A Theory of Intelligence Metrics: Measuring the Degree of Intelligence in Search

William A. Dembski
Discovery Institute's Center for Science and Culture
Cornerstone University

ABSTRACT: The study of intelligence raises two key questions: (1) How can intelligence be detected or inferred from its effects, and (2) how can the degree of intelligence in effects be measured? To date, the first of these questions has received the bulk of scholarly attention. Thus, a standardized inferential method has emerged for SETI, archaeology, forensic science, and the historical sciences generally in which some trace, signal, or pattern is observed and the question then becomes whether it is best explained as the result of intelligence. This paper focuses on the second of these questions. It develops a general theory of intelligence metrics based on information theory. The proposed metrics assign numbers to the intelligence exhibited in successful search. This paper therefore treats search as the universal paradigm for intelligence. By framing the measurement of intelligence in terms of search, this paper avoids presupposing any particular view of intelligent causation. Instead, the theory presented here applies across diverse causal settings, including human problem solving, large language models, and evolutionary processes. It thus provides an overarching informational framework for comparing forms of intelligence, for assessing the efficacy with which intelligences use informational resources, and specifically for determining how much intelligence is manifested in successful search.

ACKNOWLEDGMENT: This paper arose out of a talk I gave at the 2023 Conference on Engineering in Living Systems (CELS), convened by Discovery Institute's Engineering Research Group and organized by Steve Laufmann and Eric Anderson. I'm grateful to all who helped bring about this conference, to its participants, and especially to Steve and Eric for encouraging me to present the ideas in this paper when they were still in much less developed form.

Table of Contents

1	Introduction	2
2	The Connection Between Information and Intelligence	6
3	Mac the Canine Door Opener	9
4	The Connection Between Probability and Information	11
5	Interpreting Information as an Intelligence Metric	13
6	Task Specificity of Intelligence	14
7	Dominating versus Non-Dominating Intelligence	15
8	Conditional versus Unconditional Intelligence Metrics	17
9	Absolute versus Probabilistic Intelligent Performance	19
10	Search as Key to Intelligent Performance	20
11	Performance and Failure Functions	26
12	Cumulative Performance Functions	31
13	Weighted Performance Measures	35
14	Multi-Criteria vs. Single-Criterion Optimization	38
15	Percentile Performance/Failure Functions	39
16	Effect-Size Performance/Failure Functions	41
17	Random Predicates as Generalized Performance/Failure Functions	46
18	Measuring Intelligence in the Search for Targets	49
19	Negative Intelligence and Disinformation	53
20	Sequential Search and Waiting Times	54
21	Complexity as a Measure of Intelligence	59
22	Kolmogorov Complexity as Specified Complexity	62
23	The Intelligence of Natural Selection	68
24	Why LLMs Aren't Smart—Lean Versus Laden Intelligence	74
25	Measuring Intelligence in Information Packing	77
26	Conclusion	81
	Notes	84

1. Introduction

Scientists routinely engage in detecting and inferring whether something is the product of intelligence. Thus forensic science distinguishes natural causes from foul play, archeology distinguishes geofacts from artifacts, and the Search for

Extraterrestrial Intelligence (SETI) distinguishes electromagnetic noise from technosignatures.[1] The underlying method of reasoning here is from effect to cause: it identifies some trace, signal, or pattern (the effect) and from there attempts to determine whether it is best explained as the product of intelligence (the cause).

To date, the lion's share of scholarly and research attention has focused on whether something is the product of intelligence, less so on determining the degree of intelligence needed to produce it. The role of intelligence has thus tended to be treated in all-or-nothing terms: either it's the product of intelligence or it's not. This distinction can be invidious in that it may presuppose an underlying view of intelligence and thus set a sharp distinction between things that are and are not intelligently caused, obscuring that the real question may instead be over differences in degrees of intelligence.

This all-or-nothing approach to detecting and inferring intelligence affects how scientists use probabilities to attribute intelligence. Thus, in determining whether a signal from outer space reveals a technosignature, a SETI researcher (whether explicitly or tacitly) will set a probability threshold so that salient events or objects whose probability is smaller than that threshold trigger an inference to intelligence, but not otherwise. Failing to meet the threshold cannot rule out the activity of intelligence because intelligences can purposely mimic processes and outcomes that appear unintelligent—as when photographers physically rearrange war scenes to make them look more chaotic.[2] But failing to meet the threshold will fail to confirm the activity of intelligence.[3]

To illustrate design detection, imagine a group of exactly five people seated at a restaurant table. The birthday of each of them is January 1. No one would think that these people just randomly congregated at this table. Assuming birthdays are equally distributed throughout the year and ignoring leap years (the actual empirical probabilities are close to this), the probability of them all having this birthday is roughly 1 in 6.48 trillion. Such a level of improbability is much more plausibly ascribed to intelligence rather than happenstance, which is to say that these tablemates were seated together on purpose.

The concern might now be raised that this probability was miscalculated because there's nothing special about January 1. Instead, any calendar date would have served equally well because the people seated at the restaurant table might have shared a birthday for any other day of the year. For other days of the year, such a coincidence would have been equally salient. But factoring in those additional days would only raise the probability to 1 in 17.75 billion, leaving us still disinclined to ascribe this coincidence to chance.

Of course, if there were exactly ten people at the table sharing a common birthday, the probability would get much smaller still. Such a chance coincidence would seem even more implausible, with the probability in this case being 1 in 114 sextillion, and the inference to intelligence being that much more compelling. Yet in an example like this, we are not so concerned about the degree of intelligence needed to bring together people sharing a common birthday. Instead, we're simply concerned about whether an intelligence did so.

Probability thresholds seek to determine whether an intelligence has acted, not the degree of intelligence that acted. These thresholds are important because the first thing we want to know when the activity of intelligence stands in question is whether what seems to be the product of intelligence is in fact so. Nonetheless, when we reflect on whether an intelligence has acted, we think of it not merely as all-or-nothing but also as coming in degrees. In many instances where the role of intelligence stands in question, it's important to know not only whether intelligence was involved but also the degree of intelligence involved.

Was the intelligence in question a superintelligence, well beyond the capacities of humans—a ramped-up transhumanist AGI (artificial general intelligence) or an all-wise transcendent deity? Was it a bumbling intelligence, barely competent to bring about the design in question—Hume's incompetent designer who "botched and bungled"[4] many worlds before finally producing this one, whose quality Hume also pooh-poohed? Or was the intelligence somewhere in the middle?

The aim of this paper is to lay out a theory of intelligence metrics that applies quite generally but also to the structures and processes in nature. Such metrics measure task-specific manifestations of intelligence. Notwithstanding, this paper goes beyond merely trying to ape IQ or intelligence quotient metrics in understanding the degree of intelligence behind such items.

IQ tests have lost credibility in our day for being narrowly focused on certain cognitive tasks while being advertised as capturing the essence of what it means for humans to be intelligent.[5] Intelligence is nowadays better understood as taking many distinct forms (emotional, kinesthetic, linguistic, musical, mathematical, etc.).[6] Intelligence is now seen as task- and context-specific, so that an intelligence adept in one situation may not be adept in another. Additionally, intelligence is now seen as dynamic rather than static, subject to improvement through learning and experience but also to deterioration through disuse, damage, or sabotage.[7]

The intelligence metrics developed in this paper attempt to capture a wide range of intelligence, making due allowance for the different forms that intelligence can take. One benefit of these metrics is that we will be able to use them to measure the intelligence in processes that we might otherwise think to be unintel-

ligent because they lack a conscious rational agent. For instance, it will be possible to measure the intelligence inherent in an evolutionary process (such as one driven by natural selection acting on random variations) and thereby assess its adequacy or inadequacy in bringing about the things that evolutionists commonly ascribe to evolution, such as the formation of irreducibly complex biological systems.[8] So too, these metrics can be used to measure the intelligence in generative AI.

This paper is not the first to propose a general approach to measuring intelligence. The contemporary literature on intelligence metrics ranges from formal measures that aim at universality to operational measures designed for empirical use. Shane Legg and Marcus Hutter's 2007 measure of "universal intelligence" exemplifies the formal approach, averaging the rewards an agent can extract from a range of environments. The resulting intelligence metric, because based on Kolmogorov complexity, is non-computable. Consequently, it is more a definition of intelligence than a practical measure of it. Subsequent extensions of their framework—such as Alexander and Hutter's (2021) reward-punishment symmetric version and Oswald et al.'s (2024) extension to "uncomputable environments that can be classified on the Arithmetical Hierarchy" are likewise difficult to apply in practice.[9]

François Chollet (2019) defined intelligence in terms of "skill-acquisition efficiency": systems are credited with intelligence not just for skillful performance but also for acquiring skills on difficult tasks with limited priors, experience, and training information. Chollet formalized "generalization difficulty of a task" using Kolmogorov complexity, so his intelligence metric inherits the non-computability of Kolmogorov complexity. Unlike a purely theoretical measure, however, Chollet calls for computable approximations and introduces the Abstraction and Reasoning Corpus (ARC) as a benchmark for comparing human and artificial generalization. His approach therefore intersects with concerns in this paper over balancing task difficulty with informational resources: accomplishing the same task with fewer resources signifies greater intelligence.[10]

Recent work also focuses on operational measures of intelligence. José Hernández-Orallo and colleagues (2021) see intelligence as combining capability and generality, treating performance as a function of task difficulty under bounded resources and apply their measures to humans, animals, and AI systems.[11] Fernando Martínez-Plumed and colleagues (2022) use item response theory (IRT) to estimate the difficulty of individual AI test instances, distinguishing failures due to the hardness of problems from failures due to weakness in the system.[12] Maharshi Gor and colleagues (2024) have introduced a IRT-based framework they call Content-Aware, Identifiable, and Multidimensional Item Response Analysis

(CAIMIRA) that balances agent skills and question difficulty, both for human and artificial respondents. Their analysis draws on more than 300,000 responses from roughly 70 AI systems and 155 humans.[13]

Advances in AI have accelerated this operational trend for measuring intelligence. Xiting Wang and colleagues (2026) argue for using psychometric tools (such as reliability, validity, and IRT) to evaluate general-purpose AI rather than relying on task-performance benchmarks.[14] In a related vein, Lexin Zhou and colleagues (2026) develop general demand scales that place task instances on eighteen cognitive and intellectual dimensions and derive corresponding ability profiles for AI systems. Their aim is to provide a "solid foundation for a science of AI evaluation, underpinning the reliable deployment of AI in the years ahead."[15]

In relation to this ongoing research to understand and measure intelligence, the approach taken in this paper attempts a more fundamental analysis. It seeks to understand the measurement of intelligence in terms of first principles. At the same time, it ties intelligence metrics to real-world intelligences, with the metrics mirroring our ordinary assignments of degrees of intelligence. The proposed theory of intelligence metrics aims at a broad synthesis. It treats successful search as the paradigm case of intelligent performance. Its fundamental intelligence metric is, in various guises, Shannon information, later supplemented by complexity measures such as Kolmogorov complexity. The metrics developed here are meant to apply on equal footing to human agents, LLMs, and natural processes such as evolution by natural selection.

2. The Connection Between Information and Intelligence

The key intuition behind the modern conception of information is the *narrowing of possibilities*. In this conception, drawn from Shannon's communication theory and subsequent work on the mathematical theory of information,[16] the key truth about information is this: *the more that possibilities are narrowed down, the greater the information*. This conception is consistent with Aristotle's classical conception of information through his theory of substantial forms, in which a formal cause combines with a material cause to narrow down the precise structure and function

of a thing. The modern conception, however, based as it is in mathematics, makes sense apart from a fully developed metaphysics, such as that of Aristotle.

To illustrate the modern conception of information, imagine I tell you I'm on planet earth. In that case, I haven't conveyed any information because you already knew that (let's leave aside space travel). If I tell you I'm in the United States, I've begun to narrow down where I am in the world. If I tell you I'm in Texas, I've narrowed down my location further. If I tell you I'm forty miles north of Dallas, I've narrowed my location down even further. As I keep narrowing down my location, I'm providing you with more and more information.

Information is therefore, in its essence, exclusionary: the more possibilities are excluded, the greater the information provided. As philosopher Robert Stalnaker put it: "To learn something, to acquire information, is to rule out possibilities. To understand the information conveyed in a communication is to know what possibilities would be excluded by its truth."[17] I'm excluding much more of the world when I say I'm in Texas forty miles north of Dallas than when I merely say I'm in the United States. Accordingly, to say I'm in Texas north of Dallas conveys much more information than to say I'm in the United States.

The etymology of the word information captures this exclusionary understanding of information. The word information derives from the Latin preposition *in*, meaning in or into, and the verb *formare*, meaning to give shape to. Information puts definite shape into something. But that means ruling out other shapes. Information narrows down the shape in question. A completely unformed shmoo, such as Aristotle's prime matter, is waiting in limbo to receive information. Only by being informed will it exhibit the definiteness of a real thing.

Aristotle's conception of information overlaps with but is also separate from the modern conception of information. Aristotle's conception is tied to his theory of formal causation, in which information is understood as the cause that gives shape to matter, thereby makes a material object what it is, and ultimately allows it to achieve its purpose (for Aristotle everything had a purpose). In Aristotelian thought, the formal cause determines an object's structure and properties, defining its essence.

For Aristotle, information was thus more than a narrowing of possibilities. Instead, it was an intrinsic organizing principle that turns matter into a coherent and purposeful entity. The modern conception of information, though not wedded to Aristotle's understanding of formal causation, is nonetheless consistent with it. Aristotelian information, by defining a thing's essence, makes it this and not that. It is thus inherently exclusionary, which aligns with information in its contemporary sense as the narrowing down of possibilities.

Because information rules out possibilities, its mathematical representation is naturally top-down: one begins with a space of possibilities and then identifies the subset compatible with what is observed or achieved. The space of possibilities comes first, laying out the things that might be true in detail. And then these possibilities are narrowed down, in an act or realization, realizing some possibilities to the exclusion of others. This representational order is formal, not ontological. Yet the underlying logic is opposite to materialism, which is always a bottom-up affair, starting with material constituents and then attempting to build them into coherent assemblies of parts.

