Keith Wiley Interview for Interalia Magazine

(16/02/15)

Richard Bright: Can we begin by talking about your background? What made you take up computer science and, in particular, your interest in mind uploading?

Keith Wiley: I've been programming computers since I was seven years old. The field I initially fell in love with was artificial life, i.e., evolutionary algorithms, emergence simulations (flocking), genetic algorithms, simulated ecosystems, simulations of evolutionary phenomena, cellular automata, etc. I started graduate school in C.S. intending to following a-life and then pivoted multiple times and eventually landed on a thesis that mostly focused on surface topology and user-interfaces (primarily affordances). I discovered mind-uploading in 1997 (toward the end of my Bachelor's) when I found the book '*Beyond Humanity: Cyberevolution and Future Minds'*. With the idea cemented in my awareness, I then followed it, somewhat tangentially, for the next seventeen years. Ultimately, in 2014, I wrote and published a book on the topic and I am currently pursuing some other projects in the same area.

RB: Can you give an overview of your book, *A Taxonomy and Metaphysics of Mind-Uploading*? What are its aims?

KW: This book has two aims, which are presented in its first two sections.

The first is to offer a thorough collection (a taxonomy) of thought experiments and hypothetical scenarios that have arisen over the years during philosophical consideration of mind-uploading. Many new-comers to the field derive the simpler scenarios on their own and then offer them up in debates. While I think this curiosity is fantastic, such debates remain locked at an initial level of discussion; the same ideas are presented over and over again, and the same counter-arguments are offered over and over again. It takes time to repeatedly cover the prototypical discussion, and consequently few debates have the opportunity to advance to new ground. My goal with the taxonomy is to make those common initial scenarios more accessible so that people with similar interests can enter the discussion from a more informed position. Furthermore, the deeper and more nuanced variants within the taxonomy present serious challenges to popular early positions and conclusions. By becoming aware of these subtler counter-arguments, readers should be able to avoid falling into seemingly straightforward yet fundamentally problematic or paradoxical conclusions. Besides, the subtle variants within the taxonomy are positively fascinating. It's simply a lot of fun to discover all these thought experiments gathered in a single resource.

The second section of the book presents my personal theory of what minds truly are, how they relate to brains, and what sorts of transformations should be possible on minds. The conclusion of the second section is the climax of the book, in which I argue for what I call the "primacy of the claim to the identity of the original mind". I argue that all minds resulting from a mind-uploading procedure should receive equal "primacy" in this regard. Namely, I argue that even in nondestructive scenarios from the taxonomy (those in which the biological original person and brain survive) the mind associated with that brain should not enjoy greater primacy in its claim to the original identity than any uploaded minds that resulted from the same procedure. Rather, all minds following the procedure should be regarded as equal descendants of a common ancestral mind. This is not a popular position in such debates, which is precisely why I argue for it so thoroughly in the book. I see this issue as a form of prejudice in fact, one I am striving to alleviate.

RB: Historically, conceptions of the brain have included various mechanical devices, telephone and telegraph metaphors and, currently, the model of the computer. Does the brain 'behave' like a computer, or is it a computer**?**

KW: The word "computer" poses significant problems to such analogies. A widely known and humorous anecdote is that "computer" used to refer not to mechanical devices, but to people (mostly women) who performed rote mathematical calculations for other scientists and engineers. So, the word has evolved dramatically already, and it will continue to do so. Does the brain behave like the Von Neumann architectures we have relied on since the 1960s? There are both similarities and differences, but is it literally "the same thing as" a Von Neumann device? In one sense, no, quite simply because the brain is not a Von Neumann architecture itself, but in another sense that doesn't really matter and the answer can be interpreted as yes. Whether the brain is, for all intents and purposes, the same thing as the kinds of computers we currently use, hinges on whether the brain is Turing Complete (or even better, Turing Equivalent). Since all Turing Complete computers are, in a particular way, identical to each other, then if the brain is Turing Complete, it is definitively identical to all other Turing Complete computers. And we know that our conventional computers are Turing Complete. So, is the brain TC then? Some researchers, most famously Penrose and Hameroff, have explicitly argued that it isn't. But a lot of people disagree with them. If we gave equal credence to every counter-theory that is ever presented, then we could never claim to known anything. While there are detractors to the argument that the brain is TC, I personally think it is. And that means that in a mathematical sense, the brain and a conventional computer are synonymous in every way that really fundamentally matters.

