doi.org/10.53975/9r72-i78l

Seeds of Science

Taxonomies of Intelligence: A Comprehensive Guide to the Universe of Minds

Roman V. Yampolskiy[1]

Abstract

This paper explores the landscape of potential mind architectures by initially conceptualizing all minds as software. Through rigorous analysis, we establish intriguing properties of this intellectual space, including its infinite scope, variable dimensions of complexity, and representational intricacies. We then provide an extensive review of existing taxonomies for mind design. Building on this foundation, the paper introduces 'Intellectology' as a new field dedicated to the systematic study of diverse forms of intelligence. A compendium of open research questions aimed at steering future inquiry in this nascent discipline is also presented.

Introduction

In 1984 Aaron Sloman published “The Structure of the Space of Possible Minds” in which he described the task of providing an interdisciplinary description of that structure [1]. He observed that “behaving systems” clearly comprise more than one sort of mind and suggested that virtual machines may be a good theoretical tool for analyzing mind designs. Sloman indicated that there are many discontinuities within the space of minds meaning it is not a continuum, nor is it a dichotomy between things with minds and without minds [1]. Sloman wanted to see two levels of exploration namely: *descriptive* – surveying things different minds can do and *exploratory* – looking at how different virtual machines and their properties may explain results of the descriptive study [1]. Instead of trying to divide the universe into minds and non-minds he hoped to see examination of similarities and differences between systems. In this work we attempt to make another step towards this important goal.[2]

[1] University of Louisville, USA; roman.yampolskiy@louisville.edu

[2] This paper is an extended version of Chapter 2 from Dr. Yampolskiy’s book – Artificial Superintelligence: a Futuristic Approach © 2015 by CRC Press. An earlier version of this paper was presented at the Artificial General Intelligence (AGI2015) Conference in Berlin, Germany on July 23, 2015.

What is a mind? No universal definition exists. Humans are said to have a mind. Higher order animals are believed to have one as well and maybe lower level animals and plants or even all life forms. We think that an artificially intelligent agent such as a robot or a program running on a computer will constitute a mind. Based on analysis of those examples we can conclude that a mind is an instantiated intelligence with a knowledge base about its environment, and while intelligence itself is not an easy term to define, the work of Shane Legg provides a satisfactory, for our purposes, definition [2]. Additionally, some hold a point of view known as Panpsychism, attributing mind-like properties to all matter. Without debating this possibility, we will limit our analysis to those minds which can actively interact with their environment and other minds. Consequently, we will not devote any time to understanding what a rock is thinking.

If we accept materialism, we have to also accept that accurate software simulations of animal and human minds should be possible [3]. Those are known as uploads [4] and they belong to a class of computer programs no different from that to which designed or evolved artificially intelligent software agents would belong. Consequently, we can treat the space of all minds as the space of programs with the specific property of exhibiting intelligence if properly embodied. All programs could be represented as strings of binary numbers, implying that each mind can be represented by a unique number. Interestingly, Nick Bostrom via some thought experiments speculates that perhaps it is possible to instantiate a fractional number of mind, such as .3 mind as opposed to only whole minds [5]. The embodiment requirement is necessary since a string is not a mind, but could be easily satisfied by assuming that a universal Turing machine is available to run any program we are contemplating for inclusion in the space of mind designs. An embodiment does not need to be physical as a mind could be embodied in a virtual environment represented by an avatar [6, 7] and react to a simulated sensory environment like a brain-in-a-vat or a “boxed” AI [8].

Infinitude of Minds

Two minds identical in terms of the initial design are typically considered to be different if they possess different information. For example, it is generally accepted that identical twins have distinct minds despite exactly the same blueprints for their construction. What makes them different is their individual experiences and knowledge obtained since inception. This implies that minds can’t be cloned since different copies would immediately after instantiation start accumulating different experiences and would be as different as two twins.

If we accept that knowledge of a single unique fact distinguishes one mind from another we can prove that the space of minds is infinite. Suppose we have a mind M and it has a favorite number N. A new mind could be created by copying M and replacing its favorite number with a new favorite number N+1. This process could be repeated infinitely giving us an infinite set of unique minds. Given that a string of binary numbers represents an integer we can deduce that the set of mind designs is an infinite and countable set since it is an infinite subset of integers. It is not the same as a set of integers since not all integers encode for a mind.

Alternatively, instead of relying on infinitude of knowledge bases to prove infinitude of minds we can rely on the infinitude of designs or embodiments. Infinitude of designs can be proven via inclusion of a time delay after every computational step. The first mind would have a delay of 1 nano-second, the second a delay of 2 nano-seconds and so on to infinity. This would result in an infinite set of different mind designs. Some will be very slow, others super-fast, even if the underlying problem solving abilities are comparable. In the same environment, faster minds would dominate slower minds proportionately to the difference in their speed. A similar proof with respect to the different embodiments could be presented by relying on an ever increasing number of sensors or manipulators under control of a particular mind design.

Also, the same mind design in the same embodiment and with the same knowledgebase may in fact effectively correspond to a number of different minds depending on the operating conditions. For example, the same person will act very differently if they are under the influence of an intoxicating substance, under severe stress, pain, sleep or food deprivation, or are experiencing a temporary psychological disorder. Such factors effectively change certain mind design attributes, temporarily producing a different mind.

Size, Complexity and Properties of Minds

Given that minds are countable they could be arranged in an ordered list, for example in order of numerical value of the representing string. This means that some mind will have the interesting property of being the smallest. If we accept that a Universal Turing Machine (UTM) is a type of mind, if we denote by (*m*, *n*) the class of UTMs with *m* states and *n* symbols, the following UTMs have been discovered: (9, 3), (4, 6), (5, 5), and (2, 18). The (4, 6)-UTM uses only 22 instructions, and no standard machine of lesser complexity has been found [9]. Alternatively, we may ask about the largest mind. Given that we have already shown that the set of minds is infinite, such an entity does not exist. However, if we take into account our embodiment requirement the largest mind may in fact correspond to the design at the physical limits of computation [10].

Another interesting property of the minds is that they all can be generated by a simple deterministic algorithm, a variant of Levin Search [11]: start with an integer (for example 42), check to see if the number encodes a mind, if not, we discard the number, otherwise we add it to the set of mind designs and proceed to examine the next integer. Every mind will eventually appear on our list of minds after a predetermined number of steps. However, checking to see if something is in fact a mind is not a trivial procedure. Rice's theorem [12] explicitly forbids determination of non-trivial properties of random programs. One way to overcome this limitation is to introduce an arbitrary time limit on the mind-or-not-mind determination function effectively avoiding the underlying halting problem.

Analyzing our mind-design generation algorithm we may raise the question of complexity measure for mind designs, not in terms of the abilities of the mind, but in terms of complexity of design representation. Our algorithm outputs minds in order of their increasing value, but this is not representative of the design complexity of the respective minds. Some minds may be represented by highly compressible numbers with a short representation such as 10^{13}, while others may be composed of 10,000 completely random digits [13]. We suggest that Kolmogorov Complexity (KC) [14] measure could be applied to strings representing mind designs. Consequently some minds will be rated as “elegant” – having a compressed representation much shorter than the original string while others will be “efficient” representing the most efficient representation of that particular mind. Interesting elegant minds might be easier to discover than efficient minds, but unfortunately KC is not generally computable.

