, the camera, and the computer, plus the "for loop" algorithm for repeating this setup for each book and for populating the book contents. This is not only sufficient to 3D print the entire box contents, but at the same time, this is all the information that is in principle obtainable from this box. In other words, there is *no sense in which information about Hamlet exists in this box*, even after these modifications.

Unless – and now we come to the main point – the idea of "strong AI" is true. Unless the AI in the camera is such that it is "conscious" in the sense we commonly understand this word. In the sense that "there is someone in there, looking", or that "it is something like to be that camera AI while taking photos" (paraphrasing T. Nagel's "What is it like to be a bat" [12]). For, if there is a subjective perspective of the camera, then this perspective is quite different. Take the camera facing the perfect copy of Hamlet. To help our intuition, let us call this "AI-equipped camera with a mechanism for flipping the pages" a robot, or an agent. What is this agent's description of the box contents? Curiously, this agent's description is necessarily more complex than the objective description of the entire box.

Namely, to fully specify the physical system inside the box from the perspective of this agent then, in addition to the printing instructions, we must also include the entire text of Hamlet in the description, since the Hamlet is what the agent is looking at! Or, if it might seem unclear whether the printing instructions should be considered as accessible to this agent and thus truly a part of his internal information system, we might consider the agent facing one even more interesting book. Among the books, there is also a book that contains the perfect text of Hamlet followed by the precise printing instructions used for the 3D printer. For this book to exist it is sufficient that the book size is large enough to fit both Hamlet and the instructions. Even if 410 pages is not enough for this, we can easily extend the length of the books in the thought experiment to an arbitrary required length (we can also extend the character set to include numbers, in order to have a more natural encoding of the printer instructions). It is only important to notice that the extension of the size (and the number) of the books, even for orders and orders of magnitude, would enlarge the instruction set just marginally – all we need to do is to increase the limit of the for loop (which can be done very efficiently, e.g. by setting it in the exponential form, like $10^{10^{10^{10}}}$). Therefore, inside there is a book that contains both Hamlet and the full printing instructions while, objectively, the information about everything that exists in the box fits already in the printing instructions alone. If we now consider the agent facing this special book, there is no more doubt that this agent (taking into account only what he reads from his book) subjectively possesses more information than there objectively exists in the entire box system.

But how is it possible that this agent possesses more information than there is and than could be, even in principle, extracted from this box? And how can he be aware of the text of Hamlet if we have concluded that there is no sense in which information about Hamlet exists in this box? While we can easily identify that the information about the text of Hamlet has an actual source in the information about the "self-location" of the agent, this hardly solves the informational mystery - how does this information come about? How can it originate in the objective reality of the box, if this objective reality has no trace of this information?

Let us recapitulate the situation once more: if the camera remains just a camera, so that it makes no sense to speak about its subjective perspective, everything is clear - there is no sense in which information about Hamlet exists in the box. However, if the 3D-printed AI camera has become "conscious", and "it is something like to be that camera", then there must also be an internal perspective to this camera, an internal information system, subjective for the camera. If such subjective experience of the camera exists, it somehow must incorporate information from the book the camera is facing: and that is the information about The Tragedy of Hamlet, in spite that this information objectively does not exist at all in the box (or in the entire universe, if we take nothing else to exists apart from this box).

Since the subjective perspective of a camera is not quite an intuitive concept, it might be tempting to go yet another step further in our thought experiment, and to replace the robotic camera-agent with a human agent: in principle, it should be possible to 3D print an identical human clone in front of each book. However, one should be careful about what information is introduced in the box along with the human agents. Assuming that the agent is 3D printed in adult age, she should arguably also possess standard human faculties, like the ability to read, or even some level of familiarity with works of Shakespeare. In spite of this caveat (and taking care not to contaminate box interior with much relevant information), it seems instructive to consider two hypothetical scenarios.

