AGI and the Knight-Darwin Law

Samuel Allen Alexander[1[0000−0002−7930−110X]]

The U.S. Securities and Exchange Commission samuelallenalexander@gmail.com
https://philpeople.org/profiles/samuel-alexander/publications

Abstract. This is a paper about AGIs (agents with Artificial General Intelligence), and more specifically, about the structure of the populations that result when parent AGIs create child AGIs. We argue that such populations satisfy a certain property similar to a property of biological populations, the Knight-Darwin Law. The biological Knight-Darwin Law states that it is impossible for one organism to asexually produce another organism, which asexually produces another, and so on forever: that any sequence of organisms (each one a child of the previous) must contain a multi-parent organism, or must terminate. By showing that a certain intelligence measure decreases when a parent AGI single-handedly creates a child AGI under certain assumptions, we argue that a similar Law holds for AGIs.

Keywords: Intelligence Measurement · Knight-Darwin Law · Ordinal Notations · Intelligence Explosion

1 Introduction

One thing that distinguishes AGIs from weaker AIs is that an AGI ought to be capable of programming AGIs[1]. After all, the first AGIs will, apparently, need to be programmed by humans. If AGIs are at least as smart as humans, and if humans can build AGIs, then so should AGIs.

AGIs programming AGIs is difficult to reason about. To get our hands on something solid, we have attempted to find structures that abstractly capture the core essence of AGIs programming AGIs. This led us to discover what we call the *Intuitive Ordinal Notation System* (presented in Section 2), an ordinal notation system that gets directly at the heart of AGIs creating AGIs.

We call an AGI *truthful* if the things it knows (or could eventually deduce) are true, i.e., if it does not know any falsehoods. In [3], we used the Intuitive Ordinal Notation System to argue that if a truthful AGI X creates (without external help) a truthful AGI Y, in such a way that X knows the truthfulness of Y, then X must be more intelligent than Y in a certain formal sense. The

[1] Within this paper, our approach to AGI is what Goertzel [12] describes as the Universalist Approach: we consider "...an idealized case of AGI, similar to assumptions like the frictionless plane in physics", with the hope that by understanding this "simplified special case, we can use the understanding we've gained to address more realistic cases."

argument is based on the key assumption that if X creates Y, without external help, then X necessarily knows Y's sourcecode.

Iterating the above argument, suppose $X_1, X_2, \ldots$ are truthful AGIs such that each X_i creates, and knows the truthfulness and the sourcecode of, X_{i+1}. By the previous paragraph, X_1 would be more intelligent than X_2, which would be more intelligent than X_3, and so on. If intelligence is well-founded (i.e., if it is impossible for a chain of AGIs to keep getting less and less intelligent without eventually hitting some minimum possible intelligence level), then this would imply that it is impossible for such a list $X_1, X_2, \ldots$ to go on forever: it would have to stop after finitely many elements.

At first glance, the above results might seem to suggest skepticism regarding the singularity—or regarding what Hutter [16] calls *intelligence explosion*, the idea of AGIs creating better AGIs, which create even better AGIs, and so on. But there is a loophole. Suppose X and X' are AGIs who collaborate to create Y. Suppose X contributes sourcecode for part of Y, but keeps it secret from X', and suppose X' contributes sourcecode for another part of Y, but keeps it secret from X. Then neither X nor X' knows Y's full sourcecode, and the above-mentioned argument breaks down.

The Knight-Darwin Law [10], named after Charles Darwin and Andrew Knight, is the principal (rephrased in modern language) that there cannot be an infinite sequence $X_1, X_2, \ldots$ of biological organisms such that each X_i asexually parents X_{i+1}. In other words, if $X_1, X_2, \ldots$ is any infinite list of organisms such that each X_i is a biological parent of X_{i+1}, then some of the X_i would need to be multi-parent organisms. The reader will immediately notice a striking parallel between this principal and the discussion in the previous two paragraphs.

In Section 2 we present the Intuitive Ordinal Notation System.

In Section 3 we argue that if a truthful AGI X creates a truthful AGI Y, such that X knows the truthfulness and the sourcecode of Y, then Y is less intelligent than X. This argument appeared in a fully rigorous form in [3], but in this paper we attempt to make it more approachable by informally giving the key idea behind it.