In ruling out possibilities, information is fundamentally about contraction and diminution, about starting with more and ending with less. Many things may be possible, but only few things end up being realized. With information one needs to start at the top and work down. What is the very top? It is the collection of all possible worlds. The ultimate act of information is therefore to identify the actual world, the world we inhabit, in all of its specificity, to the exclusion of all other worlds. All other acts of information can be viewed as subsidiary—as taking subsets of possible worlds and narrowing them further.[18]

The fundamental intuition of information as narrowing down possibilities matches neatly with the concept of intelligence. The word *intelligence* derives from two Latin words: the preposition *inter*, meaning between, and the verb *legere*, meaning to choose. Intelligence thus, at its most fundamental, signifies the ability to choose between. But when a choice is made, some possibilities are actualized to the exclusion of others, implying a narrowing of possibilities. And so, an act of intelligence is also an act of information.

If we trace the etymology of *intelligent* back still further, the *l-i-g* that appears in it derives from the Indo-European root *l-e-g*. This root appears in the Greek verb *lego*, which by New Testament times meant to speak. Its original Indo-European meaning, however, was to lay, and from there to pick up and put together. Still later, it came to mean to choose and arrange words, and from there to speak. The root *l-e-g* has several variants, appearing as *l-o-g* in *logos* and as *l-e-c* in *intellect* and *select*. This is all just basic etymology, easily confirmed in a good dictionary.

As a side note, this brief etymological tour reveals that Darwin's great coup was to co-opt the term *selection*, previously associated with the conscious choice of purposive agents, and combine it with the term *natural* as a way of negating the role of such intelligent agency. In the term *natural selection*, Darwin therefore attempted to recover all the benefits of choice as traditionally conceived, and yet without requiring the services of a traditional intelligence capable of choosing.

Thus, to this day we read such claims, as by Francisco Ayala, that Darwin's greatest discovery was to give us "design without designer."[19] Likewise, Richard Dawkins describes Darwin's great coup as explaining the appearance of design in the absence of actual design.[20]

Traditionally conceived, choice requires having multiple live possibilities and then actualizing some possibilities to the exclusion of others in order to accomplish an end or purpose. A synonym for the word *choice* is therefore *decision*, with the corresponding verb forms being *choose* and *decide*. The words *decision* and *decide* are likewise from the Latin, combining the preposition *de*, meaning down from, and the verb *caedere*, meaning to cut off or kill (compare our English word *homicide*).

Decisions, in keeping with this etymology, raise up some possibilities by cutting down, or killing off, others. When you decide to marry one person, you cut off all the other people you might marry (assuming traditional one-to-one marital relationships). An act of decision is therefore always a narrowing of possibilities. It is an informational act. But given the definition of intelligence as choosing between, it is also an intelligent act.

Given the etymology of information and intelligence, it's obvious that these two notions are related. It therefore makes sense to measure intelligence via information. This is not to say there might not be other ways to measure intelligence, but on its face using information to measure intelligence should at the very least seem promising. To start the ball rolling, we consider a simple example of intelligence in action and a simple information-theoretic way of measuring its degree. This example suggests that we are on the right track in measuring intelligence via information, but it also raises questions about intelligence metrics that require further investigation.

3. Mac the Canine Door Opener

A friendly white Labrador retriever named Mac belongs to a neighbor of ours. Our subdivision consists of acreages, so this dog occasionally leaves his property and visits ours, where he makes himself at home. The neighborhood is safe and we tend to leave our doors unlocked when we are around. I first met Mac when he showed up inside our laundry room. He had let himself in through the side door to our house.

Mac, it turns out, is able to open doors with lever handles. In this regard, we might say that he is more intelligent than other dogs that have not learned to open doors with lever handles. At the same time, I am able to open not just doors with lever handles, but also doors with knob handles as well as sliding doors—doors Mac is unable to open. My intelligence at opening doors, we might therefore say, exceeds Mac's. But when it comes to door-opening intelligence, locksmiths have the advantage over me, able to open locked doors that I'm unable to unlock. And since no door can stay closed to God, we might say that God has the ultimate door-opening intelligence.

Frivolous as this example may seem, it raises interesting questions about the use of information to measure intelligence. Information, as noted, is about narrowing possibilities. Any formal account of information will therefore require a space of possibilities. Let's denote this space by the capital Greek letter Ω. In this example, we can let Ω consist of all the doors in the world. Consider now the following subsets of Ω (we use CD to denote *Closed Doors*):

- CD_{Mac}: the doors that "Mac" can't open.
- CD_{me}: the doors that I can't open.
- $CD_{locksmiths}$: the doors that locksmiths can't open.
- CD_{God}: the doors that God can't open.

Since God is presumed to be omnipotent, $CD_{God} = \emptyset$. In other words, the set of all doors that God can't open is the empty or null set. Because the null set is contained in every set, the set of doors God can't open is contained in the set of doors that the locksmiths can't open, which in turn is contained in the set of doors that I can't open, which in turn is contained in the set of doors that Mac can't open. In set-theoretic notation:

$$\emptyset = CD_{God} \subset CD_{locksmiths} \subset CD_{me} \subset CD_{Mac}$$

Because information consists of the narrowing of possibilities, it follows that there's more information in CD_{God} than in $CD_{locksmiths}$ than in CD_{me} than in CD_{Mac}. Moreover, CD_{Mac} will be contained in the doors closed to the average dog because, unlike Mac, the average dog won't be able to open lever-handled doors. Consequently, CD_{Mac} will be contained in $CD_{average\ dog}$, implying that the latter exhibits less information than the former. Any information metric will therefore need to represent these inclusions among closed-door sets via the following inequality (I here denotes the information metric):

$$I(CD_{God}) > I(CD_{locksmiths}) > I(CD_{me}) > I(CD_{Mac}) > I(CD_{average\ dog}).$$

In this example, we gauged intelligence in terms of the doors closed to various agents. We represented these closed doors with subsets like CD_{me} and CD_{Mac} of the possibility space Ω. Yet we might also have tried to gauge door-opening intelligence by using subsets like OD_{me} and OD_{Mac} of the possibility space Ω, where OD denotes the set of doors capable of being opened by me, Mac, etc. In that case, the bigger the set OD_{me} compared to OD_{Mac}, the greater my intelligence compared to Mac's in opening doors. But because the prime intuition underlying information is the narrowing or reduction of possibilities, to measure intelligence in terms of information requires, in this case, focusing on sets like CD_{me} and CD_{Mac} rather than their complementary sets OD_{me} and OD_{Mac}.

The difference between the sets CD and OD may seem to be strictly a matter of convenience and convention. But there's something deeper at stake here. We don't measure intelligence by determining where people get everything right but where they get something wrong. Tests where everyone scores 100 percent are meaningless. The point of an intelligence metric is to distinguish among levels of intelligence, and that happens by finding points of failure for one individual that are not points of failure for another. Intelligence is not determined by perfection, as in what people get invariably right. Rather, intelligence is determined by imperfection, as in where some people but not others fall short. The focus on CD_{me} and CD_{Mac} rather than on OD_{me} and OD_{Mac} with respect to door-opening intelligence exemplifies this deeper point.

4. The Connection Between Probability and Information

Typically, an information measure I on a possibility space Ω is defined in terms of a probability measure P on Ω. Thus, for a set E in Ω, $I(E)$ is, by definition, set equal to $-\log_2 P(E)$. An information measure is therefore a negative logarithm to the base two of a probability measure. Information measures, so defined, are therefore disguised probability measures in which the scale and direction of the probability are transformed, resulting in smaller probabilities corresponding to larger information values.

The concept of information just described is Claude Shannon's.[21] Shannon characterized it in the late 1940s in his groundbreaking work on communication theory. Shannon information is the most basic mathematical type of information.

Other types of information have since been developed, such as Kolmogorov information, which we will discuss later in this paper. An averaged version of Shannon information, known as entropy, received in-depth treatment from Shannon, and the theorems he proved about information as applied to communication theory focused on it. As a communications engineer, Shannon focused on entropy because, as an average, it could gauge global effectiveness and reliability of communication systems, ignoring particular features of messages, which are secondary from an engineering viewpoint.

Information as a logarithmic transformation of probability makes independent events, which are multiplicative under probability measures, additive under corresponding information measures because the logarithm of a product is the sum of the logarithm of the factors. Thus, for events E and F that are probabilistically independent (E and F being represented as subsets of Ω), $P(E \cap F) = P(E) \times P(F)$. This is just the definition of what it means for events to be independent. But that in turn is equivalent to $I(E \cap F) = I(E) + I(F)$.

Probabilities are numbers between zero and one. Shannon information measures, by taking a negative logarithm of probability, are therefore numbers between zero (corresponding to a probability of one) and infinity (corresponding to a probability of zero). In consequence, because the probability of the empty set is always zero, it follows that in the earlier example about opening and closing doors, $P(CD_{God}) = P(\emptyset) = 0$ and so $I(CD_{God}) = I(\emptyset) = -\log_2 P(\emptyset) = \infty$ (the probability of the empty set is always 0, and its negative logarithm is therefore ∞).

In general, for a being with no points of failure—as with an omnipotent, omniscient, and omnicompetent God—the information, and therefore intelligence, associated with such a being for the tasks it attempts to complete will be infinite. In general, when the points of failure in the performance of a given task constitute the empty set, the information value associated with its performance will equate to ∞. Granted, this means that anyone capable of performing some trivial task perfectly will be assigned infinite intelligence at the task. Unimpressive as "infinite intelligence" may be in such cases, it still makes good intuitive sense.

The rationale for converting probability measures to information measures via a logarithm to the base two is that it equates probabilities with bits. In many contexts, it is more convenient to work with bits rather than probabilities, bits being the currency of communication and computer technology. To see how this works, imagine we represent coin tosses as 0s and 1s, 0 for tails, 1 for heads. Three coin tosses therefore correspond to three bits. But three coin tosses also correspond to a particular probability. Consider, for instance, three heads in a row —call this the event E (represented as 111). Given the usual assumptions about

coin tosses being equiprobable and independent, we thus calculate the probability of $P(E)$ as equal to $\frac{1}{8}$ ($= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$).

This probability is readily reinterpreted information theoretically. Given that $\log_2 P(E) = -3$, by putting a negative sign in front of the logarithm, we get that $I(E) = -\log_2 P(E) = 3$ bits. In this way, probabilities of coin tosses correspond to the number of times the coin was tossed. This move puts all probabilities in one-to-one correspondence with bit lengths provided we allow for real-valued (and not just integer-valued) bit lengths. For instance, a probability p of $\frac{1}{10}$ corresponds to a bit length of $-\log_2 p = -\log_2 \frac{1}{10} \approx 3.322$ bits. Any probability thus corresponds to a real-valued number of bits.

5. Interpreting Information as an Intelligence Metric

The example of our neighbor dog Mac raises a number of important questions about using information as an intelligence metric. The most obvious question raised is how to interpret the information values such as $I(CD_{Mac})$, $I(CD_{me})$, etc. The rank ordering among these information values is obviously correct: $I(CD_{God}) > I(CD_{locksmiths}) > I(CD_{me}) > I(CD_{Mac}) > I(CD_{average\ dog})$. Indeed, this inequality better hold if information here is measuring intelligence and intelligence here is characterizing the ability to open doors.

But how should we interpret these numbers? If, for instance, $I(CD_{locksmiths})$ is twice that of $I(CD_{me})$, does that mean that locksmiths, in their capacity to open doors, are twice as smart as me? What would such a claim even mean? Information measures, as we've said, are typically disguised probability measures. Any interpretation of such information values as measuring intelligence will therefore depend on the underlying probability measure P over the space of possibilities Ω (in this example, the space of all doors).

But what's P going to be in that case? Is it going to be a uniform probability across all doors, assigning the same small probability to each? That assignment of probability might be appropriate in some circumstances. But without a convincing rationale, that assignment will seem arbitrary. Perhaps the vast majority of doors that I'm unable to open are also beyond the ability of locksmiths to open. If so, $I(CD_{locksmiths})$ and $I(CD_{me})$ may be quite close even though the actual difference in intelligence (as gauged by the ability to open challenging doors) may

by any objective standard be considerable, overwhelmingly favoring locksmiths compared to me.

To the extent that intelligence metrics are grounded in information measures and information measures are disguised probability measures, intelligence metrics will depend on the underlying probability measures. This means that in any valid interpretation of intelligence metrics based on probability measures, we need a convincing rationale for why the underlying probability measures induce metrics that can rightly be regarded as appropriate for assessing intelligence.

Must every intelligence metric be grounded in a probability-based information measure? Such a requirement would unduly restrict intelligence metrics. Rankings of people in competitive activities can be regarded as measuring intelligence at those activities, with a higher rank indicating greater intelligence (such as people ranked in the top five worldwide in tennis or golf or chess).

Certain standardized tests assign numbers that at first must have seemed quite arbitrary, such as 1 through 36 for the ACT and 400 through 1600 for the SAT, with higher numbers signifying greater aptitude or intelligence. Through common use, we now know that a 35 or 36 is a great ACT score and that 1550 or better is a great SAT score (both ranges presumably signifying high intelligence). Such numbers, though bearing on intelligence, have no immediate connection with probabilities. Yet, as we shall see later, the connection to probability here can be made explicit via percentiles.

6. Task Specificity of Intelligence

The next question we turn to is the task specificity of intelligence. In the example of our neighbor dog Mac, we focused on the ability to open doors. In this respect, I have the advantage over Mac, as is reflected in $CD_{me} \subset CD_{Mac}$ or, correspondingly, in $I(CD_{me}) > I(CD_{Mac})$. Thus, with respect to door-opening, I may rightly be regarded as more intelligent than Mac. But door-opening is only one type of ability on which the intelligence of Mac and me might be compared. When it comes to sniffing around to find food, Mac has the advantage over me and other humans in general.