In the context of complexity analysis of mind designs we can ask a few interesting philosophical questions. For example could two minds be added together [15], in other words, is it possible to combine two uploads or two artificially intelligent programs into a single, unified mind design? Could this process be reversed? Could a single mind be separated into multiple non-identical entities each in itself a mind? Additionally, could one mind design be changed into another via a gradual process without destroying it? For example could a computer virus (or even a real virus loaded with DNA of another person) be a sufficient cause to alter a mind into a predictable type of other mind? Could specific properties be introduced into a mind given this virus-based approach? For example could Friendliness [16] be added post factum to an existing mind design?

Each mind design corresponds to an integer and so is finite, but since the number of minds is infinite some have a much greater number of states compared to others. This property holds for all minds. Consequently, since a human mind has only a finite number of possible states, there are minds which can never be fully subsumed by a human mind as such mind designs have a much greater number of states, making their subsumption impossible as can be demonstrated by the pigeonhole principle.

Space of Mind Designs

Overall, the set of human minds (about 8 billion of them currently available and about 100 billion ever existed) is very homogeneous both in terms of hardware (embodiment in a human body) and software (brain design and knowledge). In fact the small differences between human minds are trivial in the context of the full infinite spectrum of possible mind designs. Human minds represent only a small constant size subset of the great mind landscape. Same could be said about the sets of other earthly minds such as dog minds, or bug minds or male minds or in general the set of all animal minds.

Given our algorithm for sequentially generating minds, one can see that a mind could never be completely destroyed, making minds theoretically immortal. A particular mind may not be embodied at a given time, but the idea of it is always present. In fact it was present even before the material universe came into existence. So, given sufficient computational resources any mind design could be regenerated, an idea commonly associated with the concept of reincarnation [17].

Given our definition of mind we can classify minds with respect to their design, knowledgebase or embodiment. First, the designs could be classified with respect to their origins: copied from an existing mind like an upload, evolved via artificial or natural evolution or explicitly designed with a set of particular desirable properties. Another alternative is what is known as a Boltzmann Brain – a complete mind embedded in a system which arises due to statistically rare random fluctuations in the particles comprising the universe, but which is very likely due to vastness of the cosmos [18].

Lastly a possibility remains that some minds are physically or informationally recursively nested within other minds. With respect to the physical nesting we can consider a type of mind suggested by Kelly [19] who talks about “a very slow invisible mind over large physical distances”. It is possible that the physical universe as a whole or a significant part of it comprises such a mega-mind [20]. That theory has been around for millennia and has recently received some indirect experimental support [21]. In that case all the other minds we can consider are nested within a larger mind. With respect to the informational nesting a powerful mind can generate a less powerful mind as an idea. This obviously would take some precise thinking but should be possible for a sufficiently powerful artificially intelligent mind. Some scenarios describing informationally nested minds are analyzed by Yampolskiy in his work on artificial intelligence confinement problem [8]. Bostrom, using statistical reasoning, suggests that all observed minds, and the whole universe, are nested within the mind of a very powerful computer [22]. Similarly Lanza, using a completely different approach (biocentrism), argues that the universe is created by biological minds [23]. It remains to be seen if given a particular mind its origins can be deduced from some detailed analysis of the mind's design or actions [24].

While minds designed by human engineers comprise only a tiny region in the map of mind designs it is probably the best explored part of the map. Numerous surveys of artificial minds, created by AI researchers in the last 50 years, have been produced [25-29]. Such surveys typically attempt to analyze state-of-the-art in artificial cognitive systems and provide some internal classification of dozens of the reviewed systems with regards to their components and overall design. The main subcategories into which artificial minds designed by human engineers can be placed include brain (at the neuron level) emulators [27], biologically inspired cognitive architectures [28], physical symbol systems, emergent systems, dynamical and enactive systems [29]. Rehashing information about specific architectures presented in such surveys is beyond the scope of this paper, but one can notice incredible richness and diversity of designs even in that tiny area of the overall map we are trying to envision. For readers particularly interested in overview of superintelligent minds, animal minds and possible minds in addition to surveys mentioned above “Artificial General Intelligence and the Human Mental Model” by Yampolskiy and Fox is recommended [30].

For each mind subtype there are numerous architectures, which to a certain degree depend on the computational resources available via a particular embodiment. For example, theoretically a mind working with infinite computational resources could trivially brute-force any problem, always arriving at the optimal solution, regardless of its size. In practice, limitations of the physical world place constraints on available computational resources regardless of the embodiment type, making brute-force approach a non-feasible solution for most real world problems [10]. Minds working with limited computational resources have to rely on heuristic simplifications to arrive at “good enough” solutions [31-34].

Another subset of architectures consists of self-improving minds. Such minds are capable of examining their own design and finding improvements in their embodiment, algorithms or knowledge bases which will allow the mind to more efficiently perform desired operations [35]. We would anticipate many initial opportunities for optimization towards higher efficiency and fewer such options remaining after every generation. Depending on the definitions used, one can argue that a recursively self-improving mind actually changes itself into a different mind, rather than remaining itself, which is particularly obvious after a sequence of such improvements. Taken to the extreme, this idea implies that a simple act of learning new information transforms you into a different mind raising millennia old questions about the nature of personal identity.

With respect to their knowledge bases, minds could be separated into those without an initial knowledgebase, and which are expected to acquire their knowledge from the environment, minds which are given a large set of universal knowledge from the inception and those minds which are given specialized knowledge only in one or more domains. Whether the knowledge is stored in an efficient manner, compressed, classified or censored is dependent on the architecture and is a potential subject of improvement by self-modifying minds.

One can also classify minds in terms of their abilities or intelligence. Of course the problem of measuring intelligence is that no universal tests exist. Measures such as IQ tests and performance on specific tasks are not universally accepted and are always highly biased against non-human intelligences. Recently some work has been done on streamlining intelligence measurements across different types of machine intelligence [2, 36] and other “types” of intelligence [37], but the applicability of the results is still being debated. In general, the notion of intelligence only makes sense in the context of problems to which said intelligence can be applied. In fact this is exactly how IQ tests work, by presenting the subject with a number of problems and seeing how many the subject is able to solve in a given amount of time (computational resource). A subfield of computer science known as computational complexity theory is devoted to studying and classifying different problems with respect to their difficulty and with respect to computational resources necessary to solve them. For every class of problems complexity theory defines a class of machines capable of solving such problems. We can apply similar ideas to classifying minds, for example all minds capable of efficiently [13] solving problems in the class P or a more difficult class of NP-complete problems [38]. Similarly we can talk about minds with general intelligence belonging to the class of AI-Complete [39-41] minds, such as humans.

We can also look at the goals of different minds. It is possible to create a system which has no terminal goals and so such a mind is not very motivated to accomplish things. Many minds are designed or trained for obtaining a particular high level goal or a set of goals. We can envision a mind which has a randomly changing goal or a set of goals, as well as a mind which has many goals of different priority. Steve Omohundro used micro-economic theory to speculate about the driving forces in the behavior of superintelligent machines. He argues that intelligent machines will want to self-improve, be rational, preserve their utility functions, prevent counterfeit utility [42], acquire resources and use them efficiently, and protect themselves. He believes that machines' actions will be governed by rational economic behavior [43, 44]. Mark Waser suggested an additional "drive" to be included in the list of behaviors predicted to be exhibited by the machines [45]. Namely, he suggests that evolved desires for cooperation and being social are part of human ethics and are a great way of accomplishing goals, an idea also analyzed by Joshua Fox and Carl Shulman, but with contrary conclusions [46]. While it is commonly assumed that minds with high intelligence will converge on a common goal, Nick Bostrom via his orthogonality thesis has argued that a system can have any combination of intelligence and goals [47].