In Section 4 we adapt the Knight-Darwin Law from biology to AGI and speculate about what it might mean for AGI, including implications about the creation of AGI.

In Section 5 we address some anticipated objections.

In Section 6 we summarize and make concluding remarks.

2 The Intuitive Ordinal Notation System

We do not know what, exactly, an AGI is, but we want to study AGIs anyway. Based on our conviction that an AGI should be capable of programming AGIs, we would like to come up with a more concrete structure, easier to reason about, which we can use to better understand AGIs.

What sort of structure would capture the essence of an AGI's AGI-programming capability? As an initial attempt, how about: "computer program that prints computer programs"? That seems like the right direction to go, but as written, there is no constraint on the printed computer programs, so this initial attempt seems to capture the essence of an AGI's capability of writing *computer programs*, rather than of writing *AGIs*.

As a second attempt, how about: "computer program that prints computer programs that print computer programs"? This brings us closer, but this second attempt seems to capture the essence of an AGI's capability of writing *first-attempt programs*, rather than of writing *AGIs*.

As a third attempt, how about: "computer program that prints computer programs that print computer programs that print computer programs"? This brings us closer still, but this third attempt seems to capture the essence of an AGI's capability of writing *second-attempt programs*, rather than of writing *AGIs*.

We need to short-circuit the above process. We need to come up with a notion X which is equivalent to "computer program that prints members of X".

Definition 1 *(See the following examples) We define the Intuitive Ordinal Notations to be the smallest set $\mathcal{P}$ of computer programs satisfying the following property:*

– *For every computer program p, if $\mathcal{P}$ contains every output that p prints when p is executed, then $p \in \mathcal{P}$.*

Example 2 *(Some simple examples)*

1. *Let P_0 be "End", a program which immediately stops without printing any outputs. Vacuously, every output printed when P_0 is executed is in $\mathcal{P}$ (there are no such outputs). So P_0 is an Intuitive Ordinal Notation.*
2. *Let P_1 be "Print('End')", a program which prints "End" and then stops. By (1), every output printed by P_1 is an Intuitive Ordinal Notation, therefore, so is P_1.*
3. *Let P_2 be "Print('Print('End')')", which prints "Print('End')" and then stops. By (2), every output printed by P_2 is an Intuitive Ordinal Notation, therefore, so is P_2.*

Example 3 *(A more interesting example) Let P_ω be the program:*

Let X = 'End'; While(True) { Print(X); X = "Print('" + X + "')"; }

When executed, P_ω prints "End", "Print('End')", "Print('Print('End')')", and so on forever. As in Example 2, all of these are Intuitive Ordinal Notations. Therefore, P_ω is an Intuitive Ordinal Notation.

Because of the recursive way the Intuitive Ordinal Notations are defined, there is a natural way to assign computable ordinal numbers to them. How do we assign a computable ordinal to an Intuitive Ordinal Notation x? We can assume (by induction) that we've already assigned computable ordinals to all the outputs printed when x is executed (this makes sense because all such outputs are Intuitive Ordinal Notations, by definition). So the answer is clear: assign x the first computable ordinal bigger than all the computable ordinals assigned to the outputs of x.

Definition 4 *For any Intuitive Ordinal Notation x, we define a computable ordinal $|x|$ inductively as follows: $|x|$ is the smallest computable ordinal α such that $\alpha > |y|$ for every output y that gets printed when x is executed.*

Example 5
- *Since P_0 (from Example 2) has no outputs, it follows that $|P_0| = 0$, the smallest computable ordinal.*
- *Likewise, $|P_1| = 1$ and $|P_2| = 2$.*
- *Likewise, P_ω (from Example 3) has outputs notating $0, 1, 2, \ldots$—all the finite natural numbers—and no other outputs. It follows that $|P_\omega| = \omega$, the smallest infinite ordinal number.*
- *Let $P_{\omega+1}$ be the program "Print(P_ω)", where P_ω is as in Example 3. It follows that $|P_{\omega+1}| = \omega + 1$, the next ordinal after ω.*

The Intuitive Ordinal Notation System is a more intuitive simplification of Kleene's $\mathcal{O}$ ordinal notation system [18].