However, TC systems can come in myriad forms. All sorts of systems have been determined to possess Turing Completeness, including systems as different from Von Neumann architectures as cellular automata, slime molds, and of course neural networks. One of the biggest distinctions is surely between serial and parallel systems. The brain is parallel on the order of 100 billion neurons, or better yet, a quadrillion simultaneously functioning synapses. We haven't built computers like that yet. Furthermore, the brain's parallel organization is different from our paltry attempts at parallel computers thus far. Our engineered parallel architectures have generally sacrificed at least one dimension of potential power. For example, they may be organized into a hierarchy of master and slave nodes, thus losing resilience to single-point-failures, or they may be designed as single-instruction-multiple-data systems (SIMD), which perform the same operation on multiple pieces of data but cannot perform varied operations across those data, or they may be truly MIMD so as to achieve full generality of function, but only at the cost of housing far fewer processing units than SIMD architectures (MIMD units are often counted in the tens or hundreds while SIMD units are often counted in the thousands and up). Our efforts to create truly massively parallel systems are still confined to extremely large and expensive super-computers, and as yet, even these behemoths have not quite matched common estimates of the brain's computing power. We simply have not encountered computerized brains yet, not as I intend the term — but we will.

With regard to brain-like architectures, we have simulated brain-like systems in software-level neural networks (some people criticize the simplicity of such models, but that misses the point, which is to contrast them with conventional functional or declarative algorithmic approaches which are even less brain-like), but the hardware on which such neural networks run, although often parallel, is still strictly structured, managed, overseen, and controlled. I am really excited for the day when we start building massively parallel brain-like computers. Neuromorphic chips hint at such prospects, but are nascent. I want to see true computerized brains at scale, essentially artificial constructions of brains. That's what I'm holding out for.

Alternatively, while one approach is to build computerized brains, another is to attain sufficient computing power to run whole brain emulations (WBE) on slightly more conventional architectures (but at greater computing power than our current systems) so as to achieve software implementations of brain functionality as opposed to hardware analogs to the brain.

RB: What is 'Brain State Space' and 'Mind State Space' and how is this related to 'Whole Brain Emulation'?

KW: The first two terms are introduced in my book. They aren't common outside the book yet (although I hope they catch on). In section two, I present my theory of mind and how it relates to the brain. Brain-state space and mind-state space refer to combinatorial collections of independent variables such that the entire space contains all possible combinations of those variables. The original such space was Borges' Library of Babel, the "space of all possible books". Another popular example is the Dennett's Library of Mendel, the "space of all possible genetic sequences". Confine a book to 410 pages, forty lines per page, eighty characters per line, drawing each character from an alphabet of 25 symbols (Borges' prescription). A finite 1,312,000 character sequence now represents the book. The Library of Babel simply contains all possible sequences that could fill that book (25 ^ 1,312,000 combinations). Dennett's space contains all possible DNA sequences of a given length. Brain-state space is the space of all possible physical configurations of a brain. We have to choose a physical resolution, and I leave it to the reader since I see no need to prescribe such a parameter in the book. It could be neural, molecular, atomic, subatomic, whatever. Once you pick a spatial resolution, there is now a finite number of ways you can configure matter within the volume of a brain to actually build a brain. Put an atom here, or put it there. Configure a synapse this way or the next possible way in the space of all possible synapse configurations. Etc. One may balk at the sheer number of possible configurations, but that is beside the point. Brain-state space simply indicates the concept of a (barely) limited number of physical brain configurations that could ever exist, such as your brain's material configuration at this very moment — which is then different a moment later, and so on, such that your life consists of a temporal sequence of brain-states.