Thus, we can imagine a possibility space with sets NSF_{me} and NSF_{Mac} where *NSF* stands for "non-sniffable food" and the subscript indicates the agent for whom the food is non-sniffable. As usual, we focus on points of failure in

assessing intelligence, with smaller inability sets indicating greater intelligence. In this case, therefore, $NSF_{me} \supset NSF_{Mac}$, implying that $I(NSF_{me}) < I(NSF_{Mac})$, underscoring that Mac is more intelligent than me in sniffing out food. Mac is therefore more intelligent at one task (sniffing out food) but I'm more intelligent at the other task (opening doors).

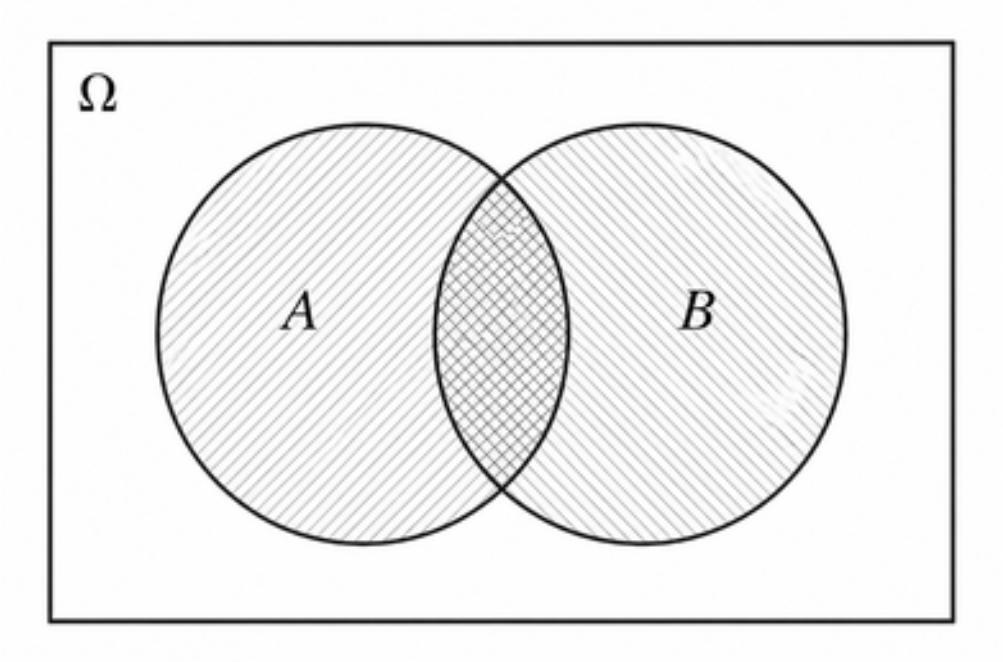
Intelligence in such examples is task specific. But is intelligence always task specific? It is, but this answer requires that we be clear about what the task in question is. We can always expand the range of tasks so that multiple tasks can be included within a single composite task. Thus, we might imagine a composite task that requires both opening doors and sniffing for food. This composite task might be aimed at finding food by opening doors, which would then depend on both abilities. With such a composite task, neither Mac nor I would excel the other in every respect. My ability to smell certain types of food behind doors open to me but closed to Mac may allow me to locate food inaccessible to Mac. On the other hand, Mac's ability to smell food unblocked by any door but also undetectable to my nose would allow Mac to find food inaccessible to me.

7. Dominating versus Non-Dominating Intelligence

Given multiple tasks with separate intelligence metrics, it can happen, as we just saw, that one agent excels another in one task but that the other agent excels the first in a different task. Separate intelligence metrics will then capture this difference. Thus I would score higher than Mac on an intelligence metric for the task of opening doors, and Mac would score higher than me on an intelligence metric for sniffing out food.

But what if we have a single (non-composite) task where one agent is better at some aspects of it, and another agent is better at other aspects of it? In the original Mac door-opening example, that was not the case because in any pairwise comparison of intelligence, one intelligence always strictly *dominated* another in the sense that the one was in every way superior or at least equivalent to the other. Thus, it was assumed I could open all doors that Mac could open, and locksmiths could open all doors that I could open. My intelligence in opening doors therefore dominated Mac's, and the locksmiths' intelligence in opening doors dominated mine.

But what if, instead, I were better at opening some doors than Mac, and Mac were better at opening other doors than me—perhaps leather doors closed to me but that he could successfully chew through (assume Mac and I are operating on our own with no special tools)? In that case, intelligence would be *non-dominating*. To see what non-dominating intelligence can look like, suppose that Ω is the underlying possibility space and that A and B are subsets of Ω representing the inability sets of two agents (i.e., the respective possibilities where the ability of the two agents breaks down). Suppose further that A and B intersect and that neither includes the other, as in the following Venn diagram:



Such non-domination of ability adds a wrinkle to interpreting the information values $I(A)$ and $I(B)$ as measures of intelligence. Perhaps $I(A)$ is much bigger than $I(B)$, suggesting that the agent associated with A is, on balance, much better at the task in question, and thus much more intelligent at performing the task, than the agent associated with B. And yet, it could be that for certain outliers or anomalies, the agent associated with B might have the advantage. In such cases, there will be possibilities where each agent outperforms the other.

Any interpretation of $I(A)$ and $I(B)$ as intelligence metrics will thus need to factor in the degree to which the agents whose intelligence is being measured at once outperform and are also outperformed by other agents. If agents outperform other agents on some possibilities but are outperformed by them on others, then

the underlying intelligence metric will be non-dominating, and this fact will need to be factored into how to interpret the values delivered by the intelligence metric.

8. Conditional versus Unconditional Intelligence Metrics

The information-based intelligence metrics we've been considering focus on where an agent's ability at a task breaks down. Thus, in the Mac example, Mac's intelligence was captured in the subset CD_{Mac} of the possibility space Ω. CD_{Mac} is an inability set, and the information value $I(CD_{Mac})$ captures numerically those possibilities that demonstrate inability. The smaller the range of possibilities where a given ability breaks down (here CD_{Mac}), the greater the information (here $I(CD_{Mac})$) and thus the higher the intelligence.

Such an intelligence metric applies to the entire possibility space Ω. It is unconditional in the technical sense of probability and information theory: the probabilities are evaluated directly on subsets of Ω one at a time rather than two at a time, as when one subset is conditioned on another. Given two subsets E and F of Ω, the conditional probability $P(E \mid F)$ is defined as $P(E \cap F)/P(F)$. Applying negative logarithms to the base two now yields a corresponding definition of conditional information: $I(E \mid F) = I(E \cap F) - I(F)$ (because division turns to subtraction under logarithmic transformation).

The point of conditional probability is to assess probability in light of a reduction of the possibility space based on new knowledge. That reduction represents a narrowing of possibilities brought on by the new knowledge, and so we typically refer to that new knowledge as new information (often treated as evidence). Conditional probability is all about changing the underlying probability structure based on new information. Conditional probability is the mechanism by which probability theory accommodates new evidence/knowledge.

To see in practical terms how conditional probability works, take Ω to consist of all sequences of the 26 alphabetical characters. Imagine one now learns that the key to the letter e was not working on a keyboard that's generating characters. Now the space of relevant alphabetical characters needs to be redefined as all sequences of 25 alphabetical characters excluding e. Call this reduced possibility space F. For a subset E of Ω, its probability, in light of the new knowledge about the letter e being unavailable, is then $P(E \mid F)$.

But it could also be that we lack such additional knowledge. Thus, in place of F, which reduces the possibility space Ω, we may simply need to work with the original probability space Ω. In that case, for a subset E of Ω, we'll simply want to calculate $P(E)$ and $I(E)$, as before. But note, these unconditional probability and information values are consistent with conditional probability and information values where the set being conditioned upon is the entire space Ω. That's because $P(E \mid \Omega) = P(E \cap \Omega)/P(\Omega) = P(E)$. This equality follows from the axioms of probability: the probability of the entire possibility space Ω is 1 (i.e., $P(\Omega) = 1$) and for any subset E of Ω, $E \cap \Omega = E$, making $P(E \cap \Omega)$ equal to $P(E)$. Accordingly, $P(E \mid \Omega) = P(E \cap \Omega)/P(\Omega) = P(E)/1 = P(E)$ and so $I(E \mid \Omega) = I(E)$.

To use unconditional information to gauge intelligence is to compare intelligence in an across-the-board way that doesn't privilege any aspects of the possibility space Ω. Unconditional information thus renders all the information values $I(CD_{God})$, $I(CD_{locksmiths})$, $I(CD_{me})$, $I(CD_{Mac})$, etc. as comparable to each other in gauging intelligence, whatever the actual numbers turn out to be. That still leaves the question of interpreting the numbers, but at least we have comparability. The intelligence so measured is unconditional.

But we could also consider the conditional intelligence between two agents. For instance, we might not care about the door-opening ability of a rabbit or a slug, but we might want to know the relative door-opening ability (and thus intelligence) of Mac and me. Thus, we might want to know how much more intelligent I am at opening doors than Mac (confined just to this task or activity). In information-theory, such a relativized measure would be denoted, as defined above, by the conditional information $I(CD_{me} \mid CD_{Mac}) = -\log_2 P(CD_{me} \mid CD_{Mac})$.

But note, we are assuming in this example that intelligence is dominating, and so in particular that $CD_{me} \subset CD_{Mac}$. Thus $I(CD_{me} \mid CD_{Mac}) = I(CD_{me}) - I(CD_{Mac})$, which will be strictly less than $I(CD_{me})$, which measures my intelligence at opening doors irrespective of any particular agent with which my door-opening intelligence might be compared. Compared to Mac, my door-opening intelligence is less than my door-opening ability compared across the board, which will include the doors closed to rabbits and slugs as well as locksmiths and gods. My relative intelligence compared to Mac's and as measured by $I(CD_{me} \mid CD_{Mac})$ thus shows me to have the advantage in intelligence.

Yet we can also flip the conditional information here. From $CD_{me} \subset CD_{Mac}$ it follows that $P(CD_{Mac} \mid CD_{me}) = 1$ and so $I(CD_{Mac} \mid CD_{me}) = 0$. Thus, even though $I(CD_{me} \mid CD_{Mac})$ can capture my relative greater intelligence than Mac's, $I(CD_{Mac} \mid CD_{me})$, being zero, will also capture that Mac's intelligence in opening

doors has no advantage over mine. But note, $I(CD_{Mac} \mid CD_{me}) = 0$ tells us nothing about the degree of Mac's intelligence at opening doors in relation to mine. In that regard, $I(CD_{Mac})$, the unconditioned information, as related to $I(CD_{me})$, is more revealing. That said, $I(CD_{Mac})$ tells us nothing about the way Mac is less intelligent at opening doors compared to me.

Conditional intelligence metrics can also arise with non-dominating intelligence. We find such non-dominating intelligence occasionally in competitive games where one player is, objectively speaking, superior to another, and yet because of a mental block or other idiosyncrasy succumbs under certain circumstances to the other player.[22] In such cases, because neither A nor B is included in the other, $I(A|B)$ and $I(B|A)$ will signify non-trivial relative intelligence advantages of the one agent over the other.

9. Absolute versus Probabilistic Intelligent Performance

Intelligence, as noted, can be treated both as a matter of kind (all-or-nothing) and as a matter of degree (allowing gradations). Accordingly, we can ask whether an intelligence was involved; but we can also ask how much intelligence was involved. So far, however, we've treated the intelligence of agents as succeeding or failing at a task with all-or-nothing certainty. In particular, we've only considered the case where for any possibility in the possibility space Ω, agents will either with certainty succeed at it or with certainty fail at it.

The success or failure of intelligent performance has therefore till now been treated as absolute rather than probabilistic. Our focus on door opening has played into this absolutist approach to intelligent performance. Doors, after all, call to mind only two options, being open or being shut. Thus, the inability sets CD_{me} and CD_{Mac} that we used to gauge the door-opening intelligence of me and Mac simply presupposed that for any door it would definitely open or definitely stay closed to either of our door-opening efforts.

What we didn't consider is whether a door's opening to me or Mac might be probabilistic. What if 99 percent of the time a given door opened to me and 1 percent of the time the same door opened to Mac? My ability to open that door would therefore be superior to Mac's. But that superiority would be probabilistic rather than absolute. Consequently, it could not be adequately represented by

inclusion or exclusion in the sets CD_{me} and CD_{Mac}, nor in the complementary sets OD_{me} and OD_{Mac}, these being respectively the doors I and Mac would be guaranteed to open.

Instead of inability and ability sets such as CD_{Mac} and OD_{Mac}, what we need is a function from the possibility space Ω to the unit interval [0,1] that for each door assigns the probability of Mac, or whoever, opening it. Thus, in place of CD_{me} and CD_{Mac}, we need *performance functions* f_{me} and f_{Mac} that map from Ω to [0,1], assigning to each door in Ω the probability that I or Mac will, respectively, open it. Corresponding to these performance functions are *failure functions* that flip the probabilities of the performance functions. We use a superscripted c (for complementation) to denote these failure functions: f^c_{me} $(= 1 - f_{me})$ and f^c_{Mac} $(= 1 - f_{Mac})$.

Note that CD_{me} and CD_{Mac} can be treated as a special case of this more general scheme by associating with those sets a failure function that assigns 1 to doors that stay closed and 0 to those doors that open (which corresponds to performance functions that assign 0 to the doors that stay closed and 1 to those that open). Mathematicians call such zero-one valued functions *indicator functions* (or *characteristic functions*) and denote them by 1_A for an arbitrary set A ($1_A(x) = 1$ for $x \in A$, and $1_A(x) = 0$ for $x \notin A$). Thus, in the Mac example, the set CD_{me} would correspond to the indicator/failure function $1_{CD_{me}}$, and CD_{Mac} would correspond to the indicator/failure function $1_{CD_{Mac}}$. We will lay out the mathematics of such performance and failure functions in section 11.