3 Inductive Ordinal Intelligence

Whatever an AGI is, presumably an AGI knows certain mathematical facts (or could eventually deduce them). The following is a notion of an AGI's intelligence based solely on the mathematical facts that the AGI knows. In [3] we argue that this notion captures key components of intelligence such as pattern recognition, creativity, and the ability to abstract and to generalize. We will say more to justify this definition in Section 5.1.

Definition 6 *The* Intuitive Ordinal Intelligence *of a truthful AGI X is the smallest computable ordinal α such that $\alpha > |p|$ for every Intuitive Ordinal Notation p such that X knows (or could eventually deduce) that p is an Intuitive Ordinal Notation.*

The following theorem provides a relationship between Intuitive Ordinal Intelligence and AGI creation of AGI.

Theorem 7 *Suppose X is a truthful AGI, and X creates a truthful AGI Y, in such a way that X knows Y's sourcecode and truthfulness. Then X has higher Intuitive Ordinal Intelligence than Y.*

Proof. Suppose Y were commanded to spend eternity enumerating the biggest Intuitive Ordinal Notations Y could think of. This would result in some list L of Intuitive Ordinal Notations enumerated by Y. Since Y is an AGI, L must be computable, because L can be enumerated by a robot (namely by Y). Thus, there is some computer program P whose outputs are exactly L. Since X knows Y's sourcecode, and presumably X is capable of elementary reasoning about sourcecodes, it follows that X can infer P—that is, X knows (or could eventually deduce) the statement: "P outputs L, the list of things Y would print if Y were commanded to enumerate Intuitive Ordinal Notations for all eternity". Since X knows Y is truthful, X can deduce that L contains nothing except Intuitive Ordinal Notations, thus X can deduce that P's outputs are Intuitive Ordinal Notations, and so X knows (or could deduce) that P is an Intuitive Ordinal Notation. So X's Intuitive Ordinal Intelligence is $> |P|$. By construction, $|P|$ is the least ordinal bigger than $|Q|$ for all Q printed by L, in other words, $|P|$ is the Intuitive Ordinal Intelligence of Y. □

4 The Knight-Darwin Law

> "...it is a general law of nature that no organic being self-fertilises itself for a perpetuity of generations; but that a cross with another individual is occasionally—perhaps at long intervals of time—indispensable." (Charles Darwin)

In his Origin of Species [9], Darwin devotes many pages to the above-quoted principal, later called the Knight-Darwin Law [10]. In [1] and [4], we translate the Knight-Darwin Law into mathematical language.

Principle 8 *(The Knight-Darwin Law) There cannot be an infinite sequence $x_1, x_2, \ldots$ of organisms such that each x_i is the lone biological parent of x_{i+1}. If each x_i is a parent of x_{i+1}, then some x_{i+1} must have multiple parents.*

Of course Darwin was aware of seemingly asexual plant species, but a key motivation for the above law was the observation that even seemingly asexual plant species occasionally reproduce sexually. For example, a plant in which pollen is ordinarily isolated, might release said pollen into the air if a storm damages the part of the plant that would otherwise shield said pollen. Even prokaryotes can be considered to occasionally have multiple parents if lateral gene transfer is taken into account.

A key fact about the ordinal numbers is they are *well-founded*: there is no infinite sequence $o_1, o_2, \ldots$ of ordinals such that each $o_i > o_{i+1}$. In Theorem 7 we showed that if truthful AGI X creates truthful AGI Y, in such a way as to know the truthfulness and sourcecode of Y, then X has a higher Intuitive Ordinal Intelligence than Y. Combining this with the well-foundedness of the ordinals yields a theorem extremely similar to the Knight-Darwin Law.

Theorem 9 *(The Knight-Darwin Law for AGIs) There cannot be an infinite sequence $x_1, x_2, \ldots$ of truthful AGIs such that each x_i creates x_{i+1} in such a way as to know x_{i+1}'s truthfulness and sourcecode. If each x_i creates x_{i+1} (intending x_{i+1} to be truthful), then some x_{i+1} must be co-created by multiple creators (assuming that creation by a lone creator implies the lone creator would know x_{i+1}'s sourcecode).*

Proof. By Theorem 7, the Intuitive Ordinal Intelligence of $x_1, x_2, \ldots$ would be an infinite strictly-descending sequence of ordinal numbers, which would violate the well-foundedness of the ordinals. □

In the remainder of this section, we will non-rigorously speculate about three implications Theorem 9 might have for AGIs and for AGI research.