Mind-state space is more vague, but as I will show, it can conform to the same principles. Your mind is in some all-encompassing cognitive state at this very moment. That state represents your genetically innate personality characteristics coupled with the totality of your life's experiences and memories leading up to this point in time, but in a sense the actual history is incidental. What matters is the instantaneous momentary state, you right now, even though that instantaneous state feels like it encodes your history. Mind-state space proposes the space of all possible such mind-states. While mind-states admittedly feel fuzzier than raw material configurations like brain-states we can actually reduce them to brute brain-states by observing that brain-states instantiate mind-states in a very concrete fashion. Every brain-state yields precisely one corresponding mind-state (although I wouldn't say the opposite is necessarily true). Since mind-states rely on some physical brain-state to instantiate them in reality, we can see that mind-states are actually just as concrete an idea as brain-states.

After defining brain-states and mind-states in the book, I then define a mind as being a temporal sequence of mind-states.

On the original question, I wouldn't say that any of this is too crucially relevant to WBE, although everything fits together of course. If WBE can successfully emulate a brain's state, then it will, by implication, instantiate the same mind-state as well.

RB: Do you see embodiment as an important aspect of mind uploading?

KW: We are creatures of our bodies as has been our lineage for billions of years. It is often proposed that the human mind positively requires a bodily experience. However, our sense of self clearly survives astounding physical truncations and biological failures, as exhibited by amputees, paraplegics, the blind, deaf, etc. I'm not sure how much of a person's body or sensory experience you can take away before their mind somehow fails to be justifiably human. We have arguably never discovered such a limit, given that we have not denied full personhood to any such victim of injury (or does allowing both the state and next of kin to make termination-of-life-support decisions about coma and/or vegetative patients qualify as precisely such a judgement?).

However, one might ask a related question. While adults can survive bodily losses with their minds intact, to what extent can an infant grow into an adult from an initially physically impoverished state? Again, we know people can grow up with intact minds despite the loss of both peripheral limbs and internal organs early in life, and those born blind and deaf can grow up just fine too, but let us consider the extreme case for a moment: a human born with no sensory experience at all, including tactile feedback from physical interaction. They would have no way to receive any information from the world, there would be no feed-back loop of experience, learning, modeling, and refinement of cognitive development. I assume that a human brain, born an infant and provided absolutely no external world with which to interact, would essentially fail to develop into what we consider an adult human mind (note that even this dramatic conclusion is merely a presumption; it could be wrong). So while embodiment may be less crucial to the maintenance of an established adult, it is probably more crucial to the on-line process of development.

Finally, I would argue that while human minds might require some sort of embodied cognitive self-image or embodied feedback, perhaps minds in general don't. Perhaps AI can be created in such a way that it does not require an embodied experience. This possibility would argue that while our uploaded selves might require some form of embodiment, simulated or otherwise, AI may nevertheless not require anything of the sort. Rodney Brooks and some leading AI researchers have proposed that even abstract nonhuman-originated AI will require embodiment as well, so it would seem we have a lot left to discover on this topic.

Since my response seems to argue both for and against embodiment, I am apparently not fully committed to an answer to this question yet.

RB: Can you say something about Identity between the original and the upload?

KW: The central thesis of my book is that while the original and all resulting uploads (there could be more than one of course) should certainly be regarded as possessing distinct identities, they should be granted equal primacy in their claim to the original identity. This raises a paradox. How can there have been one identity to start, then there be multiple identities later, such that the multiple identities are regarded as distinct from one another, yet such that they are all considered equally valid continuations of the original? As it stands, this phrasing is paradoxical. The resolution is what I call splitting of a mind (what others have called divergence, fission, and branching; it's the same idea every time). When a cell divides via mitosis, it yields two daughter cells. Which one is the original? There was one to begin with but now there are two. How do we resolve the paradox of cellular identity? Is one cell more original than the other? Does that question even make sense? In my opinion, this is exactly the same issue that occurs with minds during a nondestructive uploading procedure.