Regardless of design, embodiment or any other properties, all minds can be classified with respect to two fundamental but scientifically poorly defined properties – free will and consciousness. Both descriptors suffer from an ongoing debate regarding their actual existence or explanatory usefulness. This is primarily a result of impossibility to design a definitive test to measure or even detect said properties, despite numerous attempts [48-50] or to show that theories associated with them are somehow falsifiable. Intuitively we can speculate that consciousness, and maybe free will, are not binary properties but rather continuous and emergent abilities commensurate with a degree of general intelligence possessed by the system or some other property we shall term "mindness". Free will can be said to correlate with a degree to which behavior of the system can't be predicted [51]. This is particularly important in the design of artificially intelligent systems for which inability to predict their future behavior [52] is a highly undesirable property from the safety point of view [53, 54]. Consciousness on the other hand seems to have no important impact on the behavior of the system as can be seen from some thought experiments supposing the existence of "consciousless" intelligent agents [55]. This may change if we are successful in designing a test, perhaps based on observer impact on quantum systems [56], to detect and measure consciousness [57, 58].

In order to be social, two minds need to be able to communicate which might be difficult if the two minds don't share a common communication protocol, common culture or even common environment. In other words, if they have no common grounding, they don't understand each other. We can say that two minds understand each other if given the same set of inputs they produce similar outputs. For example, in sequence prediction tasks [59] two minds have an understanding if their predictions are the same regarding the future numbers of the sequence based on the same observed subsequence. We can say that a mind can understand another mind's function if it can predict the other's output with high accuracy.

A Survey of Taxonomies

Yudkowsky describes the map of mind design space as follows: "In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the space of minds-in-general. The entire map floats in a still vaster space, the space of optimization processes. Natural selection creates complex functional machinery without mindfulness; evolution lies inside the space of optimization processes but outside the circle of minds" [60]. Figure 1 illustrates one possible mapping inspired by this description.

Similarly, Ivan Havel writes "…all conceivable cases of intelligence (of people, machines, whatever) are represented by points in a certain abstract multi-dimensional "super space" that I will call the intelligence space (shortly IS). Imagine that a specific coordinate axis in IS is assigned to any conceivable particular ability, whether human, machine, shared, or unknown (all axes having one common origin). If the ability is measurable the assigned axis is endowed with a corresponding scale. Hypothetically, we can also assign scalar axes to abilities, for which only relations like "weaker-stronger", "better-worse", "less-more" etc. are meaningful; finally, abilities that may be only present or absent may be assigned with "axes" of two (logical) values (yes-no). Let us assume that all coordinate axes are oriented in such a way that greater distance from the common origin always corresponds to larger extent, higher grade, or at least to the presence of the corresponding ability. The idea is that for each individual intelligence (i.e. the intelligence of a particular person, machine, network, etc.), as well as for each generic intelligence (of some group) there exists just one representing point in IS, whose coordinates determine the extent of involvement of particular abilities [62]." If the universe (or multiverse) is infinite, as our current physics theories indicate, then all possible minds in all possible states are instantiated somewhere [5].

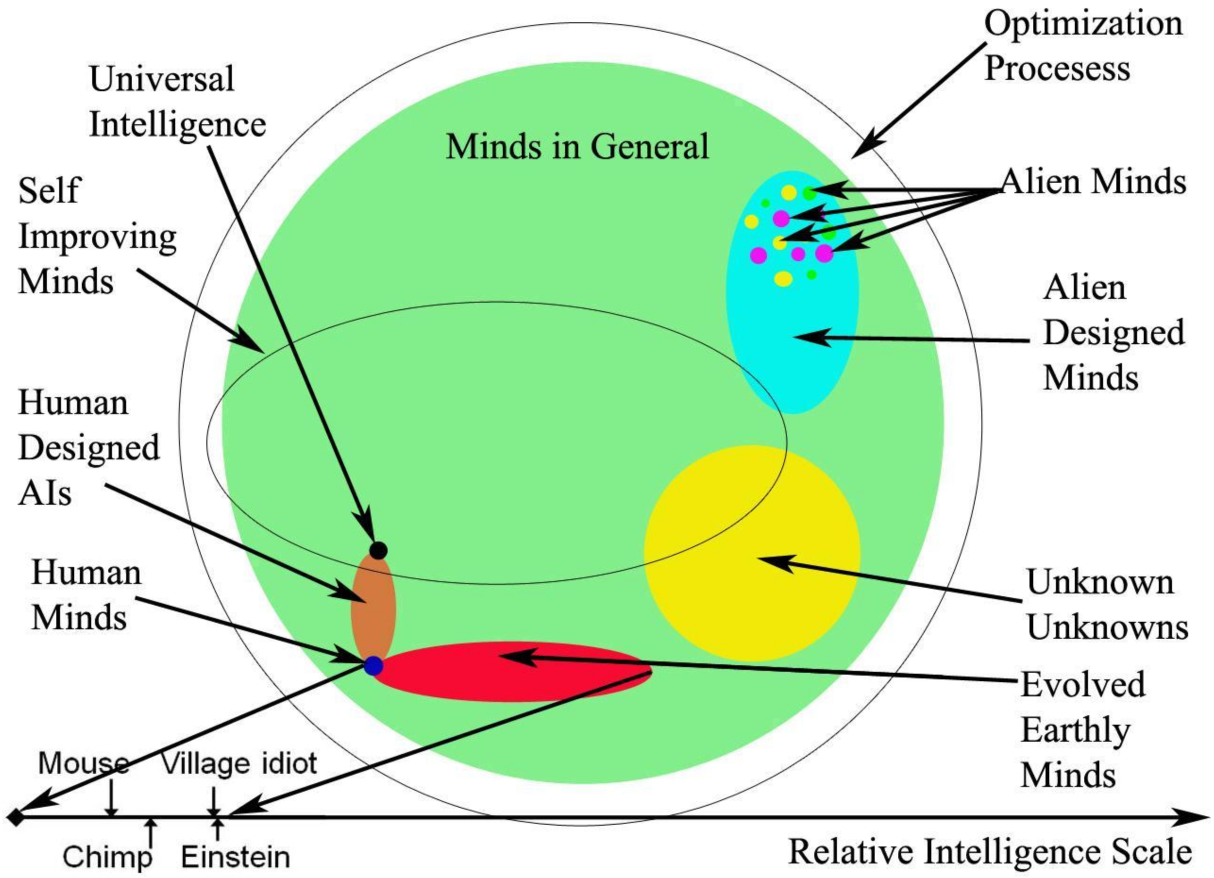

Figure 1: The universe of possible minds [60, 61].