4.1 Motivation for Multi-agent Approaches to AGI

In the Introduction, we argued that an AGI ought to be capable of programming AGIs. If so, Theorem 9 suggests that a fundamental aspect of AGI should be the ability to collaborate with other AGIs in the creation of new AGIs.

One interesting idea along these lines is that there should be no such thing as a *solipsistic* AGI, or at least, solipsistic AGIs would be limited in their ability to reproduce. For, if an AGI were solipsistic, it seems like it would be difficult for this AGI to collaborate with other AGIs to create child AGIs.

Along more practical lines, Theorem 9 suggests it may be worth prioritizing research into multi-agent systems, for example [5], [15], [19], and similar work.

4.2 Motivation for AGI Variety

The Knight-Darwin Law was important to Darwin because it was a foundation for a broader thesis that the survival of a species depends on the inter-breeding of a variety of members of that species. By analogy, we might speculate that if our goal is to create robust AGIs, then we ought to focus on creating a wide variety of different kinds of AGIs, so that those AGIs can collaborate together[2] to co-create even stronger AGIs.

On the other hand, if we want to reduce the danger of AGI getting out of control, perhaps the lesson to take away is to limit AGI variety. At the most extreme end of the spectrum, if humankind were to limit itself to only creating one single AGI, then Theorem 9 would constrain the extent to which that AGI could reproduce.

4.3 AGI Genetics

If, as Theorem 9 seems to suggest, AGI collaboration is a fundamental requirement for an AGI "population" to propagate, then it might someday be possible to view AGI through a genetic lens[3]. If AGIs X and X' co-create child AGI Y by means of X contributing proprietary sourcecode for one submodule, concealed from X', while X' contributes proprietary sourcecode for another submodule, concealed from X, perhaps there might be some notion of AGI "genes". For example, if X runs operating system O, and X' runs operating system O', perhaps their joint offspring will somehow exhibit traces of both O and O'.

[2] Anticipated by [8].
[3] Anticipated by [7].

5 Objections

In this section, we address some anticipated objections.

5.1 What does Definition 6 have to do with intelligence?

Whatever intelligence is, it has core components like pattern-matching, creativity, and the ability to abstract and to generalize. I claim that these core parts of intelligence are indispensable if one wants to competitively name large ordinals. If α is an ordinal whose notation was obtained using certain techniques, then *any* AGI who used those techniques to notate α should also be able to iterate those same techniques. Thus, to advance from α to a larger ordinal which not just *any* α-knowing AGI could obtain, must require the creative invention of some new technique or tactic, and this invention of a new technique requires some amount of creativity, pattern-matching, etc. This becomes clear if the reader attempts to go beyond Example 3 and notate qualitatively larger ordinals; see the more extensive examples in [3].

We will say more. For analogy's sake, imagine a ladder which different AGIs can climb, and suppose advancing up the ladder requires exercising different components of intelligence. Since we have so little other idea how to measure intelligence, one way to measure (or at least estimate) it would be to measure how high an AGI can climb on said ladder.

Not all ladders are equally good. A ladder would be particularly poor if it had a top rung which many AGIs could reach: for then, the ladder would fail to distinguish between different AGIs who could reach that top rung, even if one AGI reaches it with ease and another AGI reaches it with difficulty. Even if the ladder was infinite and had no top rung, it would still be suboptimal if there were many different AGIs capable of scaling the whole ladder (i.e., of ascending however high they like, on demand)[4]. A good ladder should have the property that for every particular AGI, there is some rung which that AGI cannot reach.

Definition 6 provides a good ladder. The rungs which an AGI does manage to reach, we have argued, require the usage of core components of intelligence to reach. And no particular AGI can scale the whole ladder, because no AGI can enumerate all the Intuitive Ordinal Notations: it can be shown that the set of Intuitive Ordinal Notations is not computably enumerable. Neither is our ladder overly-high, because for any particular ordinal notation, an AGI could presumably be programmed with knowledge of that notation hard-coded.