Some people will counter-argue that a crucial difference from the mitosis analogy is that one resulting mind is housed in the original and unharmed biological brain while the other resides in a wholly different substrate, a new physical system. However, that point confuses the distinction between brains and minds. We can admit that one of the physical brains in question is more original than the other, but the metaphysical minds they instantiate do not fall to the brute matters of atoms. Much in the same way that different physical copies of a book may all be said to instantiate the same metaphysical story, regardless of the age of the various tomes or whether one hardcopy was produced by visual and scripted copying of another, we can say the same thing of minds. All minds are equal moment-to-moment descendants of their recent (and ancient) mind-ancestors. The raw and tangible issues concerning the age of various molecular compositions (i.e., physical things) are simply irrelevant to metaphysical abstractions like minds (or any nonphysical identity for that matter). As I see it, the preference to associate identity with particular physical instantiations is essentially a category error.

Some people will remain unconvinced by my counter-argument. One of my primary goals is the alleviation of paradoxes in the numerous thought experiments posed in the taxonomy. The taxonomy poses a wide variety of hypothetical scenarios that pose noteworthy challenges to all identity theories except branching identity. That is the only theory that seems to survive the paradoxical challenges unscathed. For example, as I show in the third section of the taxonomy, what if the brain is physically cleaved in two, each half duplicated in computational substrate, and the halves finally reattached to one another, thereby producing two brains, each consisting of half original biological substrate and half uploaded computerized substrate? This scenario represents a literal analogy to the cellular mitosis described above, and thus vastly weakens the claim that uploading procedures and debates fail to replicate other branching analogies (such as mitosis). What theory of mind and identity could possibly accommodate this uploading procedure except allowing a true splitting of one mind into two equal descendants regardless of physical substrate? But once we break the reliance of mind upon the underlying physical instantiation, all physical challenges crumble and we find ourselves forced to grant equal primacy of identity even in those scenarios which were originally more resilient to such equality, such as nondestructive scan-and-copy uploading. The book goes over these issues in much more detail than I can reasonably offer here.

I would love for readers to give the book a fair chance and not judge it on short interviews, for only in its full exposition do I believe I have done the argument justice. If I could have completed the argument in a shorter space, then I would have written a shorter book.

RB: What is the 'White Room Thought Experiment' and what are its implications?

KW: In the White Room I propose a thought experiment in which a subject submits to an unconscious nondestructive mind-uploading procedure (i.e., they undergo anesthesia and then both the original and upload awaken from unconsciousness after the procedure is completed). The thought experiment further prescribes that both minds (both people) later awaken in minimalist white rooms, essentially isolated tiny physical universes from which they can be observed for the remainder of their lives. If we further stretch our concept of reality to include physical determinism (I say stretch because modern physics is undecided as to whether our reality is physically deterministic) we can then draw two interesting conclusions about these circumstances. First, the two people would act in physical concert not only for seconds or minutes following their awakening, but for the rest of their lives. Years after the procedure occurred, they would physically move in perfect tandem from one moment to the next. It would be uncanny to observe. Furthermore, they would be mentally identical as well. They would experience the same sequence of thoughts for the rest of their lives. If one thought about broccoli on the 15,000th day following the procedure (forty years), the other would think about broccoli at the exact same moment. The second conclusion is that we may as well regard these two people as actually having the same mind, not just two identical minds or something to that effect.

RB: Inevitably, do you think a machine can ever be conscious? And, would determining whether a machine is conscious rely on its behavioural aspects only?

KW: I hold the fairly common position that consciousness is — for whatever reason and in whatever way — an emergent consequence of the brain's functionality. Chalmers does an excellent job of showing just how difficult it is to properly associate physical phenomena like neural behavior with our higher conscious experience, and I agree that there is a marvelous mystery there awaiting further discovery. But nonetheless, I deem consciousness to be the result of that massively parallel neural symphony. Since I believe it is fundamentally the functionality of the brain that produces the emergent conscious experience, I clearly believe that similar functionality in other substrates should yield similar effects, so yes, I think a "machine" (something of a pejorative term) can be conscious. I don't think the brain's ribosomes, proteins, or lipids are crucial to consciousness. I don't see how biological material could possibly be of relevance (I believe Searle would disagree). I think it's the behavior that is relevant. Walking is a function that some biological systems happen to exhibit, but it is in no way a solely biological endeavor. Non-biological machines with mechanized legs can walk and it is true walking when they do it, not artificial or simulated walking. Walking is walking. If physical structures like what we call limbs engage in particular sequences of motions, then walking has occurred. Real walking. The same logic applies to brains, their functionality, and any consequent minds and consciousness.