The first is that, for any reason, such printed agents are philosophical zombies, with no internal perspective. In spite the fact that each of the agents might have a certain behavioral reaction to the information presented in her book (e.g. get prompted to read parts of the text aloud), this will only be an act of an impersonal mechanism and there still would be no information about Hamlet in the box, in any possible sense. Namely, since the time evolution is deterministic, the behavior of the agents cannot change the information initially present in the box - and this initial information contained only the for-loop and no mention of Hamlet. This initial information is merely transcribed into the objective behavior of the agents, more precisely, into their motions. This is no different than copying book content into the digital form by the camera-agents since, by the definition of philosophi-

cal zombies, we can speak of no other perspectives here apart from the "objective" one, i.e. there are no subjective systems of information. In this zombie-case, there would be no informational puzzle of how did the information about Hamlet come into being, since there still would be no trace of the information about Hamlet in the box.

The other scenario is that the human clones would be conscious in the usual sense. While there are much vagueness and dispute about what "being conscious" exactly means, it is at least commonly agreed that it entails the existence of some internal representation of the external world, in other words, the existence of an internal information system. We are somehow "aware" of this internal information system and that is the basic experience of "self-presence" or "self-existence" (or of "being"). This experience in certain sense precedes other manifestations of subjective experience (such as qualia, internal dialogue, etc.) and, when combined with sensory inputs, it necessitates the information about "self-location" in the universe. Namely, that entity we name "consciousness" is strictly localized to only one of the agents and is accessing sensory perceptions of that agent alone. This fact of localization therefore effectively picks one of the books (by picking one of the agents) and, from the perspective of this internal information system, it results in the appearance of a huge amount of (random) information from that book. Consequently, this information from the book is now an integral part of the data in this internal system, i.e. of the full description of the entire universe (or of the entire box) from the perspective of that agent. If we again consider the human agent who is facing that Hamlet copy supplemented with the full printing instructions, it is clear that the Kolmogorov complexity of information in her internal subjective reference system surpasses Kolmogorov complexity of the objective, external description of the box (exactly for the length of the compressed text of Hamlet).

Could it be that this excess of information simply vanishes when we take into account the global picture? Namely, if the copy of Hamlet was the only book in the box, the existence of the information about Hamlet would be indisputable. It is only in the totality of the box, with the myriads of books taken together, that this information vanishes. Could it be a similar case with the internal knowledge of Hamlet, possessed by the agent facing it? May it be that, once we take into account all agents together, that this information also vanishes? Again, this is exactly what happens in the hypothetical case of philosophical zombies. While each clone-agent may have a distinct behavioral response to the presented book, it is due to the totality of these responses and the fact that, as a whole, they merely represent the same "for loop" used to populate books only translated into a "for loop" describing agents' motions, that objectively, there is still no trace of "Hamlet" in that collective motion. However, it is much different if the agents are not zombies, and if we allow that each of them possesses a personal, subjectively experienced system of information. In that case, each of the agents can take into account perspectives of other agents (recognizing that "each of the clones sees *her* book"), but this still is not going to cancel-out the first-person realization that "*I see* the Tragedy of Hamlet *in front of me*". There is simply no "for loop" that will reproduce information existing in the subjective system of the agent facing Hamlet, either in the presence or in the absence of other agents.

IMPLICATIONS AND CAVEATS

To summarize, in the "Library of Babylon" situation that we have analyzed, the presence of the subjective experience of the involved agents was revealed through the existence of informational systems that possessed more information than the totality of information objectively present in the entire physical system. Thus we arrive at the main conclusion of this paper: *Self-locating information possessed by a conscious agent can be of arbitrary high Kolmogorov complexity, while this information nevertheless may not exist at all in the objective description of the physical reality.* As the result, agent's subjective representation of the universe (or of the physical system of which the agent is part) may be of greater Kolmogorov complexity than the full impersonal and objective description of reality (i.e. of the physical system).

Immediate consequence is that, in these situations, this discrepancy between the subjective and objective amounts of information can be used as a way to clarify and define the difference between philosophical zombies and conscious agents (in spite of indistinguishability of behavior).

Next, it is worthwhile to discuss if, and to what extent, this conclusion is further relevant for the various attempts to solve the hard problem of consciousness. We believe that, at the very least, the analyzed informational aspect lets us appreciate more the mystery of the subjective experience. While notions such as "belief", or even "pain" might be sort of vague, unmeasurable and difficult to scientifically define, so that some philosophers see them as nothing but "folk psychology" that should be simply eliminated from a serious description of reality [13], the details about the fate of the Danish prince seem far more factual and difficult to ignore. And these details just do not exist in our settings if there is no subjective experience, since they were never put in the box. Only if the subjective experience truly exists, so does the story Hamlet, in minds of some of the agents.