[4] Hibbard's intelligence measure [14] is an example of an infinite ladder which is nevertheless short enough that many AGIs can scale the whole ladder. Those whole-ladder-scalers are the AGIs which do not "have finite intelligence" in Hibbard's words (defined implicitly, see Hibbard's Proposition 3). It should be possible to use a *fast-growing hierarchy* [13] [21] to extend Hibbard's ladder into the transfinite and reduce (or eliminate?) the set of AGIs capable of scaling the whole ladder. This transformation would make Hibbard's measurement ordinal-number-valued rather than natural-number-valued (perhaps Hibbard himself intuited this, since his abstract uses the word "ordinal" in its everyday sense as a synonym for "natural number").

5.2 Shouldn't intelligence depend on environmental interaction?

We do not claim that Definition 6 is the "one true measure" of intelligence. Maybe there is no such thing: maybe intelligence is inherently multi-dimensional. Definition 6 measures one type of intelligence, a type based on mathematical knowledge[5]. Some AGI might be good at video games but poor at inventing ordinals, or vice versa. But any truly broad AGI should at least be capable (if properly instructed) of attempting any reasonable well-defined task, including the task of notating ordinals. So while Definition 6 is not the "one true measure", it is a measure of something having to do with an AGI's abilities. It is possible that a different word, such as "knowledge-level", would be more appropriate than "intelligence": but that would not qualitatively change Theorems 7 or 9, nor invalidate the parallel between the biological Knight-Darwin Law and the AGI Knight-Darwin Law.

It's worth mentioning that it is perfectly consistent with Theorem 7 that X's creation Y might perform computations faster than X, thereby performing better in realtime environments (as in [11]). In this light, Theorems 7 and 9 are profound because they suggest that children and grandchildren might appear superficially more intelligent initially, yet nevertheless, without outside help, they will eventually degenerate. The situation parallels the *hydra game* of Kirby and Paris [17], where a certain hydra seems to become enormously stronger as the player hacks off its heads, and yet the hydra inevitably dies (possibly after a very long time) if the player keeps hacking.

5.3 Can't an AGI just print a copy of itself?

If an AGI knows its own sourcecode, then there's nothing stopping that AGI from printing a copy of itself. However, if it does so, then it necessarily cannot know the truthfulness of that copy, lest it would know the truthfulness of itself. Gödel's incompleteness theorems imply that a suitably idealized AGI cannot know its own sourcecode and its own truthfulness (see [2]).

5.4 Prohibitively expensive simulation

The reader might object that Theorem 7 breaks down if Y is prohibitively expensive for X to simulate or emulate. But Theorem 7 and its proof have nothing to do with simulation or emulation. A person can reason about sourcecode without running it. In the same way, if X knows the sourcecode of Y, then X can reason about that sourcecode without executing a single line of it.

5.5 Computable enumerability of mathematical theorems

"Isn't the set of all mathematical theorems computably enumerable? Coundn't an AGI just enumerate them all by brute force, and wouldn't that AGI have

[5] Wang has correctly pointed out [22] that an AGI consists of much more than merely a knowledge-set of mathematical facts.

unbeatable intelligence according to Definition 6?" The answer is no. The set of mathematical theorems is only computably enumerable when a computably enumerable base set of axioms is agreed upon in the background, in which case, the theorems provable *from those axioms* are computably enumerable. For example, the theorems of Peano arithmetic are computably enumerable, as are the theorems of ZFC, etc. By Gödel's incompleteness theorem, for any computably enumerable true set of axioms capable of basic arithmetic, there is always a true mathematical statement not provable by those axioms. In fact, Gödel's theorem constructively tells us how to obtain such a statement, and an AGI could easily follow Gödel's recipe. Thus, it is unlikely that there is some fixed computably enumerable set of theorems which all AGIs universally are limited to.

6 Conclusion

We reviewed, from [3], an intelligence measurement notion for a truthful AGI based on that AGI's mathematical knowledge. In short: the Intuitive Ordinal Intelligence of a truthful AGI is the supremum of the computable ordinals which have notations the AGI knows to be notations of computable ordinals. We discussed why we think this counter-intuitive intelligence measurement notion indeed measures (at least a type of) intelligence.