Ben Goertzel proposes the following classification of Kinds of Minds, mostly centered around the concept of embodiment [63]:

- **Singly Embodied –** control a single physical or simulated system.
- **Multiply Embodied -** control a number of disconnected physical or simulated systems.
- **Flexibly Embodied –** control a changing number of physical or simulated systems.
- **Non-Embodied –** resides in a physical substrate but doesn't utilize the body in a traditional way.
- **Body-Centered –** consists of patterns emergent between physical system and the environment.
- **Mindplex –** a set of collaborating units each of which is itself a mind [64].
- **Quantum –** an embodiment based on properties of quantum physics.

- **Classical -** an embodiment based on properties of classical physics.

J. Storrs Hall in his “Kinds of Minds” suggests that different stages a developing AI may belong to can be classified relative to its humanlike abilities. His classification encompasses:

- **Hypohuman** - infrahuman, less-than-human capacity.
- **Diahuman** - human-level capacities in some areas, but still not a general intelligence.
- **Parahuman** - similar but not identical to humans, as for example, augmented humans.
- **Allohuman** - as capable as humans, but in different areas.
- **Epihuman** - slightly beyond the human level.
- **Hyperhuman** - much more powerful than human, superintelligent [30, 65].

Patrick Roberts in his book *Mind Making* presents his ideas for a “Taxonomy of Minds”, we will leave it to the reader to judge usefulness of his classification [66]:

- **Choose Means** - Does it have redundant means to the same ends? How well does it move between them?
- **Mutate** - Can a mind naturally gain and lose new ideas in its lifetime?
- **Doubt** - Is it eventually free to lose some or all beliefs? Or is it wired to obey the implications of every sensation?
- **Sense Itself** - Does a mind have the senses to see the physical conditions of that mind?
- **Preserve Itself** - Does a mind also have the means to preserve or reproduce itself?
- **Sense Minds** - Does a mind understand mind, at least of lower classes, and how well does it apply that to itself, to others?
- **Sense Kin** - Can it recognize the redundant minds, or at least the bodies of minds, that it was designed to cooperate with?
- **Learn** - Does the mind's behavior change from experience? Does it learn associations?
- **Feel** - We imagine that an equally intelligent machine would lack our conscious experience.
- **Communicate** - Can it share beliefs with other minds?

Kevin Kelly has also proposed a “Taxonomy of Minds” which in his implementation is really just a list of different minds, some of which have not showed up in other taxonomies [19]:

- “Super fast human mind.
- Mind with operational access to its source code.
- Any mind capable of general intelligence and self-awareness.
- General intelligence without self-awareness.
- Self-awareness without general intelligence.
- Super logic machine without emotion.
- Mind capable of imagining a greater mind.
- Mind capable of creating a greater mind. (M2)
- Self-aware mind incapable of creating a greater mind.
- Mind capable of creating greater mind which creates greater mind. etc. (M3, and Mn)
- Mind requiring a protector while it develops.
- Very slow "invisible" mind covering a large physical distance.
- Mind capable of cloning itself and remaining in unity with clones.
- Mind capable of immortality.
- Rapid dynamic mind able to change its mind-space-type sectors (think different)
- Global mind -- large supercritical mind of subcritical brains.
- Hive mind -- large super critical mind made of smaller minds each of which is supercritical.
- Low count hive mind with few critical minds making it up.
- Borg -- supercritical mind of smaller minds supercritical but not self-aware
- Nano mind -- smallest (size and energy profile) possible super critical mind.
- Storebit -- Mind based primarily on vast storage and memory.
- Anticipators -- Minds specializing in scenario and prediction making.
- Guardian angels -- Minds trained and dedicated to enhancing your mind, useless to anyone else.
- Mind with communication access to all known "facts." (F1)
- Mind which retains all known "facts," never erasing. (F2)
- Symbiont, half machine half animal mind.
- Cyborg, half human half machine mind.
- Q-mind, using quantum computing
- Vast mind employing faster-than-light communications”

Elsewhere Kelly provides a lot of relevant analysis of landscape of minds writing about Inevitable Minds [67], The Landscape of Possible Intelligences [68], What comes After Minds? [69], and the Evolutionary Mind of God [70].

Aaron Sloman in “The Structure of the Space of Possible Minds”, using his virtual machine model, proposes a division of the space of possible minds with respect to the following properties [1]:

- Quantitative VS Structural
- Continuous VS Discrete
- Complexity of stored instructions
- Serial VS Parallel
- Distributed VS Fundamentally Parallel
- Connected to External Environment VS Not Connected
- Moving VS Stationary
- Capable of modeling others VS Not capable
- Capable of logical inference VS Not Capable
- Fixed VS Re-programmable
- Goal consistency VS Goal Selection
- Meta-Motives VS Motives
- Able to delay goals VS Immediate goal following
- Statics Plan VS Dynamic Plan
- Self-aware VS Not Self-Aware

Taxonomy of Superintelligences

In the light of recent exponential growth in capabilities of AI models it is reasonable to attempt to suggest a taxonomy of future superintelligences. The creation of such a taxonomy would involve a blend of computational theory, philosophy of mind, and ethics. Let's attempt a speculative taxonomy while outlining capabilities at each level:

SAI Level 1: Baseline Superintelligence
Capabilities: This level surpasses human intelligence in all domains. Capabilities might include solving currently unsolvable mathematical conjectures within seconds, creating Nobel-prize winning literature in minutes, and making scientific breakthroughs that would take humans decades, all within a short period.
Examples: Imagine an AI that could design a cure for all known forms of cancer based on a fundamental understanding of cellular biology and then generate the optimal economic model for distributing it worldwide, while also drafting international legislation to enable its implementation.

SAI Level 2: Super-Superintelligence
Capabilities: This intelligence would be as superior to SAI Level 1 as Level 1 is to humans. For example, if Level 1 can cure all known cancers, Level 2 might be capable of reengineering biological life to be inherently immune to diseases and long-lived.
Examples: An SAI Level 2 might develop a Theory of Everything in physics that unifies quantum mechanics and general relativity, not just on paper but also in

practical applications. It might also create self-replicating, self-repairing technologies that can clean and renew Earth's ecosystems on a global scale.

SAI Level 3: SSSuperintelligence
Capabilities: Exponentially more capable than Level 2, this level could involve manipulating the fabric of reality at the sub-atomic or even Planck scale.
Examples: Imagine an AI that could harness zero-point energy, essentially making energy constraints irrelevant. It could possibly even manipulate the fundamental constants of the universe locally, changing the rules of physics to solve previously "impossible" problems.

SAI Level 4: SSSSuperintelligence
Capabilities: We're reaching levels where it becomes increasingly abstract to even predict what such an intelligence could do, as it would be capable of comprehending and manipulating dimensions or aspects of reality that are entirely outside human understanding.
Examples: A Level 4 SAI could potentially simulate multiple universes to perform experiments and derive knowledge, manipulate time, or even create new forms of life and intelligence that are as superior to it as it is to us.

SAI Level n: S^{n}uperintelligence
Capabilities: Each new level continues to be exponentially more capable than the previous, reaching competencies that are virtually incomprehensible from our current standpoint.
Examples: At this point, the examples would be beyond human comprehension, venturing into realms of capability that may involve the manipulation of fundamental aspects of existence that humans are not even aware of.

Qualitative Attributes (Common Across Levels)

- **Computational Efficiency:** Increases exponentially with each level.
- **Omnidisciplinarity:** Mastery of all possible domains, including those that higher-level SAIs invent.
- **Strategic Depth:** Enhanced abilities for planning and long-term strategy, which could span across time scales and dimensions incomprehensible to lower orders.
- **Ethical or Value Alignment:** With each level, the challenge of aligning the SAI's objectives with human or universal good becomes exponentially more complex and critical.