10. Search as Key to Intelligent Performance

If a human powerlifter can deadlift 800 pounds, that indicates a high level of performance in the ability to move heavy weight. But most of us would be uninclined to say that this level of performance indicates a high level of intelligence, except perhaps in the preparatory work by the powerlifter in planning and executing a training regimen that strengthens him sufficiently to move such a massive amount of weight.

Thus far, we've described intelligence in terms of ability/performance (or inability/failure). For instance, we've used inability sets within the possibility space Ω to define information measures that assess intelligence (and in the next section

we'll be extending this definition to include performance/failure functions). Yet it seems that performance in general need not count as intelligent performance. Deadlifting 800 pounds, though impressive, seems not to deserve to be counted as intelligent. Yet if someone can prove a complicated mathematical theorem, that, it seems, should count as intelligent.

So what is the difference in such cases? Even though there may be different ways to answer this question, a promising answer that's congruent with our general intuitions about intelligence is *search*. A powerlifter in deadlifting 800 pounds is engaged in search at best peripherally, as in finding the bar with the weight to lift (the search for which is trivial). But consider the following widely recognized acts of intelligence where in each case search seems intrinsic to the act of intelligence:

1. In opening a door, as with Mac, an agent is searching for the right method to open it, which may be quite complicated if, for instance, the door has a combination lock that opens only to the right combination.
2. In taking an IQ or aptitude test, the test-taker is searching for the right answers.
3. In putting together a puzzle, the puzzle solver is searching for the right way to join together all the pieces.
4. In making a violin, the violin maker is searching for the right raw materials, how to prepare them, and then how to assemble them.
5. In proving a complicated mathematical theorem, the mathematician is searching for the right steps in the proof.
6. In trying to sink a golf ball, the golfer is searching for the right way to connect the club with the ball so that it will have the greatest probability of immediately, or mediately, getting into the hole.
7. In arguing a case before a court, an attorney is searching for the right words, tone of voice, and body language with which to convince a judge and jury.

Is search coterminous with everything we might want to include under intelligent activity? Are there successful searches that we would not want to count as intelligent acts? Are there intelligent acts that we would not want to count as the outcome of search? I want to urge that the two are coterminous at the level of representation: intelligent performance can be represented or modeled as successful search without implying that every intelligence internally operates by a search process.[23] Of course, search needs to avoid a retrospective fallacy in which

any physical history can be redescribed as a successful search for whatever state it happened to reach. The success criterion must be specified independently of the realized outcome, a point made explicit in the formulation of specified complexity (see section 22).

To illustrate the generality of search for intelligent performance, consider an MLB batter who consistently bats above .333—which would be an impressive batting average and is Hall of Fame caliber. Is his performance to be regarded as merely skillful but not as intelligent because we tend not to use the word intelligent to describe the skilled performances of athletes?

Yet such a skill seems to align quite nicely with our definition of intelligence as the ability to choose between: this batter has a finely honed ability to choose among bat swings those that will lead to a high percentage of hits. Granted, the actual choices made at the plate in swinging at pitches will be only partially under conscious control, most of the ability being unconscious and automatic in the form of muscle memory and learned hand-eye coordination. In any case, our batter with the .333 average will have a greater success rate than most in successfully finding, or searching for, how to connect with a pitched ball to get a hit.

Other sports confront similar search-like abilities. A basketball player must be able to search among different ways of shooting the ball to increase the probability of making a basket. A quarterback must be able to search among different ways to throw the football effectively to a receiver. A hockey player must be able to search among different ways to hitting the puck with a stick at the goal, which will include choosing among different types of shots, such as a slap shot, wrist shot, snap shot, and backhand shot. Idiosyncratic as this language of search may sound when applied to sports, successful search nonetheless is presupposed in athletic skill.

Skill at competitive mental games, such as chess, likewise presupposes successful search. The chess player is at each move in the game searching for a strong (ideally the strongest) move to play. At the level of the entire game, the chess player is searching for a sequence of moves, adapted move by move to the opponent's play, that will bring about victory. And at the level of the chess player building skills to enhance one's game, the chess player is searching for and attempting to engage in an effective training regimen to build those skills. Search thus enters at every level of competitive mental games. Moreover, on the last point regarding finding and engaging in an effective training regimen, competitive mental games intersect with the process of learning.

But note, learning itself can be conceived in terms of search. As Osherson et al. put it in *Systems That Learn*:

> Learning typically involves 1. a learner, 2. a thing to be learned, 3. an environment in which the thing to be learned is exhibited to the learner, [and] 4. the hypotheses that occur to the learner about the thing to be learned on the basis of the environment. Learning is said to be successful in a given environment if the learner's hypotheses about the thing to be learned eventually become stable and accurate.[24]

Obviously, these stable and accurate hypotheses form the target of such a "learning search." Moreover, the intelligence of the learner is then gauged by whether the learner's hypotheses about the thing to be learned do indeed become stable and accurate and also by how quickly the learner's hypotheses converge on those that are stable and accurate. For human learners, it matters greatly whether learning happens in minutes, hours, days, weeks, months, years, or decades. Greater intelligence here is correlated with speed of learning.

Problem solving can likewise be understood as a form of search. A problem presents an agent with a range of possible solutions, only some of which actually work. Those that work form a target, and solving the problem consists in locating that target within a space of possibilities. A problem-solving search may depend on resources that help the problem solver navigate the possibility space. Background knowledge, representations of the problem, partial solutions to it, computational capacity, etc. may all facilitate problem solving ability. Not all resources, however, need be productive. Misleading information or too much information (making it difficult to sort through) may undermine finding the problem's solution. Ideally, resources reduce the number of possibilities to be considered, identify promising paths, rule out dead ends, or make otherwise inaccessible solutions attainable. Measuring the intelligence displayed in problem solving therefore depends not only on whether the target is reached, but also on how effectively available resources are exploited in reaching it.

In general, it's hard to see how anything that we might regard as intelligent could fail to fall under search. Intelligence, as we have stressed, always involves a choice—or at least a contingency—among possibilities, which is always an informational act. And any informational act by definition involves a narrowing of possibilities. And any narrowing of possibilities involves, if only tacitly, locating those possibilities (by searching for them) to the exclusion of others. It therefore makes sense to broaden under the umbrella of search all that we mean by intelligent performance.

But is search broad enough to include natural processes and is making search the paradigm for intelligence a legitimate way to measure the intelligence of at

least some natural processes, such as evolution by natural selection? Search need not presuppose a conscious agent pursuing a consciously represented purpose. The theory of search developed in this paper requires a space of possibilities, some process that samples that space, and some subset of outcomes that counts as success. Nothing in this formal structure requires successful outcomes to be specified in advance by the process itself.

A target may be imposed prospectively, as when a mathematician searches for a proof, but it may also be identified retrospectively (such as functionally), as when a process arrives at configurations capable of replication, metabolism, flight, or some specified level of catalytic activity. In such cases the target is the subset of the possibility space exhibiting the property in question. Daniel Dennett therefore, without attributing foresight to it, described natural selection as an "algorithmic search processes," and Stuart Kauffman similarly treated mutation, recombination, and selection as searching a fitness landscape.[25]

A biological target need not coincide with a particular nucleotide sequence, a preordained organism, or a unique global optimum, as though the evolutionary process somehow incorporated a precise representation of them and were specifically aiming at them. A biological target may instead consist of all configurations that exhibit a certain degree of function or fitness. Robert Hazen and his colleagues take this approach to targets in defining *functional information*, which, for a specified degree of function, is the fraction of all possible configurations that achieve at least that degree of function.[26] This is a probability and easily converted to an intelligence metric. These functional configurations clearly specify a target. Yet neither Hazen nor his colleagues claim that evolution intended or foresaw the targets exhibiting functional information.

H. Allen Orr's objection that Darwinian evolution is not a search because it has no "pre-set target" assumes too narrow a view of targets.[27] Nothing in the present theory requires that a target must be an exact outcome fixed in advance or that an evolutionary process that attains it must be inherently teleological, somehow internally representing it as a goal. That would be orthogenesis, the view that organisms possess an innate drive to evolve in specific directions. Targets as defined in the present theory place no restrictions on the evolutionary processes that might attain them.

That said, even without pre-set targets, any evolutionary theory worth its salt must account for how populations come to occupy biologically significant regions of configuration space. These would be regions corresponding to replication, metabolism, gene expression, protein function, adaptive performance generally, and other biological structures and functions whose attainment can be assessed

probabilistically, thereby becoming grist for an intelligence metric. To call these regions targets leaves open whether they arose through natural selection, self-organization, genetic drift, symbiogenesis, natural genetic engineering, epigenetic inheritance, intelligent agency, or some combination of these and possibly other causes. The theory developed here is metaphysically minimalist. It imposes no restrictions on scientific inquiry. But it does require evolutionary search processes to prove their intelligence.

In closing this section, I offer an important caveat: We need always to remind ourselves of the pitfalls that may lead us mistakenly to credit an agent (or process) with intelligence simply because it has successfully located a target, and even if the target is intrinsically difficult to attain. When an agent locates a difficult target, we are naturally inclined to credit the agent with intelligence. But the success must result from the agent's own causal powers. It's not enough for the agent simply to have gotten lucky. A winning lottery ticket may hit a fantastically small target, but the winner obviously should not be credited with "lottery-solving intelligence." We should similarly rule out the file drawer effect, as when someone tries repeatedly to perform a task and ignores all failures until success is reached (the failures ending up in a "file drawer"). The reported success may conceal thousands of unsuccessful attempts, making a chance hit look like an impressive search.

The target itself needs to be independently given rather than identified only after it was reached. An otherwise ordinary outcome should not become extraordinary only in retrospect. To be legitimately credited with intelligence, the agent must not enjoy undisclosed advantages. These could include a favorable starting point, inside information, a copied solution, outside assistance, extra time, or access to computational resources. Such advantages can reduce the difficulty of the task and thus diminish the degree of intelligence we would otherwise credit to the agent.

Bottom line: Intelligence should be credited to an agent or process only in the absence of unfair privileges or advantages contributing to its success and only to the degree that its own capacities account for reaching the target.

11. Performance and Failure Functions

We introduced performance and failure functions in section 9 to make sense of performance that is probabilistic rather than absolute. Absolute performance is all-or-nothing: for a given possibility, success or failure is guaranteed. But it could also happen that for a given possibility, success or failure is to varying degrees uncertain. We therefore distinguished between probabilistic and absolute performance in the Mac example, contrasting doors that open with probability either one or zero versus those that open with probability anywhere between one and zero. In this section, we'll provide a formal account of probabilistic performance, laying it out in the context of a test-taking example.

Consider a short math test. Let's say it consists of five questions of varying difficulty, and the test-taker has one minute per question to solve it and then must go on to the next question. Alice is now going to take the test, and let's assume that for each question, she has probability p_n of correctly answering it ($1 \leq n \leq 5$). The possibility space Ω then consists of these five questions, and Alice's performance function f_{Alice} then assigns to the n^{th} problem in Ω the probability p_n of her answering this question correctly. Suppose Bob is likewise going to take this test, and his performance function f_{Bob} assigns to the n^{th} problem in Ω the probability q_n of his answering this question correctly.

Performance functions can be deterministic in the sense that all the probabilities p_n and q_n are either zero or one. In that case, the performance functions f_{Alice} and f_{Bob} would be indicator functions of the form 1_A and 1_B where A comprises the questions on the test that Alice can with certainty answer correctly and B the questions that Bob can with certainty answer correctly—the probability of any question being answered correctly or incorrectly thus being zero or one.

The underlying mathematics here is instructive, especially in the way it extends from indicator functions (which only take on the values zero and one) to performance functions in general (which can take on any probability values). Given a probability measure P on Ω and subsets A and B of Ω, the unconditional probability of A is by definition $P(A)$ and the conditional probability of A given B is by definition $P(A \mid B) = P(A \cap B)/P(B)$. Given that A and B can be respectively represented by indicator functions 1_A and 1_B, it follows that

$P(A) = \int_\Omega 1_A dP$, and that
$P(A \mid B) = \left[\int_\Omega 1_{A\cap B} dP\right] / [\int_\Omega 1_B dP] = \left[\int_\Omega 1_A \cdot 1_B dP\right] / [\int_\Omega 1_B dP] inist.$

The last equation holds because the indicator function $1_{A\cap B}$ for the intersection of A and B is equal to the product of the indicator functions, namely, $1_A \cdot 1_B$ (this product is nonzero only where A and B are both nonzero, which is to say at their intersection). These equations suggest a straightforward generalization for performance functions f and g (such as f_{Alice} and f_{Bob}) from Ω to [0,1]. Thus, in direct imitation of what we did with indicator functions, we define the unconditional probability of f and the conditional probability of f given g as follows:

$P(f) = \int_\Omega f dP$, and that
$P(f \mid g) = \left[\int_\Omega f \cdot g dP\right] / [\int_\Omega g dP] = \left[\int_\Omega f \cdot \overline{g} dP\right]$

where $\overline{g}$ is the probability density function associated with g, which is to say,

$\overline{g} = g/[\int_\Omega g dP]$.

Because performance functions are nonnegative, $\overline{g}$ is likewise nonnegative. Moreover, dividing by the integral of g (we assume this integral is not zero) normalizes $\overline{g}$, ensuring that its integral is 1, thereby making $\overline{g}$ a probability density function. This means that the conditional probability for f given g is the integral of f with respect to the probability density $\overline{g}$. Corresponding information measures are thus likewise defined straightforwardly:

$I(f) = -\log_2 P(f)$, and
$I(f \mid g) = -\log_2 P(f \mid g)$.

Let's now apply these probabilistic and information-theoretic definitions to the example of Alice and Bob taking that math test. Assume that the probability measure P on Ω is the uniform probability, and so assigns probability 1/5 to each question on the test. This makes sense if, as we'll assume, all the questions count equally in scoring the test. Let's also assume that the questions get progressively more difficult for anyone who takes the test.