For this reason, we think that the discrepancy between the subjectively perceived and objectively existing information makes the illusionist (eliminativist) attempts to explain (or explain away) the consciousness (e.g. [14–17]) further implausible. Again, insisting that there is no truly such thing as subjective perspective, or that the latter is in certain sense merely an illusion and that the objective, third-person account is sufficient to describe

reality, is here equivalent to insisting that there is just *no sense* in which any information about Hamlet exists in the box. But, this is hardly plausible because, contrary to this, we just know that if we happened to be that one of these clones facing Hamlet and leafing through the copy, we would, beyond any doubt, obtain some information from the book and that information would be very much real – *at least in some sense*. We know "what-it-is-like" to know about Hamlet. Unlike other more basic phenomenal states, this one also entails a lot of information, so there can hardly be any confusion about its existence.[5] Hence, if the clones have the standard human properties, some of them must also be in the state of "knowing about Hamlet", i.e. there must exist a perspective in which information about Hamlet exists. Therefore, there is less and less room to deny the existence of the subjective experience, as of something puzzling and qualitatively different from the strictly objective reality. Accordingly, the problem to explain why this internal perspective exists – known, in essence, as the hard problem – becomes even more difficult to ignore in this context.

And the prospect to explain this subjective perspective as emerging from the objective physical properties alone also gets grimmer in the light of this information paradox. Namely, in situations of the "library of Babylon" type, the problem with the emergence paradigm becomes more pronounced since this internal perspective is even quantitatively transcending the external, objective one - it contains a part of the information that is not objectively there. How comes that someone can subjectively have more information than objectively exist in the entire universe, if the subjectivity is merely a logical and necessary consequence of the objective reality? Besides, can the experience of contemplating about Hamlet's dilemma be truly derived from the Hamletless objective reality? As in our earlier example with Beethoven's symphony: lack of information about Hamlet in our view demonstrates that the objective reality of the box alone has no explanatory power to explain the appearance of the subjective perspective of the agent reading Hamlet. We see this puzzle as yet another illustration of the same "explanatory gap", often pointed out by the promoters of the hard problem: the gap between the objective physical processes in the brain and the subjective experiences which these processes purportedly should explain. Surely, it is possible to deduce (from the objectively available data) what this subjective perspective of the given agent would contain (in this case, that it would indeed contain information about Hamlet), if we first take for granted that this subjective perspective exists. But this has nothing to do with an explanation of why this perspective should exist in the first place, or how it could come about. And this explanation is farther away once we recognize that the appearance of that subjective perspective would have to bring the story of Hamlet into existence, out of "informational" thin air.

We are aware that a number of relatively obvious objections can be given both to our main conclusion and to our inferences regarding the physicalist approaches to the hard problem. To start with, our entire analysis is contingent upon the ideal uniformity both of the array of the books and of the environment. Indeed, if there is any contamination of the box interior by unwanted information, no matter random or not, the entire construction becomes dubious. While this is not only precluding any real-life realization of the described setup[6] (e.g. as it requires total isolation, perfect vacuum, zero temperature), but also raises questions of whether, even in principle, it makes sense to speak of ideal precision of the initial configuration (personified by the imaginary ideal 3D printer). For example, all distances must be exact to infinite precision, for the Kolmogorov complexity of the system to stay constrained and precise. These requirements further narrow down the list of physical theories to which our conclusions apply – not only that the theories should be deterministic, but also must not contain hidden variables (e.g. these considerations probably make no sense in the context of de Broigle-Bohm pilot wave interpretation of QM).