The intelligence measure which we introduced measures intelligence not using real numbers, but rather using computable ordinal numbers. These have the so-called *well-ordering property*: there is no infinite, strictly-descending sequence of computable ordinals. We proved that if a truthful AGI creates a child truthful AGI, in such a way as to know the child's truthfulness and sourcecode, then the parent must necessarily be more intelligent than the child according to the intelligence notion introduced here. Together with the well-foundedness of the ordinals, this allowed us to establish a structural property for AGI populations, a property which closely resembles a property of biological populations[6], the Knight-Darwin Law. We speculated about implications of this remarkable parallel between biology and AGI (our speculations here are not intended to be exhaustive, and we expect there are many more implications besides). Our hope is that by better understanding the process by which AGIs must create new AGIs, we can make progress toward a more ambitious goal: the creation of AGIs by humans.

References

1. Alexander, S.A.: Infinite graphs in systematic biology, with an application to the species problem. Acta Biotheoretica **61**(2), 181–201 (2013)
2. Alexander, S.A.: A machine that knows its own code. Studia Logica **102**(3), 567–576 (2014)

[6] For some other interesting work exploring AGI connections to animals and their populations, see [6], [20], and [23].

3. Alexander, S.A.: Measuring the intelligence of an idealized mechanical knowing agent. In: Cognition, Interdisciplinary Foundations, Models, and Applications (CIFMA). Springer (2019)
4. Alexander, S.A., de Bruin, A., Kornet, D.J.: An alternative construction of internodons: the emergence of a multi-level tree of life. Bulletin of mathematical biology **77**(1), 23–45 (2015)
5. Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., Mordatch, I.: Emergent complexity via multi-agent competition. arXiv preprint arXiv:1710.03748 (2017)
6. Beyret, B., Hernández-Orallo, J., Cheke, L., Halina, M., Shanahan, M., Crosby, M.: The animal-ai environment: Training and testing animal-like artificial cognition. arXiv preprint arXiv:1909.07483 (2019)
7. Buchanan, B.G.: Artificial intelligence as an experimental science. In: Aspects of Artificial Intelligence, pp. 209–250. Springer (1988)
8. Dai, P., Weld, D.S., et al.: Artificial intelligence for artificial artificial intelligence. In: Twenty-Fifth AAAI Conference on Artificial Intelligence (2011)
9. Darwin, C.: On the origin of species. John Murray, London, 6th edn. (1872)
10. Darwin, F.: The knight-darwin law. Nature **58**, 630–632 (1898)
11. Gavane, V.: A measure of real-time intelligence. Journal of Artificial General Intelligence **4**(1), 31–48 (2013)
12. Goertzel, B.: Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence **5**(1), 1–48 (2014)
13. Hardy, G.H.: A theorem concerning the infinite cardinal numbers. Quarterly journal of Mathematics **35**, 87–94 (1904)
14. Hibbard, B.: Measuring agent intelligence via hierarchies of environments. In: International Conference on Artificial General Intelligence. pp. 303–308. Springer (2011)
15. Hibbard, B.: Societies of intelligent agents. In: International Conference on Artificial General Intelligence. pp. 286–290. Springer (2011)
16. Hutter, M.: Can intelligence explode? Journal of Consciousness Studies **19**(1-2), 143–166 (2012)
17. Kirby, L., Paris, J.: Accessible independence results for peano arithmetic. Bulletin of the London Mathematical Society **14**(4), 285–293 (1982)
18. Kleene, S.C.: On notation for ordinal numbers. The Journal of Symbolic Logic **3**(4), 150–155 (1938)
19. Lazaridou, A., Hermann, K.M., Tuyls, K., Clark, S.: Emergence of linguistic communication from referential games with symbolic and pixel input. In: International Conference on Learning Representations (2018)
20. Strannegård, C., Svangård, N., Lindström, D., Bach, J., Steunebrink, B.: Learning and decision-making in artificial animals. Journal of Artificial General Intelligence **9**(1), 55–82 (2018)
21. Wainer, S., Buchholz, W.: Provably computable functions and the fast growing hierarchy. In: Simpson, S. (ed.) Logic and Combinatorics, vol. 65. American Mathematical Society (1987)
22. Wang, P.: Three fundamental misconceptions of artificial intelligence. Journal of Experimental & Theoretical Artificial Intelligence **19**(3), 249–268 (2007)
23. Yoshida, N.: Homeostatic agent for general environment. Journal of Artificial General Intelligence **8**(1), 1–22 (2017)