Assume now we find that for question n, Alice has probability $p_n = 1/n$ of correctly answering it (for $1 \leq n \leq 5$). Thus Alice answers these questions correctly with probabilities 1, 1/2, 1/3, 1/4, and 1/5, or in decimal terms, 1, .5, .333 (approximately), .25, and .2. Let's say that Bob is a lot worse on this test than Alice. Specifically, Bob has probability $q_n = 1/10^n$ of correctly answering it (for $1 \leq n \leq 5$). Thus Bob answers these questions correctly with probabilities .1, .01, .001, .0001, and .00001.

Next let's calculate the probability and information values associated with f_{Alice} and f_{Bob}. The calculations are straightforward, with the integrals above

becoming weighted summations from 1 to 5 of probabilities or products of probabilities, each multiplied by a weight of 1/5. For ease of notation, we'll let f denote f_{Alice} and g denote f_{Bob}:

$P(f) = .4567$ and $I(f) = 1.1308$
$P(g) = .0222$ and $I(g) = 5.4919$
$P(f \mid g) = .9483$ and $I(f \mid g) = .0765$
$P(g \mid f) = .0462$ and $I(g \mid f) = 4.4347$

These calculated values, however, fail to measure the intelligence of Alice and Bob or their relative intelligence. The problem is that f and g, as performance functions, characterize the probabilities of Alice and Bob in *successfully* answering the math test questions used to gauge their mathematical intelligence. But as we noted in the Mac example, intelligence is better gauged informationally not by ability but by inability, which is why to assess Mac's door opening intelligence, we focused not on OD_{Mac}, the doors Mac can open, but on the complementary set CD_{Mac}, the doors Mac can't open.

In general, for a subset A of the possibility space Ω, its complement is defined as $A^c = \{x \in \Omega \mid x \notin A\}$. For the indicator function 1_A, the indicator function of the complement, namely 1_{A^c}, can then be expressed in terms of the original indicator function by subtracting it from 1 (where 1 is now treated not as the number 1 but as the function that takes on the value 1 everywhere on Ω). In other words, $1_{A^c} = 1 - 1_A$. It follows that $P(A^c) = P(1_{A^c}) = P(1) - P(1_A) = 1 - P(A)$, as is always the case with complementation of sets since the probability of a complement is one minus the probability of the original set.

Subtracting an indicator function from 1 to get the indicator function of the complement suggests a natural way to form complements of performance functions. Thus, for performance functions f and g, we define the complementary performance functions f^c and g^c respectively as $1 - f$ and $1 - g$. We call these complementary performance functions *failure functions*.

Subtracting a performance function from 1 yields a performance function in the technical sense in that it too is a function from the possibility space Ω to the unit interval [0,1]. But with these complementary performance functions, what we mean by performance is flipped. In fact, they might rightly be regarded as anti-performance functions (corresponding to the inability sets that we used to assess intelligence in the Mac example). Indeed, this is why we call them failure functions.

These failure functions are the key to assessing the intelligence of Alice and Bob in this math test example. Specifically, $f^c = 1 - f$ $(= 1 - f_{Alice})$ gives the

probability that Alice will fail to answer each math test question correctly and likewise $g^c = 1 - g$ $(= 1 - f_{Bob})$ that Bob will fail to answer each correctly. The smaller the probabilities (and thus correspondingly the greater the information values) associated with these anti-performance functions, the greater the intelligence. Here are the probability and information values associated with these failure functions (exactly in parallel with the values we calculated for the original performance functions):

$P(f^c) = .5433$ and $I(f^c) = .8802$
$P(g^c) = .9778$ and $I(g^c) = .0324$
$P(f^c|\ g^c) = .5545$ and $I(f^c|\ g^c) = .8507$
$P(g^c|\ f^c) = .9979$ and $I(g^c|\ f^c) = .00297$

Even if an exact interpretation of these numbers requires more context, the numbers themselves make sense, having the right relative sizes. Alice is clearly more intelligent than Bob at this test, which is reflected in the unconditional information values satisfying $I(f^c) > I(g^c)$, indicating that Alice is in absolute terms better at this test than Bob. But it is also reflected in the conditional information values satisfying $I(f^c|\ g^c) > I(g^c|\ f^c)$, indicating that Alice is, in a head-to-head comparison, better at this test than Bob.

Note that the information values calculated here are not very high because they are based on probabilities that are not very low. The fact is, neither Alice nor Bob is anywhere near "acing" this math test. If, for instance, Carol had a probability of .9999 of correctly answering each of the five questions—and thus was likely to ace it—her performance function $h = h_{Carol}$ would yield a corresponding failure function h^c that equals .0001 for each test question. Because h^c is a constant function, $P(h^c) = P(h^c \mid f^c) = P(h^c \mid g^c) = .0001$, and so $I(h^c) = I(h^c \mid f^c) = I(h^c \mid g^c) = 13.288$, which is much larger and more impressive than the information/intelligence metric values calculated for Alice and Bob.

One way to appreciate more clearly the difference in intelligence between Alice, Bob, and Carol in this example is to take ratios of the corresponding unconditioned information values. Consider, therefore, the following ratios:

$I(f^c)/I(g^c) = .8802/.0324 = 27.167$
$I(h^c)/I(f^c) = 13.288/.8802 = 15.097$
$I(h^c)/I(g^c) = 13.288/.0324 = 410.123$

These numbers indicate that Alice's performance on this math test is over 27 times more information intensive than Bob's and that Carol's performance is over 15 times more information intensive than Alice's. Moreover, when pitting the

most intelligent on this test (Carol) with the least intelligent (Bob), we find that Carol's performance is over 410 times more information intensive than Bob's.

Such ratios therefore help us to understand the relative intelligence of agents even if in absolute terms the amount of information calculated may not seem all that impressive. Alice, for instance, does not seem particularly good at this math test. Given her probabilities of correctly answering the test's five questions (1, 1/2, 1/3, 1/4, and 1/5), her expected percentage score on the test is 46 percent, which in most US classrooms would be considered failing (60 percent is typically needed for a D–). But relative to Bob, Alice is a genius on this test. And Carol, relative to Alice, is likewise a genius on this test.

To the extent that these information-based intelligence metrics become more prevalent and widely used, researchers are likely to become more adept at making sense of these numbers and what they convey intuitively. In the present example, Carol's high information score suggests that she is really good at correctly answering the questions of this test. But her actual intelligence at math when considered more broadly (i.e., not limited to a five-question test) seems unlikely to be captured by this information score. In fact, if the test is too easy, it may tell us virtually nothing about Carol's mathematical intelligence.

For another test that's far more difficult, a much lower information score may indicate much greater mathematical intelligence. For instance, in the Putnam math exam, taken by US undergraduates, the median score is zero—in other words, half the students get no credit on any of its questions.[28] The performance function of most undergraduates on the Putnam exam will be uniformly low (close to zero, and for at least half it is equal to zero), meaning that the corresponding failure function will be uniformly high (close to one), which in turn will correspond to extremely low information values (close to zero bits), as with Bob in the preceding example.

As a postscript to this section, it's worth distinguishing performance functions from fuzzy sets. For traditional sets, inclusion is all-or-nothing: a thing is either in the set or not in the set. With fuzzy sets, inclusion comes in degrees.[29] Fuzzy sets come up in contexts where inclusion is hard to justify in all-or-nothing terms. Consider the "set" of all bald men. Someone without a hair on his head is clearly bald and will belong to such a set. But what about a man with a single strand of hair on his head? Clearly, we would want to consider that person to be bald as well. But keep adding hairs, and eventually, it's not so clear whether a person should be considered bald. Fuzzy sets capture this type of fuzzy or vague inclusion.

Performance functions, as developed here, are mathematically equivalent to type-1 fuzzy sets.[30] The mathematics of the two are identical, with both perfor-

mance functions and fuzzy sets being functions from some space Ω into the unit interval [0,1]. Nonetheless, performance functions are interpreted differently from fuzzy sets, the point at issue being probability of successful performance rather than fuzzy inclusion. Hence the difference in terminology even though the equivalence of the mathematics.

12. Cumulative Performance Functions

With the performance functions considered in the last section, the possibility space consisted of individual tasks and the performance functions assigned to each task the probability of an agent successfully completing it. Often, however, success depends on the completion of multiple tasks taken jointly where performance on each individual task may carry little to no weight. A baseball player may swing wildly at a first pitch but still get a hit. In that case, the missed swing carries no weight. In a multiple-choice test of 100 questions where each question counts the same and 90 or above is an A, a single missed question carries some weight but not much. In such situations, what matters is cumulative performance across multiple individual tasks that jointly make up a composite task. Cumulative performance on the composite task constitutes the criterion of success.

Such cumulative performance is gauged by cumulative performance measures, and these behave differently from the performance measures considered in the previous section. For non-cumulative performance functions, the possibility space Ω consists of individual tasks, and the performance function simply assigns a probability of success to each task. But for cumulative performance functions, the possibility space Ω is structured differently. It consists of individual tasks that jointly make up a composite task. Performance on the composite task will then signify degrees of success. Moreover, each of the individual tasks making up the composite task will exhibit dependencies so that a high probability of success on one composite task may entail a still higher probability of success on a related composite task.

Consider, for instance, a multiple-choice test with 100 questions. We could, as in the previous section, set up a possibility space so that it simply contains each question and where for each question we assign a probability that a given agent will answer it correctly. Performance across such questions, even though there

are more than one, would then be treated unitarily as a non-composite task. The task, in this case, would be to answer all 100 questions, treated as a unit.

But in general with such tests, we are interested in the total number of questions answered correctly, so our possibility space Ω could also look as follows: $\Omega = \{0, 1, 2,..., 99, 100\}$. Each number in Ω here will then represent not the number of a question on the test but rather a possible score on the test, cumulated across all the test questions, with 100 being best.

A performance function f will now assign to each number n in Ω a probability p_n, which as in the last section we write $f(n) = p_n$. But note, p_n must not denote the probability of scoring exactly n on the multiple-choice test, otherwise the intelligence metric constructed from these probabilities will be misleading and useless. To see this, consider a genius test-taker who with probability 1 will ace the test, scoring 100 percent. For any n less than 100, p_n would then equal 0, ensuring that the associated failure function f^c would be 1 everywhere except at 100, implying that the information associated with this failure function would be close to 0 (i.e., $P(f^c)$ would be very close to 1, making $I(f^c)$ very close to 0). Low information of failure functions indicates high probability of failure, in turn indicating low, not high, intelligence. Yet the test-taker in question here is as intelligent as could be on this test.

The way around this problem is to interpret the elements of $\Omega = \{0, 1, 2,..., 99, 100\}$ cumulatively. Thus the number 98 in this possibility space would be interpreted as a score of 98 or better. A 65 would be interpreted as a score of 65 or better. Etc. Given this interpretation of the elements of Ω, it follows that for $m < n$, $p_m = f(m) \geq f(n) = p_n$. Thus for the test-taker guaranteed to ace this test, not only does p_{100} equal 1, but for all n less than 100, p_n equals 1. In that case, the associated failure function f^c would equal 0 everywhere, entailing that the associated information value would equal ∞. On this test, the test-taker would have zero probability of failure and thus exhibit infinite intelligence. Of course, the usual caveat applies: this infinity is ascribed to perfect performance independently of intrinsic difficulty of the test or impressiveness in performing well on it.

Let's put some more concrete numbers on this test-taking example. Assume that Brian is a B student and Carla is a C student, and that on a 100-question test, Brian is guaranteed to score in the range 80 to 89 and Carla to score in the range 70 to 79. Granted, these assumptions are artificial, but let's go with them for the purposes of illustration. Let f_B denote Brian's performance function and f_C denote Carla's performance function. It follows that f_B will be 1 on the values 0 to 80 and that f_C will be 1 on the values 0 to 70. Moreover, f_B will decline to

0 on the values 80 to 90 (after which it will remain 0), and f_C will decline to 0 on the values 70 to 80 (after which it will remain 0).

The failure functions f_B^c and f_C^c will then look as follows: f_B^c will equal 0 from 0 to 80, go up by increments of 1/10 from 81 to 90 (let's assume these increments to keep things simple), and then be 1 from 90 to 100. Similarly, f_C^c will equal 0 from 0 to 70, go up by increments of 1/10 from 71 to 80 (let's likewise assume these increments to keep things simple), and then be 1 from 80 to 100. Let's also assume (again to keep things simple) that $\Omega = \{0, 1, 2,..., 99, 100\}$, the space of test outcomes, has a uniform probability P (note that the underlying space Ω here has 101 elements). It then follows, by doing the math, that $I(f_B^c) \approx 2.704$ and $I(f_C^c) \approx 1.986$, implying that $I(f_B^c) > I(f_C^c)$, as it should be since Brian is, objectively, more intelligent on this test than Carla.

But note that the actual difference between these two information values does not seem numerically significant (merely around .7). Even the ratio of these information values, which we found in the last section to be diagnostic of significant intelligence differences, fails to underscore such a difference here: $I(f_B^c)/I(f_C^c) \approx 2.704/1.986 \approx 1.362$. Granted, as a B student, Bob is not a towering genius compared to Carla, who is a C student. Yet in terms of class grades, the difference between a B and a C is significant, and so these differences in information values and ratios of them is likewise significant.

Even an A student may not seem all that distinguished compared to B and C students. Suppose David is an A student with failure function f_D^c equaling 0 from 0 to 90 and going up by increments of 1/10 from 91 to 100. In that case, $I(f_D^c) \approx 4.199$. Compared to the previous information values, David's information is not vastly greater. And yet, as an A student, he will be regarded as a considerably stronger student than either Bob or Carla.