Nonetheless, we do not think that there is anything in general that precludes the possibility, at least in principle, to have this limited or at least a well-controlled amount of information in a physical system. For example, we find our conclusions applicable in the context of classical physics. But much more interestingly and importantly, there is a serious contemporary scientific view of the universe, even quite a popular one, that assumes the "library of Babylon" type of situation to be a regular, daily occurrence. This is the many-worlds interpretation (MWI) of quantum mechanics (actually, it was the search for the roots of missing information problem in MWI that initially led us to consider this more general case). Namely, according to MWI, not only that the imagined library of Babylon can be actually generated, but it is quite practically feasible, easy even with contemporary technology. What we need is only a relatively modest number of random quantum events with two possible outcomes of equal probability (e.g. 410x40x80x5 of z-spin measurements of a particle with spin along the x-axis). These outcomes are then interpreted as bits, converted into letters of a book (5 bits are sufficient to encode 29 characters, 3 extra combinations can also encode space symbol) and read by an agent (either a robot or

[5] Admittedly, in principle it is possible even to deny this internal reality of "knowing about Hamlet" state, in a similar manner as some philosophers can doubt the existence of pain. But due to the well-defined and nontrivial informational content that the former carries, this seems additionally unconvincing.

[6] We stress that, however, the length of the array is not a real issue - the same conceptual problem with information arrises already with two books, it is only less intuitively obvious.

human). According to MWI, this relatively simple procedure will result in the generation of stunning $32^{410x40x80}$ copies of the agent reading the book, where each instance is absolutely identical in every minute detail apart from the contents of the books (and the latter take all possible sequences of 410x40x80 letters). And these copies are, by the MWI proponents, usually imagined quite literally (as equally coexisting in the richness of the Hilbert space) – essentially in a similar way as human clone-agents coexist in our library of Babylon example.

But the same problem in MWI is actually far more drastic and intuitively striking than in classical settings. From the viewpoint of the MWI formalism, it is in principle possible for the initial state of the universe (even of our universe) to be of extremely low Kolmogorov complexity. It could be such an orderly configuration of particles/fields that its description can fit on a CD (or even on an old floppy drive) and still be – according to MWI – consistent with observational data. Namely, billions of years later such a simple initial state nevertheless may evolve into a state representing a well-defined superposition of myriads of worlds, of which one could be exactly like ours. But since MWI assumes strictly deterministic (and time inversion invariant) laws of motion, the description of this myriad of worlds, each of them being of unfathomable complexity taken alone, when taken together can still fit on that single CD (along with an algorithm that computes the time evolution). This is possible in exactly the same way as each of the billions of books in our library of Babylon contains a lot of information when taken separately, but very little when taken all together. Therefore, many-worlds interpretation, surprisingly, allows that our present universe objectively contains no trace of information not only about Hamlet, but also no information about any human affairs, about any human at all (e.g. about human or any other DNA) or even about Earth itself.[7] If information about all the richness we see around does not objectively exist, and yet we perceive it, it must be that all this information exists only subjectively, encoded in our "self-location" within the Hilbert space containing uncountable worlds. The most surprising is actually that MWI proponents are either ignoring or overlooking this problem, boldly claiming that what they defend is an objectivist interpretation that introduces no further assumptions apart from the main deterministic law of motion (i.e. Schrodinger's equation). In the context of our thought experiment, it already becomes obvious that the MWI idea existentially relies upon a very concrete proposal about the solution to the hard problem. Namely, they tacitly assume that subjective perspective and subjective experience – i.e. those responsible to bring into existence information about all these rich visible structures around us – must spontaneously and necessarily arise as a consequence of specific changes in values of certain parts of the total wavefunction. These "parts of the wavefunction" which exhibit certain patterns of form and dynamics, are then named "conscious agents". Curiously, this counter-intuitive belief is not in MWI taken to be an additional postulate, but is understood to be a logical necessity which MWI supporters do not deem necessary even to plausibly substantiate, let alone prove, in spite that entire interpretation crucially hinges upon this.

This information problem in many world interpretation of quantum mechanics actually has many more important and specific aspects, and thus deserves to be discussed in its own right – which we intend to do elsewhere. But a takeaway relevant for the present paper is that the amount and content of information that is absent from the objective reality and present in subjective experience can be far more drastic than a content of a literary novel: objective reality may not contain even any indication of a human or a brain, and yet there can be a subjective reality existing and observing all the intricate details of human biology (with or without mention of Hamlet).