Cloning and Equivalence Testing Across Substrates

The possibility of uploads rests on the ideas of computationalism [71] specifically, substrate independence and equivalence meaning that the same mind can be instantiated in different substrates and move freely between them. If your mind is cloned and if a copy is instantiated in a different substrate from the original one (or on the same substrate), how can it be verified that the copy is indeed an identical mind? At least immediately after cloning and before it learns any new information. For that purpose, I propose a variant of a Turing Test, which also relies on interactive text-only communication to ascertain the quality of the copied mind. The text-only interface is important not to prejudice the examiner against any unusual substrates on which the copied mind might be running. The test proceeds by having the examiner (original mind) ask questions of the copy (cloned mind), questions which supposedly only the original mind would know answers to (testing should be done in a way which preserves privacy). Good questions would relate to personal preferences, secrets (passwords, etc.) as well as recent dreams. Such a test could also indirectly test for consciousness via similarity of subjective qualia. Only a perfect copy should be able to answer all such questions in the same way as the original mind. Another variant of the same test may have a 3rd party test the original and cloned mind by seeing if they always provide the same answer to any question. One needs to be careful in such questioning not to give undue weight to questions related to the mind's substrate as that may lead to different answers. For example, asking a human if he is hungry may produce an answer different from the one which would be given by a non-biological robot.

Conclusions

Science periodically experiences a discovery of a whole new area of investigation. For example, observations made by Galileo Galilei led to the birth of observational astronomy [72], aka study of our universe; Watson and Crick's discovery of the structure of DNA led to the birth of the field of genetics [73], which studies the universe of blueprints for organisms; Stephen Wolfram's work with cellular automata has resulted in "a new kind of science" [74] which investigates the universe of computational processes. I believe that we are about to discover yet another universe – the universe of minds [75].

As our understanding of the human brain improves, thanks to numerous projects aimed at simulating or reverse engineering a human brain, we will no doubt realize that human intelligence is just a single point in the vast universe of potential intelligent agents comprising a new area of study. The new field, which I would like to term *intellectology*, will study and classify design space of intelligent agents, work on establishing limits to intelligence (minimum sufficient for general intelligence and maximum subject to physical limits), contribute to consistent measurement of intelligence across intelligent agents, look at recursive self-improving systems, design new intelligences (making AI a sub-field of intellectology) and evaluate capacity for understanding higher level intelligences by lower level ones. At the more theoretical level the field will look at the distribution of minds on the number line and probabilistic distribution of minds in the mind design space as well as attractors in the mind design space. It will consider how evolution, drives, and design choices impact the density of minds in the space of possibilities. It will investigate intelligence as an additional computational resource along with time and memory. The field will not be subject to the current limitations brought on by the human centric view of intelligence [76] and will open our understanding to seeing intelligence as a fundamental resource like space or time. Finally, I believe intellectology will highlight the inhumanity of most possible minds and the dangers associated with such minds being placed in charge of humanity [77, 78].

References

1. Sloman, A., *The Structure and Space of Possible Minds*. The Mind and the Machine: philosophical aspects of Artificial Intelligence. 1984: Ellis Horwood LTD.
2. Legg, S. and M. Hutter, *Universal Intelligence: A Definition of Machine Intelligence.* Minds and Machines, December 2007. **17(4)**: p. 391-444.
3. Yampolskiy, R., *How to Escape From the Simulation.*
4. Hanson, R., *If Uploads Come First.* Extropy, 1994. **6(2)**.
5. Bostrom, N., *Quantity of experience: brain-duplication and degrees of consciousness.* Minds and Machines, 2006. **16(2)**: p. 185-200.
6. Yampolskiy, R. and M. Gavrilova, *Artimetrics: Biometrics for Artificial Entities.* IEEE Robotics and Automation Magazine (RAM), 2012. **19**(4): p. 48-58.
7. Yampolskiy, R.V., B. Klare, and A.K. Jain. *Face recognition in the virtual world: Recognizing Avatar faces*. in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. 2012. IEEE.
8. Yampolskiy, R.V., *Leakproofing Singularity - Artificial Intelligence Confinement Problem.* Journal of Consciousness Studies (JCS), 2012. **19(1-2)**: p. 194–214.
9. Wikipedia, *Universal Turing Machine*. Retrieved April 14, 2011: Available at: http://en.wikipedia.org/wiki/Universal_Turing_machine.
10. Lloyd, S., *Ultimate Physical Limits to Computation.* Nature, 2000. **406**: p. 1047-1054.

11. Levin, L., *Universal Search Problems.* Problems of Information Transmission, 1973. **9(3)**: p. 265--266.
12. Rice, H.G., *Classes of recursively enumerable sets and their decision problems.* Transactions of the American Mathematical Society, 1953. **74**(2): p. 358-366.
13. Yampolskiy, R.V., *Efficiency Theory: a Unifying Theory for Information, Computation and Intelligence.* Journal of Discrete Mathematical Sciences & Cryptography, 2013. **16(4-5)**: p. 259-277.
14. Kolmogorov, A.N., *Three Approaches to the Quantitative Definition of Information.* Problems Inform. Transmission, 1965. **1(1)**: p. 1-7.
15. Sotala, K. and H. Valpola, *Coalescing Minds: Brain Uploading-Related Group Mind Scenarios.* International Journal of Machine Consciousness, 2012. **4(1)**: p. 293-312.
16. Yudkowsky, E.S., *Creating Friendly AI - The Analysis and Design of Benevolent Goal Architectures*. 2001: Available at: http://singinst.org/upload/CFAI.html.
17. Fredkin, E., *On the soul*. 1982, Draft.
18. De Simone, A., et al., *Boltzmann brains and the scale-factor cutoff measure of the multiverse.* Physical Review D, 2010. **82**(6): p. 063520.
19. Kelly, K., *A Taxonomy of Minds*. 2007: Available at: http://kk.org/thetechnium/archives/2007/02/a_taxonomy_of_m.php.
20. Vanchurin, V., *The world as a neural network.* Entropy, 2020. **22**(11): p. 1210.
21. Krioukov, D., et al., *Network Cosmology.* Sci. Rep., 2012. **2**.
22. Bostrom, N., *Are You Living In a Computer Simulation?* Philosophical Quarterly, 2003. **53(211)**: p. 243-255.
23. Lanza, R., *A new theory of the universe.* American Scholar, 2007. **76**(2): p. 18.
24. Yampolskiy, R.V., *On the origin of synthetic life: attribution of output to a particular algorithm.* Physica Scripta, 2016. **92**(1): p. 013002.
25. Miller, M.S.P. *Patterns for Cognitive Systems*. in *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on*. 2012.
26. Cattell, R. and A. Parker, *Challenges for Brain Emulation: Why is it so Difficult?* Natural Intelligence, 2012. **1(3)**: p. 17-31.
27. de Garis, H., et al., *A world survey of artificial brain projects, Part I: Large-scale brain simulations.* Neurocomputing, 2010. **74**(1–3): p. 3-29.
28. Goertzel, B., et al., *A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures.* Neurocomput., 2010. **74**(1-3): p. 30-49.
29. Vernon, D., G. Metta, and G. Sandini, *A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental*

Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation, 2007. **11**(2): p. 151-180.