For the numbers here to be impressive, we need a student who is virtually guaranteed to ace the test. Suppose Esther, for instance, is able to guarantee a score of at least 99 on the test. In that case, Esther will have a failure function f_E^c equaling 0 from 0 to 99. Whatever Esther's probability of getting a perfect score, $I(f_E^c)$ will then be at least 6.658, which is the information associated with Esther if she is guaranteed to score a 99 but cannot score a 100. For $I(f_E^c)$ to become really big, Esther's probability of acing (i.e., getting a perfect score on) the test will therefore need to be very high.

The conditional information associated with these failure functions can also be calculated. $I(f_C^c|\ f_B^c) = 0$, as it must if Brian is in every way better at scoring well on this test than Carla, which he is in this case. On the other hand, $I(f_B^c|\ f_C^c) \approx .718$. This last number, especially given that $I(f_C^c|\ f_B^c) = 0$, indicates that Brian

is more intelligent than Carla at this test. Similar numbers can be calculated for David and Esther. Do such conditional information values provide deeper insight into the degree to which one test-taker is more intelligent than another at this test compared to the non-conditional information values? That's hard to say. It will depend on examining this test-taking situation across a variety of test-takers with different levels of intelligence at taking this test.

Clearly, these information values have the right ordinal properties (they are bigger where they're supposed to be bigger and smaller where they're supposed to be smaller). But interpreting the significance of these numbers, especially in relation to each other, will depend on broad familiarity with the test and its takers. That's how it is in general with these information-based intelligence metrics. The numbers from one situation to another need not be commensurable. But within a situation, they can become commensurable and interpretable.

Cumulative performance measures, as in the test just considered, arise in many competitive contexts. Take a game of golf on a course with 18 holes. Let's imagine par for this course is 72 and that no one ever requires more than 200 strokes to complete the course. Consider therefore the possibility set $\Omega = \{18, 19, 20, ..., 199, 200\}$. Anyone playing on this course will have a score in this range. Moreover, we may assume that at any given moment, a person playing this course will have a probability of attaining each possible score n between 18 and 200.

But note: if lower scores are to count as more successful performances than higher scores, these probabilities need to increase as the number of strokes to complete the course increases. Thus $p_{18} \leq p_{19} \leq p_{20} \leq ...$ where p_n indicates the probability of getting a score of n or lower for any n between 18 and 200. These probabilities will indicate the performance level of a given player. Let's assume these probabilities are associated with a player named Allen. A superior player, call her Barb, will then typically exhibit probabilities $q_{18} \leq q_{19} \leq q_{20} \leq ...$, where for each n between 18 and 200, $q_n \geq p_n$.

Even so, this dominance relationship among probabilities for Barb and Allen need not hold across the board. We can, for instance, imagine that the generally superior player (Barb) is nevertheless not physically powerful enough to make holes in one whereas the less skilled player (Allen) is physically more powerful and can. In that case, it might be that $0 = q_n < p_n$ for $18 < n < 36$, but that for $n \geq 36$, $q_n > p_n$. Such details are easily worked out, though they indicate that for cumulative performance functions, probabilities associated with one agent need not uniformly dominate probabilities of another (compare section 7).

In any case, for cumulative performance probabilities, as for non-cumulative ones, the closer the associated performance probabilities are to 1, the closer

to 0 will be the complementary failure probabilities and thus the greater the associated information value, which in turn implies the greater measured intelligence. Because the performance and failure functions associated with cumulative performance are monotonic (i.e., always tending upwards or always downwards), it doesn't matter what the probability distribution of the underlying Ω space is —these functions when averaged by the underlying probabilities will give broadly the right answers about the relative intelligence of the parties involved. It therefore makes sense to default to uniform probabilities in such cases, though for simplicity and not for deeper philosophical reasons (such as a principle of indifference). If there are reasons to dispense with uniform probabilities and prefer a different probability distribution, this approach can readily accommodate it.

13. Weighted Performance Measures

The mathematics of cumulative performance functions is the same as the mathematics of non-cumulative performance functions. It's just that cumulative performance functions exhibit probabilistic dependencies that restrict their shape with respect to the underlying possibility space. In the non-cumulative performance functions considered in section 11, the possibility space consisted of individual tasks and the performance function assigned to each task the probability of an agent successfully completing it. The cumulative performance function then went further by aggregating these tasks to produce a unified measure of performance across such individual tasks. But in either case, non-cumulative or cumulative, performance functions delivered probabilities and only probabilities.

Often, however, a successful demonstration of intelligence depends on the completion of multiple tasks taken jointly where the tasks come in different forms and bear with different levels or importance of weight on the intelligence in question. Consider a batter in Major League Baseball. In a given plate appearance, he may whiff at the first pitch, collect himself, and then hit a homerun in the same at bat. The whiff in itself would be deemed unsuccessful, and yet the entire plate appearance would be successful, and that's all that matters for the game and for the player's stats. It's "three strikes and you're out," which gives players leeway to whiff as many as two pitches.

It might therefore be argued that swinging or not swinging at individual pitches takes too fine-grained an approach to the task of a batter at the plate, the actual task being to get a hit (or get on base by other means, such as a walk, though let's set that aside for the moment). That seems right. So one could imagine a possibility space Ω consisting of just one element, which denotes getting a hit, and where a probability p is then assigned to that element—in other words, the probability of the batter in question getting a hit. But such a possibility space scarcely captures the richness of baseball. Players do not just get hits, but singles, doubles, triples, and home runs. They drive in runs, sometimes on sacrifice flies, where the out is not credited against the player's batting average. Players also strike out, walk, and get hit by pitches. Moreover, their probability of success at the plate will vary depending on the strength (skill or intelligence) of the pitcher they face.

All these possibilities can be represented by Ω. To keep things simple, let's consider five possibilities that gauge a player's skill/intelligence in a plate appearance: (1) gets a home run, (2) gets a triple, (3) gets a double, (4) gets a single, (5) otherwise gets on base (e.g., walk, hit by pitch, or fielding error). Professional baseball is fastidious at keeping statistics on its players, so these numbers will all be available. A performance function on Ω can here be readily defined, assigning the probability of a player hitting a home run, triple, etc. for each of the five possibilities considered (for simplicity, leave aside pitching and fielding strength). But an information value based on such performance functions, or the related failure functions, may not be very enlightening. Babe Ruth used to joke that if he were hitting for contact like Ty Cobb, he'd be hitting over .600 (Cobb hit over .400 in three of the seasons he played, though Ruth never did).[31] But the most home runs that Cobb ever hit in a season were 12. Ruth hit 60.

So who was the better, the more valuable player? Or who, as we might say, was the more intelligent player? Most historians of baseball would regard Ruth as the greater player. But can we see that from simply looking at the probabilities of Ruth getting a home run, triple, double, single, or otherwise on base versus Cobb's corresponding probabilities? The performance/failure functions here can't tell us how to gauge the difference, nor the information measures as applied to these functions.

What baseball statisticians do these days to gauge success at the plate is form a weighted performance measure that incorporates these probabilities along with various weighting factors and that thereby allows for direct comparability of performance. With a weighted performance measure, we're making a decision on how to combine apples and oranges. Do 40 home runs in a season but a below .200

batting average and with many strikeouts signify greater plate success than 2 home runs in a season but a .320 batting average with hardly any strikeouts? To answer this question requires a weighted performance measure.

Baseball statisticians therefore introduce additional measures such as WAR (wins above replacement) or SLG (slugging percentage). Greater values of these measures are supposed to signify greater value that the player is contributing. Such measures, interpreted as intelligence measures, are therefore not purely probabilistic or informational. Take slugging percentage, of SLG. It is an expected outcome value—the expected number of total bases per at-bat: $SLG = 1 \cdot P(1B) + 2 \cdot P(2B) + 3 \cdot P(3B) + 4 \cdot P(HR)$. It is called a percentage, but it is not really a percentage or probability at all. In fact, SLG can take values between 0 and 4, 4 signifying a player who always and only hits home runs. SLG assigns different weights (1, 2, 3, and 4) to different probabilistic outcomes (getting a single, double, triple, or home run respectively). It is widely regarded as a useful statistic for measuring player strength, or as we might redescribe it, player intelligence.

To sum up, performance measures may be purely probabilistic, as in previous sections, or weighted. A purely probabilistic measure reflects the relative frequency with which an agent reaches a target, assigning a value of one to success and zero to failure. A weighted performance measure distinguishes among degrees or kinds of success, assigning each possible outcome a value and averaging those values according to their probabilities. Batting average is of the first kind, slugging percentage (though not really a percentage) is of the second.

As a postscript, note that for purely probabilistic performance measures, probabilities have natural complements, because probability of success corresponds to one minus probability of failure and vice versa. But for weighted performance measures, there is no natural complement. Barry Bonds in 2001 achieved a slugging percentage of .863, the all-time MLB single-season record for slugging. But there is no natural complement to this number. In general, weighted performance measures capture successful performance and thus don't put the focus on failure or inability as do performance measures gauged purely by probabilities.

14. Multi-Criteria vs. Single-Criterion Optimization

Implicit in mixed performance functions is multi-criteria optimization. To understand multi-criteria optimization, let's first review single-criterion optimization. Single-criterion optimization is about scoring well on a given linear performance criterion (linear as in all scores being comparable by a greater than, equal to, or less than relationship). This can mean a high score, as when higher scores indicate better performance (as with sinking baskets in basketball). Or it can mean a low score, as when lower scores indicate better performance (as in minimizing the number of strokes in golf).

If sports seem too frivolous, consider respectively fitness and penalty functions in an evolutionary computing context, for which the mathematics is the same except that we optimize fitness by maximizing it and we optimize a penalty by minimizing it. Such optimization focuses on a single criterion. It is straightforward, with better scores indicating greater optimization.

Things get more complicated when the optimization needs to be done across multiple criteria. In that case, no single score suffices. Optimizing across multiple criteria requires combining apples and oranges, which in turn requires judgment about how much of an apple matches up with how much of an orange. This is multi-criteria optimization. Multi-criteria optimization was implicit in the last section, as with mixed performance functions in baseball.

In baseball, it's great when a player can get on base by swinging a bat and making contact with a ball. But it's also great when a player can get on base by not swinging a bat, such as by getting a walk or getting hit by a pitch (getting hit in this way may not seem so great to the player at the time, but it is great for his stats). These two ways of getting on base are very different. Getting on base without swinging a bat gets a runner to first base and moves a runner from first to second, but won't score a runner on third if there are no intervening runners on first and second. Getting a single, on the other hand, will almost always move all runners, scoring a runner on third

How then to decide which player is better (more intelligent) with respect to these two criteria? What's needed is to combine performance for each of these criteria into a single criterion. That, for instance, is the job of the OPS metric (on-base plus slugging percentage). It combines getting on base by swinging a bat with getting on base by not swinging a bat. Moreover, this metric rewards how many bases you get when you get on base by swinging a bat—the more bases,

the greater the reward. OPS is a weighted sum of probabilities, but it is not a probability.

One can debate the merits of how well OPS combines these two criteria to produce a single criterion that captures effectiveness at plate appearances. Yet only by reducing a multi-criteria optimization to a single-criterion optimization is it possible to compare performance across the criteria and thereby assess total performance. In reducing multi-criteria to single-criterion optimization, we are, strictly speaking, no longer in the realm of performance and failure functions induced by probabilities. And yet, we're not far from these. In fact, it's typically straightforward to derive probabilities from single-criterion optimization measures that combine multiple criteria. We examine how this works next.

15. Percentile Performance/Failure Functions

In general, if we have a set of scores (numbers) where larger scores indicate better performance or greater intelligence, we can assign percentiles to these scores. In such cases, percentiles correspond to what percent of scorers score at a given level or better. Percentiles are numbers between 0 and 100, but by putting the word "percent" after any number, we divide it by 100. Percentiles are thus in fact probabilities because dividing numbers between 0 and 100 by 100 contracts them to numbers between 0 and 1, which is the range for probabilities. Conversely, probabilities are turned into percentiles by multiplying them by 100. Because of this straightforward correspondence between percentiles and probabilities, we won't distinguish between the two very carefully. Thus we will refer to a percentile p even when for purposes of calculation we treat p as a probability.

When we have a set of scores where greater scores indicate better performance, we are able to construct percentile performance functions. Thus for an intelligence with a given score at a percentile p but not at a higher percentile, the corresponding percentile performance function assigns a probability of 1 to each score less than that given score and p for all scores greater than or equal to it.

The corresponding percentile failure function will then be 0 for each score less than the given score and 1–p for the given and higher scores. An information-based intelligence measure for that score will then be $-\log_2(1-p)$. Note, we could also reverse the ordering so that lesser scores indicate better performance

or greater intelligence (as in golf), but that situation is equivalent, except that the direction of the scores is reversed. To keep things simple, we'll focus on bigger numbers signifying better performance.

In assigning percentiles to scores, we simply rank the scores in order and then count what proportion fall at or below a given score. This type of rank order happens, for instance, with the ACT test, an aptitude/intelligence test used in the United States to guide college admissions.[32] Its composite scores vary from 1 to 36. There's nothing in these scores per se that tells us how much better a score of 35 is compared to a score of 12. To understand the difference, we need to convert ACT scores to percentiles, as in the following table. The first number here is the ACT composite score, the second is its corresponding percentile:[33]

36: 99.8435 percentile	24: 73.2698 percentile	12: 4.3192 percentile
35: 99.2244 percentile	23: 68.7370 percentile	11: 1.4749 percentile
34: 98.1491 percentile	22: 63.7906 percentile	10: 0.4690 percentile
33: 96.8252 percentile	21: 58.4691 percentile	9: 0.1837 percentile
32: 95.2892 percentile	20: 52.8388 percentile	8: 0.0841 percentile
31: 93.5489 percentile	19: 46.9981 percentile	7: 0.0400 percentile
30: 91.5619 percentile	18: 40.9931 percentile	6: 0.0181 percentile
29: 89.3408 percentile	17: 34.8364 percentile	5: 0.0076 percentile
28: 86.9025 percentile	16: 28.5792 percentile	4: 0.0027 percentile
27: 84.1217 percentile	15: 22.2046 percentile	3: 0.0007 percentile
26: 80.9471 percentile	14: 15.7409 percentile	2: 0.0001 percentile
25: 77.3474 percentile	13: 9.4335 percentile	1: 0.0000 percentile

Given these percentiles, we can now assign an information value to these scores. For instance, an ACT score of 36 corresponds to a percentile of 99.8435. These percentiles are empirically based, depending on the performance of test takers. They therefore induce empirically based probabilities. The probability of an arbitrary test-taker getting a 36 is therefore 1 – .998435 = .001565. The probability of an arbitrary test-taker getting a 35 or better is therefore 1 – .992244 = .007756. These probabilities correspond to the values taken by cumulative failure functions (see section 12), and so the information values associated with them are I(ACT=36) = I(event of probability .001565) = $-\log_2(.001565) = 9.32$ and I(ACT=35) = I(event of probability .007756) = $-\log_2(.007756) = 7.01$.