Another, ostensibly more serious objection to our conclusions would be the following. While we have established that the information content of the considered systems is a constant of motion (due to deterministic laws of physics), this still does not mean that the amount of information, or its content, is not a relative quantity. For example, while the momentum of an isolated physical system is a conserved quantity, its value depends upon the frame of reference. Could it be that any paradox here dissolves, if we take this into account? Moreover, it is very natural to see information as a relative notion. While it would be hard and unusual to define information as a function of position (i.e. coordinate), it is rather standard to see it as a property of a system relative to another system (relative information), especially in the context of physical correlations. If we *define* information in this way, then the information possessed by the agent facing Hamlet - that is, the information *relative* to this agent - simply contains the text of Hamlet. At the same time, the information existing relative to the systems external to the box (e.g. relative to the 3D printer), contains only the printing instructions, no Hamlet. Mystery solved?

We do not think so. Changing the definition can hardly remove the explanatory gap that we observed. We must reiterate our conclusions: it still holds that the only things that objectively exist in the box are just those that we have 3D printed - and there is not a trace of Shakespeare in the printed reality in the box. On the other hand, within the same box, there is, somehow, someone reading and pondering over Shakespeare's verses. Therefore, contrary to our expectations, these verses are very much real: we can easily imagine what it is like to read Hamlet (or any other literary work). If we are indeed

[7] Note that this is not possible in the context of classical physics where trivially simple initial conditions are very unlikely to lead to huge observable complexity (since there is only "one world" according to the classical physics and no additional ones that could "cancel out" the visible complexity).

reading it, there is no doubt that the information we have just read is also real in some sense (e.g. we can clearly evoke it in our mind). Thus, since this happens in the box, then information about Hamlet is in some definite sense real inside the box, in spite of the fact it was not put there in the first place and that it could not spontaneously appear. This matter-of-fact about the experienced reality of information about Hamlet *cannot be a matter of definition*. And the puzzling contradiction, therefore, remains as long as we recognize only objective reality and only the third-person perspective: information about Hamlet then must not exist at all, since - regardless of any definition - all objective information written in the books can be accounted for by a for loop, without a mention of the Shakespeare's work. Defining information as relative *per se* only means that "was there a subjective perspective of the agent facing Hamlet, this perspective would contain information about Hamlet and not something else".

On the other hand, if this information about Hamlet really exists in the box, in spite of not being a part of the *objective reality*, it only means that the information must exist within some "subjective reality". In our view, this new category cannot be introduced within the physicalist framework by a mere formal (re)definition of the notion "information". It takes a deeper modification: a postulate that a perspective different from the objective one is present (at least) in certain situations and in a relation to certain physical subsystems (i.e. to "conscious" agents). Once the existence of this new perspective, i.e. of the "subjective perspective" is *postulated*, its contents can be *recognized as the information relative to this agent*.

Therefore, one way to explain the observed disagreement between the subjective and objective amounts of information is to extend the underlying ontology in order to incorporate this internal information system (i.e. subjective perspective) as something *correlated with, but not logically derivable from* the objective description of the reality. A further step of axiomatic simplification would be to recognize that there is actually no need to postulate the existence of any "objective perspective" *per se*, since the third-person perspective is, in practice, always again a subjective perspective of some other agent. In this view, ontological reality would be granted only to subjective perspectives, and the real question would be how these perspectives (i.e. internal information systems attached to various agents) can be interrelated in a consistent manner.

Another approach would be to violate the conservation of information by giving up determinism. Indeed, if there is a frequent influx of huge amounts of random information, our entire analysis and conclusions would be no longer applicable.

Finally, we note that the collapse postulate of quantum mechanics can be seen in this light as a combination of both of these ideas. First, the collapse postulate tacitly implies the existence of an internal perspective (the one from which the measurement is performed) and, therefore, this internal perspective can be seen as effectively introduced by this postulate. Secondly, the measurement outcomes, as occurring from this internal perspective, are well defined and inherently random, thus they are introducing fresh amounts of information and are increasing Kolmogorov complexity of the internal description of reality. In this sense, the collapse postulate of QM can be recognized as a step towards the solution to the hard problem of consciousness. As for the particular puzzle of the subjective information that is objectively missing, which was the subject of this paper – it all but solves it. However, a deeper and much more difficult question, with numerous possible answers, still lingers: to which entities, systems or agents is this "subjective perspective" granted?