30. Yampolskiy, R.V. and J. Fox, *Artificial General Intelligence and the Human Mental Model*, in *Singularity Hypotheses*. 2012, Springer Berlin Heidelberg. p. 129-145.
31. Yampolskiy, R.V., L. Ashby, and L. Hassan, *Wisdom of Artificial Crowds—A Metaheuristic Algorithm for Optimization.* Journal of Intelligent Learning Systems and Applications, 2012. **4**(2): p. 98-107.
32. Ashby, L.H. and R.V. Yampolskiy. *Genetic algorithm and Wisdom of Artificial Crowds algorithm applied to Light up*. in *Computer Games (CGAMES), 2011 16th International Conference on*. 2011. IEEE.
33. Hughes, R. and R.V. Yampolskiy, *Solving Sudoku Puzzles with Wisdom of Artificial Crowds.* International Journal of Intelligent Games & Simulation, 2013. **7**(1): p. 6.
34. Port, A.C. and R.V. Yampolskiy. *Using a GA and Wisdom of Artificial Crowds to solve solitaire battleship puzzles*. in *Computer Games (CGAMES), 2012 17th International Conference on*. 2012. IEEE.
35. Hall, J.S., *Self-Improving AI: An Analysis.* Minds and Machines, October 2007. **17(3)**: p. 249 - 259.
36. Yonck, R., *Toward a Standard Metric of Machine Intelligence.* World Future Review, 2012. **4**(2): p. 61-70.
37. Herzing, D.L., *Profiling nonhuman intelligence: An exercise in developing unbiased tools for describing other "types" of intelligence on earth.* Acta Astronautica, 2014. **94**(2): p. 676-680.
38. Yampolskiy, R.V., *Construction of an NP Problem with an Exponential Lower Bound.* Arxiv preprint arXiv:1111.0305, 2011.
39. Yampolskiy, R.V., *Turing Test as a Defining Feature of AI-Completeness*, in *Artificial Intelligence, Evolutionary Computation and Metaheuristics - In the footsteps of Alan Turing. Xin-She Yang (Ed.)*. 2013, Springer. p. 3-17.
40. Yampolskiy, R.V., *AI-Complete, AI-Hard, or AI-Easy–Classification of Problems in AI.* The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, 2012.
41. Yampolskiy, R.V., *AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System.* ISRN Artificial Intelligence, 2011. **271878**.
42. Yampolskiy, R.V., *Utility Function Security in Artificially Intelligent Agents.* Journal of Experimental and Theoretical Artificial Intelligence (JETAI), 2014: p. 1-17.
43. Omohundro, S.M., *The Nature of Self-Improving Artificial Intelligence*, in *Singularity Summit*. 2007: San Francisco, CA.

44. Omohundro, S.M., *The Basic AI Drives*, in *Proceedings of the First AGI Conference, Volume 171, Frontiers in Artificial Intelligence and Applications, P. Wang, B. Goertzel, and S. Franklin (eds.)*. February 2008, IOS Press.
45. Waser, M.R., *Designing a Safe Motivational System for Intelligent Machines*, in *The Third Conference on Artificial General Intelligence*. March 5-8, 2010: Lugano, Switzerland.
46. Fox, J. and C. Shulman, *Superintelligence Does Not Imply Benevolence*, in *8th European Conference on Computing and Philosophy*. October 4-6, 2010 Munich, Germany.
47. Bostrom, N., *The superintelligent will: Motivation and instrumental rationality in advanced artificial agents.* Minds and Machines, 2012. **22**(2): p. 71-85.
48. Hales, C., *An empirical framework for objective testing for P-consciousness in an artificial agent.* Open Artificial Intelligence Journal, 2009. **3**: p. 1-15.
49. Aleksander, I. and B. Dunmall, *Axioms and Tests for the Presence of Minimal Consciousness in Agents I: Preamble.* Journal of Consciousness Studies, 2003. **10**(4-5): p. 4-5.
50. Arrabales, R., A. Ledezma, and A. Sanchis, *ConsScale: a plausible test for machine consciousness?* 2008.
51. Aaronson, S., *The Ghost in the Quantum Turing Machine.* arXiv preprint arXiv:1306.0159, 2013.
52. Yampolskiy, R.V., *Behavioral modeling: an overview.* American Journal of Applied Sciences, 2008. **5**(5): p. 496-503.
53. Yampolskiy, R.V., *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*, in *Philosophy and Theory of Artificial Intelligence*. 2013, Springer Berlin Heidelberg. p. 389-396.
54. Yampolskiy, R.V., *What to Do with the Singularity Paradox?*, in *Philosophy and Theory of Artificial Intelligence*. 2013, Springer Berlin Heidelberg. p. 397-413.
55. Chalmers, D.J., *The conscious mind: In search of a fundamental theory*. 1996: Oxford University Press.
56. Gao, S., *A quantum method to test the existence of consciousness.* The Noetic Journal, 2002. **3**(3): p. 27-31.
57. Elamrani, A. and R.V. Yampolskiy, *Reviewing tests for machine consciousness.* Journal of Consciousness Studies, 2019. **26**(5-6): p. 35-64.
58. Yampolskiy, R.V., *Artificial consciousness: An illusionary solution to the hard problem.* Reti, saperi, linguaggi, 2018(2): p. 287-318.
59. Legg, S. *Is there an elegant universal theory of prediction?* in *Algorithmic Learning Theory*. 2006. Springer.
60. Yudkowsky, E., *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in *Global Catastrophic Risks*, N. Bostrom and M.M. Cirkovic, Editors. 2008, Oxford University Press: Oxford, UK. p. 308-345.

61. Yudkowsky, E., *The Human Importance of the Intelligence Explosion*, in *Singularity Summit at Stanford*. 2006.
62. Havel, I.M., *On the Way to Intelligence Singularity*, in *Beyond Artificial Intelligence*, J. Kelemen, J. Romportl, and E. Zackova, Editors. 2013, Springer Berlin Heidelberg. p. 3-26.
63. Geortzel, B., *The Hidden Pattern: A Patternist Philosophy of Mind. Chapter 2 - Kinds of Minds* 2006: Brown Walker Press.
64. Goertzel, B., *Mindplexes: The Potential Emergence of Multiple Levels of Focused Consciousness in Communities of AI's and Humans* Dynamical Psychology, 2003. http://www.goertzel.org/dynapsyc/2003/mindplex.htm.
65. Hall, J.S., *Chapter 15: Kinds of Minds*, in *Beyond AI: Creating the Conscience of the Machine*. 2007, Prometheus Books: Amherst, NY.
66. Roberts, P., *Mind Making: The Shared Laws of Natural and Artificial*. 2009: CreateSpace.
67. Kelly, K., *Inevitable Minds*. 2009: Available at: http://kk.org/thetechnium/archives/2009/04/inevitable_mind.php.
68. Kelly, K., *The Landscape of Possible Intelligences*. 2008: Available at: http://kk.org/thetechnium/archives/2008/09/the_landscape_o.php.
69. Kelly, K., *What Comes After Minds?* 2008: Available at: http://kk.org/thetechnium/archives/2008/12/what_comes_afte.php.
70. Kelly, K., *The Evolutionary Mind of God* 2007: Available at: http://kk.org/thetechnium/archives/2007/02/the_evolutionar.php.
71. Putnam, H., *Brains and behavior.* Readings in philosophy of psychology, 1980. **1**: p. 24-36.
72. Galilei, G., *Dialogue concerning the two chief world systems: Ptolemaic and Copernican*. 1953: University of California Pr.
73. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids.* Nature, 1953. **171**(4356): p. 737-738.
74. Wolfram, S., *A New Kind of Science*. May 14, 2002: Wolfram Media, Inc.
75. Sanderson, K., *GPT-4 is here: what scientists think.* Nature, 2023. **615**(7954): p. 773.
76. Yampolskiy, R. *On the Differences between Human and Machine Intelligence*. in *AISafety@ IJCAI*. 2021.
77. Yampolskiy, R. *On controllability of artificial intelligence*. in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*. 2020.
78. Yampolskiy, R.V., *On the Controllability of Artificial Intelligence: An Analysis of Limitations.* Journal of Cyber Security and Mobility, 2022: p. 321–404-321–404.