Note that an equivalent analysis can be done for composite SAT scores, which vary between 400 and 1600. A score of 1600 corresponds to a percentile of 99.9826,

and so I(SAT=1600) = I(event of probability 1–.999826) = 12.49. On a percentile basis, the SAT is therefore better able than the ACT to discriminate intelligence at the highest level. The ACT at its highest yields an information-based intelligence value of 9.32 whereas the SAT yields a corresponding information value of 12.49.[34]

In general, a percentile performance function that for a given score gives a percentile of p corresponds to a percentile failure function that evaluates to 1–p, with an information-based intelligence metric for that score taking the value I(event of probability 1–p) = $-\log_2(1-p)$.

16. Effect-Size Performance/Failure Functions

The theory of measurement distinguishes between ordinal and interval scales.[35] Ordinal scales focus exclusively on the relative order of numbers. Interval scales include this ordinal property of numbers but also measure differences in amounts of the numbers, which can be interpreted as distances. For intelligence metrics, such distances can contain crucial information about the degree of intelligence leading to the numbers in question.

To see what's at stake here, imagine a class of 100 students where each student gets a score on a midterm. Let's say the average score for these students is a 75 and that based on past experience with this test, scores are distributed normally around the mean of 75 with a standard deviation of 5 points. Suppose that the second highest score on this exam was an 85 but that the very highest score was a 95. Assume no ties for either score.

If we were treating these scores from a purely ordinal perspective, as in the last section, the highest scoring student would have scored in exactly the 99$^{\text{th}}$ percentile and the second high student would have scored in exactly the 98$^{\text{th}}$ percentile. The information values corresponding to these percentiles would thus, as described in the last section, have been I(highest ordinal score on exam) = I(event of probability .01) = 6.64 and I(second highest ordinal score on exam) = I(event of probability .02) = 5.64.

But from the vantage of an interval scale, the performance of the best versus the second-best student on this exam looks quite different. Given that test scores are distributed normally, the second-best student scored at precisely two standard deviations above the mean. The probability of scoring that high or higher is

therefore .02275, leading to the following information calculation: I(interval score by second-best student) = I(event of probability .02275) = 5.46. But the very best student scored at precisely four standard deviations above the mean. The probability of scoring that high or higher is in that case .00003167, leading to the following information calculation: I(interval score by best student) = I(event of probability .00003167) = 14.95.

Because information is measured on a logarithmic scale, an information-based intelligence metric score of 14.95 is vastly bigger than 5.46. By taking into account the interval scale behind these test scores, we can see that with respect to this exam, the intelligence of the best student towers above that of the rest of the students who took the exam—and that includes even the number two student. Such a fine-grained distinction among levels of intelligence would be lost from a purely ordinal scale but is here captured through an interval scale. Percentiles based purely on ordinal data, while always available, can be hard pressed to make proper sense of such outliers and thus how much more extreme their performance is compared to typical performance. When interval data is available, we thus may be able to gain further insight into the underlying information-based intelligence metrics.

Information-based intelligence metrics that draw on interval rather than purely ordinal data can be thought of as measuring *effect sizes*. Effect sizes quantify the magnitude of differences by generalizing the concept of standard deviations for probability distributions. Standard deviations measure the spread or variability of a single probability distribution around its mean. Effect sizes extend this idea to quantify the separation and/or overlap between two or more probability distributions.

Consider, for instance, two groups of students that take the same exam. One group doesn't do any special preparation (the control group). The other group does special preparation to excel on the exam (the treatment group). Depending on the variability within groups and the mean difference between groups, an interval-based measure of the distance between the means of both groups will exist, and from there it will be possible to calculate an analogue to the number of standard deviations in that distance and a corresponding probability for achieving the mean in one group conditional on the other group. In this way, any increase in intelligence through the treatment as compared to the control can be measured.

Effect sizes in statistics apply to such situations, quantifying the magnitude of a relationship or difference between groups and not just their ordering by some ranking criterion. Effect sizes therefore offer a more meaningful measure than significance testing, which only determines whether an effect exists. Unlike

p-values, which indicate whether an observed result is unlikely due to chance, effect sizes such as Cohen's d (for standardized mean differences), Pearson's r (for correlation strength), and odds ratios (for relative likelihoods) provide insight into practical significance.[36]

For example, a large study may find that aspirin use significantly reduces the risk of heart attacks (p-values less than 0.00001 have been reported in some studies).[37] But, as one statistics text on effect sizes notes, the actual effect size of aspirin is small:

> The benefits of aspirin consumption expressed in correlational form are tiny, just $r = .034$. This means that the proportion of shared variance between aspirin and heart attack risk is just .001 (or .034×.034). This sounds unimpressive as it leaves 99.9% of the variance unaccounted for.[38]

Consequently, the benefit of taking aspirin, though real, is quite modest for the average individual. Moreover, aspirin has its own risks in the form of ulcers, bleeding, and the need for blood transfusions. In any case, understanding effect sizes helps researchers distinguish between statistically detectable effects and those that are practically meaningful.

A general framework for this distance-based effect-size approach to measuring intelligence now looks as follows. We start with a metric space M having a distance function d. Consider a point u in the metric space, and let's imagine a probability measure P on M that assigns probabilities to all subsets of M of the form $M_{u,r} = \{x \in M \mid d(x,u) \geq r\}$ for all nonnegative real numbers r. $M_{u,r}$ denotes all points in M that are at least a distance r from u.

As r increases, $P(M_{u,r})$ gets smaller. If we think of u as embodying average performance for elements of M and distance from u as gauging improved performance across M, then any intelligence responsible for some w in M where $d(w,u) = r$ will command an information-based intelligence metric score of $-\log_2 P(M_{u,r})$. Note that this general case also handles scoring of standard deviations above the mean for a normal distribution (i.e., in the right tail) as in the first example of this section. In this case, what's needed is to identify the metric space M with the half interval $[u,\infty)$. In this way, directionality of the scoring is respected.

The effect-size approach to intelligence is implicit in many of the intelligence metrics in use. Consider, for instance, the intelligence metric used to measure competitive advantage in chess. The ranking of chess players worldwide is based on the Elo rating system, developed by Arpad Elo. The system calculates the relative skill levels of players in two-player games, notably in chess but also in other competitive contexts.[39]

Each player starts with an initial rating that's the same for everyone starting out, and then the player's rating increases or decreases based on game outcomes (win-lose-draw). The amount of change depends on the rating of the opponent. Beating a higher-rated opponent leads to a greater rating increase than beating a lower-rated player, while losing to a lower-rated player leads to a greater rating decrease than losing to a higher-rated player. Drawing a comparably rated opponent doesn't change your rating, but drawing to a weaker/stronger opponent respectively lowers/raises your rating.

Elo ratings in chess don't just yield percentile performance functions. They also yield effect-size performance functions. Calculating the precise magnitude of those effects is built into the Elo ratings, with 400 being a convenient number in the sense that someone who has a 400-point advantage in an Elo rating over another person is 10 times as likely to win against that person, with other effects then being adjusted accordingly.[40]

While chess players may ignore the precise statistics connected to the Elo rating system, they know instinctively that more than percentiles are at play with Elo ratings. They can tell by looking at a rating how much better and more impressive one player is compared to another. In May of 2023, for instance, here were the Elo ratings of the top ten chess players in the world:[41]

1	Carlsen, Magnus	NOR	2853
2	Nepomniachtchi, Ian	RUS	2794
3	Ding, Liren	CHN	2789
4	Firouzja, Alireza	FRA	2785
5	Nakamura, Hikaru	USA	2775
6	Giri, Anish	NED	2768
7	Caruana, Fabiano	USA	2764
8	So, Wesley	USA	2760
9	Anand, Viswanathan	IND	2754
10	Radjabov, Teimour	AZE	2747

The difference of almost 60 rating points between Carlsen and Nepomniachtchi (typically shortened, thankfully, to "Nepo") suggested that at the time Carlsen still dominated the chess world. A year later, in May of 2024, Carlsen still occupied the top spot, but his rating was down to 2830 whereas the number two spot was now occupied by Fabiano Caruana, who had moved up to 2805.[42] By January 2025, Carlsen had inched up to 2831 and Caruana down to 2803.[43] This suggests

that Carlsen was no longer as dominant as he had been, though from a purely percentile perspective, nothing had changed in his "chess intelligence."

As a historical note, Carlsen's highest ranking ever was in May of 2014 and then again in August 2019, when he peaked at 2882, surpassing Garry Kasparov's ranking of 2851, which until Carlsen's had been the highest Elo chess rating ever.[44] How much better was Carlsen's high of 2882 compared to Kasparov's high of 2851? How much playing strength did Carlsen lose from his all-time high to his current (2025) Elo rating of 2830? Answering such questions may require not just deeper statistical analysis but perhaps also extensive experiential understanding of the relative playing strengths of the top echelon of chess players at a given time, including assumptions that may not be universally shared and factors that may be beyond the reach of any intelligence metric.

To continue on this historical note, how does Carlsen's chess dominance compare to that of Bobby Fischer's back in the summer of 1972, when Fischer defeated Boris Spassky for the world championship in Reykjavik, Iceland? Fischer's ranking back then was 2785, the highest ever at the time. This rating put him over a hundred rating points ahead of Spassky's 2660, the second highest rating at the time (the future world champion Anatoly Karpov had, back then, a rating of 2630).[45] Fischer's 125-point lead over the rest of the chess world may, though only briefly (Fischer retired from competitive chess right after his victory over Spassky), have signified the greatest chess dominance ever. But again, from a purely percentile point of view, we would not know that.

The prototypical example of an intelligence metric is, of course, the intelligence quotient, or IQ. Consisting of a mean and standard deviation measure, IQ exemplifies an effect-size performance function. Leaving aside the validity of the IQ test (i.e., whether it actually measures what we would want to call general intelligence—the joke is that IQ measures IQ), it is defined so that the numbers correspond to a normal distribution with a mean of 100 and a standard deviation of 15. Thus IQ scores break down as follows, in line with the underlying normal distribution:

- 130 and above: 98th percentile
  (which is the cutoff for "gifted" programs)
- 120-129: 91st-97th percentile
- 110-119: 75th-90th percentile
- 100-109: 50th-74th percentile
- 90-99: 25th-49th percentile
- 80-89: 9th-24th percentile

- 70-79: 2nd-8th percentile
- 69 and below: 2nd percentile or below

17. Random Predicates as Generalized Performance/Failure Functions

Earlier, in section 11, we defined performance functions as functions f from Ω to [0,1] where for each x in Ω, $f(x)$ is the probability of an agent succeeding in performing the task indicated by x. These are the "ordinary" performance functions. For any given task x, $f(x)$ in this case assigns a fixed static probability of x being successfully accomplished. For many situations such an assignment of probability is adequate. Yet it can also miss the full richness of performance, for which we need "generalized" performance functions.

Suppose, for instance, that x and y are in Ω, and that if an agent successfully performs x, then that agent is guaranteed to perform y successfully. In that case, success at x would entail success at y. Such a relation indicates a probabilistic dependence between the tasks x and y that (ordinary) performance functions as developed so far are unable to capture. Such a performance function f from Ω to [0,1] will only be able to partially capture this dependency, such as by requiring $f(x)$ to be less than or equal to $f(y)$. Yet for all that the performance function f cares, the agent's performance on x could be probabilistically independent from the agent's performance on y. If they were probabilistically independent and the probability of success for both tasks were strictly less than 1, the agent could be successful at x and yet still on occasion fail at y.

As it is, probabilistic dependencies in performance are widespread. Often they take the form of priming effects. Thus, even though accomplishing a task y may be difficult on its own, by first having someone successfully perform task x, task y may become much easier. Think of, for instance, a math exam where question x presents a key technique which is then just what's needed to answer question y—in math exams, often one question provides a hint for another. Or consider that the order of test questions can be crucial to overall performance. For instance, in the ACT and SAT exams, the easier questions come first, the harder later. To

present the hardest questions first would have a demoralizing effect on many test-takers and would thus undermine their overall performance.

To capture performance more fully, we therefore incorporate probabilistic dependencies into performance functions. Interestingly, unlike the performance functions of section 11, whose outputs are probabilities (of success) and which thus take values in the unit interval [0,1], generalized performance functions only need to take values in the binary set {0,1}. That ends up working because generalized performance functions, although they need to assign probabilities of success and failure, can do so by assigning a probability to obtaining 0 and alternatively to obtaining 1. The full interval [0,1] is thus, as it were, recovered by assigning probabilities to the binary set {0,1}.

The mathematics of generalized performance function goes back to an idea I had in the 1990s under what I called *random predicate logic*, and which I eventually published in 2002 with the now defunct ISCID journal.[46] Given a space Ω, an ordinary predicate F on it assigns a truth value to any x in Ω. Thus, for any x in Ω, F delivers for x a value of true or false. By letting 1 denote true and 0 denote false, we can therefore define an ordinary predicate F as a function from Ω to the binary set {0,1}.