That said, we face two challenges. One, we don't necessarily know precisely which neural functionality is tied to consciousness and which isn't. Consequently, we theoretically run the risk of creating brains in non-biological substrates that reproduce certain functions but not others, and we may fail to produce consciousness (a risk that is alleviated by ever advancing neuroscience as we continually hone our understanding of all matters brain and mind). This risk, manageable though I believe it to be, leads to the second challenge. I'm not sure how we can ever verify the consciousness of a machine. After all, we can't verify the consciousness of other humans. We merely assume other humans are conscious because we ourselves are conscious and we know everyone is biologically similar. As a side-note, some researchers have claimed to develop methods for consciousness verification in locked-in syndrome patients, but these methods actually detect intentional behavior (essentially what psychologists call verbal report), not consciousness in the sense we are discussing here; in a purely theoretical sense, such patients (or anyone for that matter) could always be Chalmers' infamous zombies, and that is the risk I pose at the beginning of this paragraph.

Since a computer or machine will be different in various potentially substantial ways, it is more difficult to grant the assumption of consciousness to such a system. How will we know that the differences weren't crucial (perhaps you can't be conscious without ribosomes after all and we were wrong to ever think otherwise!) The best solution is that we must establish certain assumptions about which functions underlie consciousness. We must make those assumptions as reasonably as we possibly can, such as the popular contemporary assumption that collective neural spiking underlies all of human thought as opposed to idiosyncratic intra-neural chemistry, which is not generally believed to be the basis of human thought. We need similar assumptions for the ultimate cause of consciousness. With those assumptions laid down, we can then label any system, biological or otherwise, as either satisfying the criteria or not. Some might argue this is arbitrary ground from which to grant or deny consciousness, but I believe that ongoing neuroscience and cognitive science will provide us with increasingly confident criteria for making such a prescription. There is an incredible amount left to learn about the brain and how it produces the mind. In the future we will have a far more complete science of the brain and mind, and it is from that vastly more knowledgeable position that we will confidently establish trustworthy criteria for determining the conscious status of various cognitive systems.

Incidentally, I think a case can be made that it might be impossible to fake conscious behavior at the organismal level. This is sort of a Turing test but for consciousness as opposed to mere intelligence. If an agent behaves in such a way as to convey consciousness, perhaps it absolutely must be conscious to accomplish such a feat. A lot of people do not seem to agree with this view, and I admit I only propose it to keep the idea alive, but not out of strong commitment. Many people seem to have no trouble imagining artificial intelligence and robots that exhibit behavior of truly human versatility but which utterly lack any inner experience. Such beings are essentially Chalmers' zombies. I am not as convinced that such machines are possible. If they seem human, perhaps they simply must be so in order to accomplish the charade. If this theory was true, then verifying the consciousness of uploads would be a cinch. We would simply interview them. As I conceded above, I'm not committed to this idea myself, but I think it's worth keeping warm.

RB: It could be said that what has determined our humanity is the trajectory of our technology. How do you think we will interface with technology in the future?

KW: While it may not sound too profound to agree with the status quo, I like the prediction that we will move toward increasingly direct mind interfaces (I say mind interface and not neural interface because I am focusing on the cognitive traits of such interfaces, not their physiological attachments, and I also don't say thought interface because I show below that we may actually skip the thought stage entirely). In the extreme, we will desire a thing, event or outcome, and it will simply come to pass. We used to write with a brush, then a pen or pencil, then a typewriter, then a computer. Recent advances have included tools like Swype which abstract typing from the character level to the word level (typing, of course, abstracted from the stroke level to the character level). Moving past typing we have made strong inroads toward speech-to-text dictation, eye-tracking, etc., to say nothing of the tools used by paraplegics, ALS patients, etc. All of this portends an obvious trajectory: we are gaining increasingly direct access from thought to action while at the same time we are representing our ideas at increasing abstract levels (from strokes to characters to words, then phrases and on up). Eventually we will speak sentences to ourselves in our heads (psychologists call this inner speech) and the writing will simply appear. And soon after that, we may even move past inner speech to having our thoughts materialize as writing faster than we can even convert those thoughts into language in our heads (the sentences in our heads occur, of course, after the fact of the thought they are expressing, so perhaps we could skip the conscious awareness of our own thoughts and go directly from a preconscious mental idea to its written linguistic expression). That would be a very strange sensation.