SUMMARY

The subject of exploration in this paper was the discrepancy, occurring under certain conditions, between the information subjectively possessed by an agent and the totality of information objectively existing in the physical system to which the agent belongs. We pointed out that, surprisingly, the amount of information existing subjectively (i.e. its Kolmogorov complexity) can surpass by an arbitrary extent the amount of information existing objectively. We explicated this phenomenon by devising the "library of Babylon" thought experiment and discussed its relation to the hard problem of consciousness.

Our main conclusions were: i) the phenomenon of the missing information is a curiously quantifiable manifestation of the hard problem of consciousness, and, as such, can be also used to clarify, from yet another aspect, the difference between conscious agents and philosophical zombies; and ii) this informational puzzle seem to further diminish the prospect of ever finding fully physicalist explanation of the subjective experience. Furthermore, we identified postulating the subjective perspective and introduction of chance in physical dynamics as two possible solutions to the presented conundrum.

We have also identified that a particularly severe and relevant manifestation of this problem, deserving a detailed analysis, arises in the context of the many-worlds interpretation of quantum mechanics. This aspect will be the subject of our future study.

[1] D. Chalmers, *Facing up to the problem of consciousness*, Journal of Consciousness Studies 2(3):200-219, (1995).

[2] A. N. Kolmogorov, "On tables of random numbers",

Sankhy Ser. A 25, 369376 (1963).
[3] A. N. Kolmogorov, “Three Approaches to the Quantitative Definition of Information”, Problems Inform. Transmission. 1 (1): 17 (1965).
[4] R. Kirk, “Sentience and Behaviour”, Mind, Volume 83, Issue 329, Pages 4360 (1974) https://doi.org/10.1093/mind/LXXXIII.329.43
[5] D. Chalmers, “The conscious mind”, New York: Oxford University Press (1996).
[6] M. G. Titelbaum, “The Relevance of Self-Locating Beliefs”, The Philosophical Review, Vol. 117, No. 4, pp. 555-605 (2008).
[7] C. T. Sebens, S. M. Carroll, “Self-locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics”, The British Journal for the Philosophy of Science, Volume 69, Issue 1, Pages 2574 (2018).
[8] J. D. Norton, “Causation as Folk Science”, Philosophers’ Imprint. 3 (4): 122. (2003) hdl:2027/spo.3521354.0003.004
[9] Hugh Everett (1955) “The Theory of the Universal Wavefunction”, Manuscript, pp 3140 of Bryce DeWitt, R. Neill Graham, eds, The Many-Worlds Interpretation of Quantum Mechanics, Princeton Series in Physics, Princeton University Press (1973), ISBN 0-691-08131-X.
[10] Vaidman, Lev, “Many-Worlds Interpretation of Quantum Mechanics”, The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.)
[11] P. W. Humphreys, “Is ‘Physical Randomness’ Just Indeterminism in Disguise?, in PSA 1978, volume 2, Peter D. Asquith and Ian Hacking (eds.), Chicago: University of Chicago Press, pp. 98113 (1978).
[12] T. Nagel, “What Is It Like to Be a Bat?”, The Philosophical Review. 83 (4): 435450 (1974).. doi:10.2307/2183914
[13] Ramsey, William, ”Eliminative Materialism”, The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.)
[14] K. Frankish, Illusionism as a Theory of Consciousness, Journal of Consciousness Studies, 23: 1139 (2016)
[15] Daniel C. Dennett, “From Bacteria to Bach and Back: The Evolution of Minds”, W. W. Norton & Company 2017, ISBN 978-0-393-24207-2
[16] Patricia S. Churchland and T. J. Sejnowski, “The Computational Brain”, (1992) . Cambridge, Massachusetts: The MIT Press.
[17] Webb TW, Graziano MS (2015). “The attention schema theory: a mechanistic account of subjective awareness”. Front Psychol. 6: 500. doi:10.3389/fpsyg.2015.00500