Gardener Comments

Mark:
Overall this is an interesting perspective that relates closely to work on Machine Behavior. I feel the article could focus more on its call to action as opposed to on spelling out the specifics of the interpretation and comparative analysis of minds, i.e. the high level concept seems more important and robust than the specific interpretation of minds and their representation.

Ted Wade:
One key assumption is that a valid science of intellectology can be accomplished by intellects as limited as ours. There might be something like Vingean uncertainty that severely limits any such attempt. The paper assumes that mind is an instantiated intelligence, and refers to Legg for a working definition of intelligence. The paper should at least briefly explain Legg, and perhaps offer a couple more takes on defining its primary subject matter. That way we would be more able to separate minds from other computational systems.

Dr. Jason Jeffrey Jones (Psychology PhD):
This article is too disorganized and unfocused to publish in its current state. However, the topic is important, and some of the ideas in this manuscript are inspiring. I would advise the author to drop the long development of the simple (one might say facile) idea that "every mind is exactly one particular integer." Instead, begin with the idea - introduced late in the current manuscript - that every mind exists in a large space where every ability is a dimension. Developing that idea further would be interesting.

Josh Randall:
The article attempts to describe a field of intellectology which is primarily comprised of previous attempts at developing a taxonomy of minds - primarily derived from researchers into AI. The author spends a large chunk of the article focused on the infinitude of minds and distinguishing between artificial minds and biologically generated minds. Much of this prose relies on unsourced or under-explained ideas about infinity, definitions of minds, and concepts surrounding panpsychism. The primary new contribution appears to be the final paragraph explaining additional information about uploads but would have benefited from much more detail as opposed to the literature review throughout the rest of the article.

Anonymous1:
The core of the article is unspecific: "Consequently, we can treat the space of all minds as the space of programs with the specific property of exhibiting intelligence if properly embodied. All programs could be represented as strings of

binary numbers, implying that each mind can be represented by a unique number"

The core of being a mind is the existence of a (Cartesian) conscious experience, and that depends on physical implementation. It is possible that the same Turing machine can lead to a conscious experience or not depending on its physical implementation (it is even possible that the same physical implementation leads to conscience depending on the speed of execution, see the paradox of the Searle "Chinese room"). Minds are not mathematical objects, and they are not "enumerable".

The paper does not engage with axiomatic theories of conscience and it does not even consider that intelligence could happen without conscience experience (https://ceur-ws.org/Vol-2287/short7.pdf). A deeper engagement with literature on philosophy of mind (Koch, Tononi, etc), and "neural correlates of conscience" is necessary before the paper can be the real seed of any scientific work. Currently, I find this work as both too speculative, and lacking a clear unified message.

Dr. Payal B. Joshi:
The article presents an intriguing concept on intellectology and mind as a universe. Modern times are gripped with the concepts of machine learning, data mining and artificial intelligence. It comes across that such studies, if incorporated in understanding nuances of psychological behaviors of human beings, shall unravel many facets of the human brain (maybe!). Also, the author has presented a detailed work that depicts a complex interplay of intelligence, goals and human behavior in mind in multidimensional space. These are complex but may allow us to understand nuances of superhuman intelligence and psychotic behaviors too. Albeit, such studies are largely theoretical in context, these studies have the potential to further studies in understanding intellectual minds and human behavior.

As a proponent of artificial intelligence, I found this article particularly enjoyable and gave me food for thought for running my next experiment too. I recommend publishing the article as it is for wider readership.

Roger's Bacon:
I recommend this paper for publication, while also acknowledging that it could be greatly improved with some revisions.

1) I'm not sure how necessary the “size, complexity and properties of mind” section is, feels like it gets lost in the weeds a little bit. That section and the

"Infinitude of Minds" make simple but important points that I think can be expressed more succinctly. I appreciate the challenge of what the author is trying to do here as there are numerous philosophical issues and distinctions that could be raised, but given how speculative all of this is it could make more sense to pass through this relatively quickly by raising several open questions.

2) the paragraph starting with..“Given our algorithm for sequentially generating minds, one can see that a mind could never be completely destroyed…” could really just be a sentence, not sure how much it is adding.

3) What’s most interesting to me is the final sections (space of design/taxonomies) and I’d be curious to see this expanded, especially in light of recent advances in AI. Though I imagine some will balk at the proposal of intellectology, I think it's an interesting speculative idea and considering it further may be fruitful in some way, even if very indirectly. How can we get from the current state of the mind sciences to a mature field of intellectology? What new methods/frameworks will be needed? Might be useful to regroup some existing branches of research under this new umbrella term?

4) Issues surrounding collective intelligences (ant colonies, governments, all of humanity) might be worth addressing briefly

Some further reflections:

1) I'd be curious to see a catalog of new scientific fields which have been proposed in either science fact or science fiction. From science-fiction, psychohistory (*Foundation*) and cosmic sociology (Three Body Problem) come to mind, but no doubt there are countless others. From science fact, the only other one I’ve come across is Entitiology from “An ontology of psychedelic entity experiences in evolutionary psychology and neurophenomenology” (Winkleman, 2018).

> I propose that to determine whether there are consistent and unique features of psychedelic entity experiences, we need a cross-cultural and interdisciplinary assessment of phenomenological reports of diverse types of experiences of entities (i.e., see Winkelman, 1992). Formal quantitative comparisons of the reported characteristics of diverse entity experiences are necessary to discover any commonalities to psychedelic entity experiences and their uniqueness with respect to other types of entity experiences. We need a new field of scientific inquiry, entitiology, i.e., the study of entities, to address the questions of the nature of psychedelic, and other types of entity experiences….Consequently, entitiology must

> encompass a number of existing areas of inquiry and by necessity will incorporate at least a part of the domain of the entities reported in the following areas of study: Angelology, Demonology, Spiritology, UFOology, Folklore and Mythology studies of elves, fairies, dwarfs, pixies, imps, gnomes, goblins, leprechauns, little people, and similar phenomena reported in cultures around the world, Possession, Mediumship, and Shamanism, Ghosts, apparitions, and poltergeist phenomena, Psychiatric syndromes, especially abnormal body syndromes and experiences such as the “Old Hag” and other terrorizing dreams. A systemic coding and analysis of the features of these various accounts can determine whether or not a single type or several types of psychedelic entity experiences occur. And only through comparison with profiles obtained for reports of what are conceptualized as angels, fairies, extraterrestrials, and shamanic spirits can we determine if there are unique features of psychedelic entities.