I have focused on writing so far, but the same direct mind interfacing with other systems could proceed along similar lines. Fully implemented, under such a system the physical world around us would appear to be in a constant state of physical flux as it materially shifts and flutters on the fly in response to our needs and desires, possibly faster than we can even consciously realize or articulate those desires as conscious inner speech. While this experience would be jarring for us today, such technologies will evolve steadily and arrive smoothly. By the time such technologies are possible, we will be fully prepared for them.

Traditionally, interfacing has been the problem of how to access human sensory and motor modalities: how do we present data as visual, auditory, or tactile presentations, and then how we convert a person’s physical movements (or secondary effects, such as vocalizations), into instructions back into the system, thereby closing the user-interface loop? My PhD research considered how to design software interfaces that manipulate a purely virtual system in the most physically analogous way possible. User interfaces prescribe strict barriers between the person and the system. Those barriers are the interface of course, and they are eventually going to dissolve. We and the systems we interact with will steadily cease to be distinct entities requiring interfaced message-passing. As we incorporate increasingly invasive technology into our brains, and then as our technologies become increasingly intelligent and exhibit increasingly thought-like properties, we will meet in the middle and become cognitively synonymous entities. Translating back and forth won’t even make sense in that context; it will all be the same connected physical and cognitive system. In so doing, we may attain that pinnacle of eastern religions: we will become one with the universe. But even more profoundly, if my thoughts merge with the physical world and your thoughts merge with the world, and if it’s all the same world, then we will not only merge with the physical world, but also with each other!

I have never had the opportunity to follow this reasoning to its conclusion before. Good interview question. Thank you.

You hinted that our humanity is determined by such systems. If the mystic conclusions I fell into above offer any insight, then clearly this will have a profound impact on our humanity.

RB: In order to fully begin to fathom the mind, some theorists have proposed we will need new sciences formed by fusing together ones we already have – for example physics, computer science, neuroscience and cognitive science, in as-yet-undreamt of ways. Do you agree with this? And are there other disciplines that should be included?

KW: We will certainly have to evolve some of our concepts. As I described above, our application of the word “computer” is going to change radically as we eventually start building true computerized brains: full-scale hardware analogs of neural brain structure and functionality. Likewise, I can envision some sort of grand synthesis of computer science, neuroscience, and cognitive psychology that we currently have no access to. And of course, in the end, everything comes down to math. There must be some unified math that will explain not only brains, but also minds, and even consciousness. Finding such a math may be harder than the physicists’ theories and goals of a unified physics, as evidenced by that fact that we have already made great strides on the latter (e.g. the tremendous progress on the Standard Model) while the former remains almost comically whimsical in our current era.

You asked which additional disciplines should be included. Without math, there is nothing.

RB: What role does artistic activity play in furthering our understanding of consciousness?

KW: Honestly, I don’t know. Art, in all its forms, even when we can’t clearly define or identify it, seems to be central to human experience. There is practically no such thing as a non-artistic human. Admittedly, people exhibit varying technical skills in regard to art, so there is variation in the balance between production and consumption, but art is simply part and parcel with the human condition. I’m sure we can’t possibly strip art and consciousness away from each other and study them in isolation. Beyond that claim, we should all spend a lot more time producing and consuming art than we currently do in our hectic workaholic society.

Lastly, and this isn’t a bad note to end on, if art and mind are as intimately tangled as I suggest, then the advanced minds of the future, either artificial or uploaded, should bring about art in the universe of a complexity and on a scale unimagined in our era.