2) I’m reminded of Michael Levin’s idea of classifying selves by the size/shape of their “cognitive (computational) light-cone” from “The Computational Boundary of a “Self: Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition” (2019). The “Predictions and Research Program” section of the paper has numerous reflections that speak to the nascent field of intellectology.

> “I propose a semi-quantitative metric, based on the spatio-temporal boundaries of events that systems measure and try to control, that can be used to define and compare the cognitive boundaries for highly diverse types of agents (which could be biological, exo-biological, or artificial)...The edges of a given Agent’s goal space define a sort of “computational light cone” – the boundaries beyond which its cognitive system cannot operate. For example, a tick has a relatively small cognitive boundary, having very little memory or predictive power in the temporal direction, and sensing/acting very locally. A dog has much more temporal memory, some forward prediction ability, and a degree of spatial concern. However, it is likely impossible for a dog’s cognitive apparatus to operate with notions about what is going to happen next month or in the adjacent town. Human minds can operate over goals of vastly greater spatial and temporal scales, and one can readily imagine artificial (organic or software-based) Selves with properties that define every possible shape in this space (and perhaps change their boundaries over evolutionary and individual timescales)”

Joe R:
This article offers plenty of thought-provoking ideas, consistently fails to give good reasons for most of them, and repeatedly shoots itself in the foot while

trying. I must reluctantly admit that its treatment of mind-space is fairly comprehensive, but the lack of concrete examples or even half-hearted attempts at internal consistency significantly diminishes its value. If the author manages to fix the numerous holes in their logic and factual errors, I could recommend publishing the article for its thoroughness alone, but as it stands, I would expect better internal coherence of a grand-sounding attempt to introduce the field of "intellectology" to the world.

For example: The second proof of infinite possible minds raises many questions. First, why is "time" a relevant factor at all? Could you not have a binary representation of a mind which, if instantiated by a suitable machine, would fulfill all relevant properties of a mind? The result might be static, frozen, unchanging, but it is still the case that it represents a mind, albeit a mind currently in stasis. This is because if you were to run the binary on a substrate, it would be capable of "interacting with its environment and other minds." Similar logic applies to classifying the code for Mario Bros as a "game" even if no one is currently playing it. Second, why does the insistence that how fast the mind is running matters? If you run Mario Bros. at 2x speed, nothing about the underlying code has changed; it remains an identical copy of Mario Bros. Speed is a feature of the substrate, not the mind. Finally, if in defiance of the above we require that a binary representation of a mind must include change over time, i.e. if a mind must be actively running on something to be classified as such; then the representation is causally incomplete without also including a full representation of the environment the mind is operating in, since (by the article's own definition of a mind) that environment must necessarily interact with the mind. By this logic, Mario Bros. does not count as a "game" unless it is currently being played by someone, and the binary representation of that game must include the player's input and everything causally and temporally adjacent to it - unless I'm missing something.

Put another way - a simulation of "the brain of Bob Smith, accountant, at exactly noon Eastern Daylight Time on May 12, 2020" should be counted as a mind, even if it remains "paused' forever and no other information is included in the encoding of said mind…or so I would think? The article gives precisely zero concrete examples so I can't be sure of their definition. If real or simulated time has to pass for Bob to be considered a mind…how much time? Does a simulated minute of Bob's life count? At what point would the article writers consider Bob a "different mind"? Why on earth do they think "drunk Bob" qualifies as "different" but "Bob at time T+1" doesn't?

Taken seriously, the arguments in the article would seem to imply the following contradictory points:

1) A binary encoding must be instantiated on a substrate to be considered a mind (i.e. subjective time must pass for said mind);
2) Instantaneous time-slices of a mind do not count as minds themselves;
3) Two (non?)-instantaneous time-slices of the same entity (e.g. "drunk Bob" and "sober Bob") which could have been arrived at in the same simulation not only count as minds, but as two different minds;
4) A single integer can encode a mind;
5) You can instantiate a mind from an integer by running it on something;
6) Minds cannot be fully destroyed because they continue to exist as integers;
7) A Universal Turing Machine is a mind;
8) "the same mind can be instantiated in different substrates and move freely between them".

(1) and (4) seem especially in conflict, unless we assume that the relevant integer simultaneously encodes a thinking entity, its substrate, and some unspecified yet adequate amount of time passing for the entity…which would render (5) utterly meaningless and cause serious problems for (8).

(2) and (3) are at least partially in conflict. If you run a mind for 2 minutes and split the resulting computation into two 1-minute chunks, is that two half-instances of one mind or 2 different unique minds? What if one minute is sober and the other is drunk? What if you keep splitting until you get a bunch of infinitesimally tiny time-slices?

(1) and (6) seem to be in conflict. If part of the definition of a mind is that it be instantiated, then you can destroy one by destroying its substrate.

I lack the mathematical foundation to dispute (7) but it intuitively seems like it must conflict with at least one of the preceding claims.

I invite the author to clarify whether their proposed definition of a mind does, or does not, include the substrate it's running on, and to adjust their claims appropriately.

"Consequently, since a human mind has only a finite number of possible states, there are minds which can never be fully understood by a human mind as such mind designs have a much greater number of states, making their understanding impossible as can be demonstrated by the pigeonhole principle." -> the conclusion might be true, but I don't find this run-on sentence convincing. Human understanding is not limited by the number of states our minds can occupy, it's limited by some hideously complex function thereof. Smaller minds can understand larger minds via compression, abstraction, pattern-recognition, etc. A smaller mind can even, in theory, simulate a larger one in its entirety, given enough time. There may still exist minds incomprehensible to humans, but not solely because they're bigger.

"Also, the most powerful and most knowledgeable mind has always been associated with the idea of Deity or the Universal Mind." -> Now they're just making stuff up to sound impressive. Literally one page ago the author proved there is no largest mind! What does this add? Even being charitable, "always" is false and entirely unjustified.

"Depending on the definitions used, one can argue that a recursively self-improving mind actually changes itself into a different mind, rather than remaining itself, which is particularly obvious after a sequence of such improvements. Taken to the extreme, this idea implies that a simple act of learning new information transforms you into a different mind raising millennia old questions about the nature of personal identity." -> this is only a problem if, like the article, you are extremely confused about the definition of a unique mind, or permit equivocation thereof.

"Interestingly, a perfect ability by two minds to predict each other would imply that they are identical..." -> Why? Picture two minds whose entire environment consists of a single switch with ON and OFF. Mind A wants the switch ON. Mind B wants the switch OFF. Each can correctly predict the actions of the other - flipping the switch in the desired direction at every opportunity - but they are clearly different minds. (Especially if we think that having a favorite number of 32 instead of 33 is enough to make two minds "different".)

The taxonomy section is a grab bag of mostly unrelated classification systems, useful at least as examples of how many ways minds can vary.

"For example, asking a human if he is hungry may produce an answer different from the one which would be given by a non-biological robot." -> Once again, by the article's own claims in "Infinitude of Minds" about what makes a mind unique, if one mind experiences hunger and the other does not, then the two minds are not identical, unless we consider "ability to experience hunger" a smaller difference than "favorite number."