Given a probability space Θ with probability measure Q, we can now define a random predicate F that maps the cartesian product $\Omega \times \Theta$ to {0,1}. Recall that Ω has the probability measure P, and, as just noted, Θ has the probability measure Q. We can think of the random predicate F in three ways:

1. As a joint 0-1-valued function of x in Ω and θ in Θ, F sends (x,θ) to $F(x,\theta)$.
2. For a fixed x in Ω, $F(x, \theta)$ is a 0-1-valued random variable for θ in Θ. In this case we write F_x for the random variable in θ ($F_x(\theta) = F(x,\theta)$).
3. For a fixed θ in Θ, $F(x,\theta)$ is a 0-1-valued function for x in Ω. In this case we write F_θ for the old-fashioned predicate on Ω induced by θ ($F_\theta(x) = F(x,\theta)$).

The third of these approaches provides the clearest representation of F as a random predicate. In standard probabilistic fashion we suppress the θ in F_θ and write F for F_θ whenever θ is unspecified. If in fact F_θ does not vary as θ varies (i.e., holds constant with changing θ), then F is just an old-fashioned predicate. In this way random predicate logic subsumes ordinary predicate logic.

To the second of these approaches, dependencies among the random variables F_x are of special interest. To illustrate, we'll consider the classic paradox of the heap. A single pebble does not make a heap. But add enough pebbles, and eventually a heap emerges. Yet there's no precise point at which a tightly packed

collection of pebbles forms a heap. At the extremes something is or is not a heap. But in the middle, there's fuzz or vagueness. Similarly with baldness. A head bereft of all hair is bald. A head full of hair is not. A head with a single hair is bald. But as hairs keep accumulating, there comes a point at which it's unclear whether the person is bald.

Consider, then, the case where F denotes the random predicate that something is a heap. Suppose Ω consists all tightly packed collections of pebbles, with all pebbles having the same size. Let a denote a collection of a thousand pebbles and b denote a collection of a million pebbles (a and b are in Ω). Clearly, whatever randomization makes F_a out to be a heap, F_b must also make b out to be a heap. It follows that the random variable F_a is always less than or equal to the random variable F_b. It follows that $Q(F_a = 1) \leq Q(F_b = 1)$ and $Q(F_a \leq F_b) = 1$. Thus, to understand the vague predicate "x is a heap" when interpreted as a random predicate F, it is essential to consider dependencies among random variables such as F_a and F_b.

Random predicate logic straightforwardly extends traditional first-order predicate logic. Moreover, its formalism subsumes type-1 fuzzy sets.[47] What interests us here, however, is that random predicates are able to generalize performance functions, the predicate in this case for x in Ω being that "the task x is performed successfully." An ordinary performance function f simply assigns a probability $f(x)$ to the successful performance of x. A generalized performance function, which is to say a random predicate, can capture that probability but also capture dependencies among those probabilities as x varies. Thus, for a generalized performance function F, integrating out by the randomizing element θ in the probability space Θ with probability Q yields an ordinary performance measure of f:

$$\int_{\Theta} F(x, \theta) dQ = f(x).$$

Generalized performance functions can structure the probability space Θ and its probability Q so that for x in Ω, the random variables are each probabilistically independent. Such probabilistic independence seems implicit with ordinary performance functions, where probability of performance on a given item doesn't depend on probability of performance for other items. Nonetheless, generalized performance functions can also model probabilistic dependence. We've already seen the ability of generalized performance functions to capture such probabilistic dependencies, as when certain test items are guaranteed to be successfully performed whenever other test items are performed successfully.

Just as an ordinary performance function f becomes a ordinary failure function by subtracting it from 1, i.e., $1 - f = f^c$, so too a generalized performance function F becomes a generalized failure function by subtracting it from 1, i.e., $1 - F = F^c$. The logic and the interpretation for generalized failure functions in relation to generalized performance functions all carry through straightforwardly: probabilities flip and apply to failure rather than success.

Corresponding informational measures also apply. With an ordinary failure function $1 - f = f^c$,

$$I(f^c) = -\log_2 P(f^c) = -\log_2(\textstyle\int_\Omega f^c dP).$$

Similarly, with a generalized failure function $1 - F = F^c$,

$$I(F^c) = -\log_2(\int_\Omega [\int_\Theta (1 - F) dQ] dP) = -\log_2(1 - \int_\Omega [\int_\Theta F dQ] dP).$$

In calculating the information associated for a generalized failure function, one first integrates out the randomizing element θ in Θ, which yields an ordinary failure function, and then integrates over the task performance space Ω to calculate overall failure, which when placed under a negative logarithm to the base 2 then yields an intelligence metric for generalized failure functions. Note that conditional probability and conditional information generalize straightforwardly.

18. Measuring Intelligence in the Search for Targets

When an intelligence acts intelligently, its challenge typically is to find a point x in a target T that's a subset of a possibility (or search) space Ω. It may be finding an Easter egg in a large field. It may be finding a combination that opens a combination lock. It may be finding the right design parameters in a space of possible artifacts that yields an artifact valuable to users and profitable to makers. It may be finding a sequence of amino acids that forms a protein exhibiting a given function.

Finding a point x in a target T that's a subset Ω is the simplest case of an intelligence doing a search. This elementary search formulation admits many extensions. The target may move, its location may be only partly known, and detection may be noisy or probabilistic. The searcher may also face limits on time,

movement, sensing, or computation, while seeking to maximize the probability of detection, minimize search time, or balance gains against costs. More complicated search frameworks allow multiple targets or searchers, outcomes of unequal value, graded rather than binary success (e.g., targets that are fuzzy sets), and strategic targets that evade or assist the searcher.

Even so, the fixed-target model remains the simplest and most widely used. Many generalizations can be recast within it by enlarging the search space. A moving target may be represented by a trajectory in a space of possible trajectories. An adaptive search may be represented by a policy in a space of policies (a policy in decision or control theory being a rule that maps prior search information to the next action). A changing target may be represented by a fixed set of successful state-time pairs. Graded success may be represented by nested targets or performance function defined over enlarged search spaces. Such reformulations may be unilluminating and computationally inefficient, but they show that many generalized searches can still be reconceptualized as locating an element of a fixed target within an expanded possibility space.[48]

In moving forward, then, with our study of intelligence in search, we go with the simplest search scenario in which a target T is clearly defined in the sense that for any x in Ω, we know whether x is or is not in T. The challenge for searches we consider will thus be to find such an x in T. An agent looking for T will then be regarded as intelligent to the degree that it readily finds T. But what does it mean to readily find T? This question can be answered in terms of probabilities: the more probable it is for the agent to find T, the greater its intelligence at doing so. Intelligence in this case becomes relative to the probability of finding the target T.

Suppose, for instance, we have two agents, call them Alice and Bob. To gauge their intelligence in finding T will then mean to determine their probability of doing so. Let's say Alice has probability p of finding T and Bob has probability q. Let's say that q is a lot bigger than p. In that case, Bob's intelligence at finding T will be greater that Alice's. But that means the information associated with q will actually be less than the information associated with p (negative logarithms of bigger numbers calculate to smaller information values, indicating less narrowing of possibilities).

Thus, in line with our guideline to measure intelligence in terms of failure rather than success, we calculate the probability of Alice failing to find T, namely $1–p$, and the probability of Bob failing to find T, namely, $1–q$. Since q is a lot bigger than p, $1–q$ is a lot smaller than $1–p$, and so $I(\text{Bob}) = -\log_2(1–q) > -\log_2(1–p) = I(\text{Alice})$. If, for instance $p = .75$ (Alice has a 75 percent chance

of finding T) and if $q = .999999$ (Bob has a 99.9999 percent chance of finding T), then $I(\text{Alice}) = -\log_2(1-.75) = -\log_2(.25) = 2$ bits whereas $I(\text{Bob}) = -\log_2(1-.999999) = -\log_2(.000001) = 19.93$ bits, or almost 20 bits. This would suggest quite a disparity between the intelligence of Alice and Bob in finding T.

An alternative approach to gauging intelligence in search arises from work on conservation of information. Circa 2010, Robert Marks and I proved a number of conservation-of-information theorems showing that when a search with small probability p of success is improved to a search with probability of q of success where $p < q$, that improvement incurs an information cost corresponding to a probability of at least p/q, the associated information cost being $-\log_2(p/q)$.[49]

With conservation of information, the improvement of search performance from a small probability of success p to a large probability of success q is captured in the ratio p/q. The key information measure within the theory of conservation of information is the negative logarithm to the base 2 of this ratio, or as we just saw $-\log_2(p/q)$. This number is called the *active information* (or also *added information*), and it describes how much information is required, or must be added, to improve the search from probability p of success to probability q of success. It therefore provides another measure of intelligence, quantifying how much more intelligent the agent associated with q is compared to the agent associated with p.

In the Alice and Bob case up above, with Alice's $p = .75$ and Bob's $q = .999999$, $p/q = .75/.999999$, which is almost exactly .75, and so the negative logarithm to the base 2, which is to say the active information, will be a mere .415 bits. This number hardly seems to capture how much more intelligent Bob is compared to Alice in finding the target, Bob being virtually assured to find it, Alice missing it 1 out of 4 times. Where active information shines, however, is when the probability p is minuscule, suggesting virtually no chance of finding T, and the probability q is by comparison much larger, suggesting a reasonable chance of finding T.

Consider, for instance, an example by Richard Dawkins where he considers the evolution of a target phrase METHINKS IT IS LIKE A WEASEL by blind search and by cumulative selection among sequences of capital Roman letters and spaces.[50] By blind search, the probability p of finding the target phrase is roughly 1 in 10^{40}. By cumulative selection, the probability q of finding the target phrase is virtually assured, say, .999999. In that case, the active information is $I = -\log_2\left(\frac{10^{-40}}{.999999}\right) = 132.877$ bits.[51]

From the vantage of active information, Dawkins' cumulative selection algorithm adds a huge amount of information—or intelligence as we might say. Does this mean that cumulative selection is more intelligent than blind search? Yes and no. Yes in the obvious sense that it was able with much higher probability to find

the target phrase than blind search. No in the less obvious sense that Dawkins needed to supplement cumulative selection with an explicit measure of proximity between sequences in general and the target phrase.

That supplementation was crucial to the success of Dawkins' algorithm. In general, when a search algorithm improves on blind search, it has to borrow information from elsewhere—it cannot generate the information from scratch. That is the upshot of conservation of information. In this example, the information was borrowed from Dawkins' mind, which is to say from his intelligence.

Active information is worth comparing with our conventional information measure that gauges intelligence via failure, less failure indicating greater intelligence. The information measure associated with Bob, now reimagined as cumulative selection, will as before evaluate to $I(\text{Bob}) = -\log_2(1-.999999) = -\log_2(.000001) = 19.93$ bits, whereas the information measure associated with Alice, now reimagined as blind search, if given by $p = 10^{-40}$, will evaluate to $I(Alice) = -\log_2(1-10^{-40}) =$

$$.00000000000000000000000000000000000001443 \text{ bits.}$$

As measured by active information, it's clear from a single number, in this case $I(\text{ActiveInfo}) = 132.877\ bits$, how much more intelligent cumulative selection (Bob) is compared to blind search (Alice). With the conventional approach that measures information in terms failure, we need two numbers, $I(\text{Alice})$ and $I(\text{Bob})$, and a comparison of them that calculates not just the difference between the two but also the proximity of $I(\text{Alice})$ to zero. The conventional approach is easier to calculate and applies more generally. Yet it lacks the convenience and interpretability of a single number.

When only one of the agents has an exceedingly small probability of hitting the target, then the active information approach makes clearer the disparity in intelligence between two agents, encapsulating in a single number what in the conventional approach requires careful comparison of two numbers. The conventional approach is strictly speaking more general, allowing reconstruction of the active information approach, though not vice versa. Yet the active information approach provides more intuitive insight into a difference in intelligence when one of the intelligences has virtually nil probability of finding the target. We continue this train of thought in the next section.

19. Negative Intelligence and Disinformation

Given the information-theoretic view of intelligence developed in this paper, it would seem that any measure of intelligence should never fall below zero. We measure intelligence in terms of performance at tasks, and specifically in terms of how probable it is that one can perform the task successfully. The lowest intelligence one could have at a task would therefore be to have a 0-probability of completing it successfully, which would correspond to a failure probability of 1. And since intelligence is measured information-theoretically by taking a negative logarithm to the base 2 of a probability, that would correspond to an information *qua* intelligence measure of 0. Probabilities fall on the closed unit interval $[0, 1]$. The corresponding information measures fall on the closed interval $[0, \infty]$, with information 0 corresponding to probability 1, and information ∞ corresponding to probability 0. It would therefore seem that there's no way to get a negative measure of intelligence.

But there is. A 0-probability of success or a 1-probability of failure corresponds to a 0-measure of intelligence in terms of information. But in many circumstances, such a 0-probability should be entirely off the table for consideration. In many circumstances, one can attempt uniform random sampling or exhaustive search or some other brute force method to locate the target. These will set a baseline below which no search performance need dip. The probability p of success here might be very small, but it typically will not be zero. There should thus never be any reason to employ a search process with probability less than p. Rather, any search process one employs to try to locate the target should have probability q where q is at least as great as p, preferably much larger than p if p is extremely small so that the search with probability q of success will stand a reasonable chance of actually being successful.

For instance, imagine an Easter egg hunt in which the egg is well hidden in a large field. Let's say the field is so large and the egg is so well hidden that the probability of successfully finding it with a purely random search (such as uniform random sampling) has probability p, which is extremely small albeit positive. Next, suppose that someone starts calling out "warm, colder, warmer, warmer, colder, cold, warm, warmer, hot, …" Those instructions, if followed, may change the probability of finding the egg and induce a new search.

If the information being shouted is accurate, the probability of q will be much greater than that of p. The probability of q will then correspond to a guided

search, the probability of p to a random search, and the intelligence of the q-search will be much greater than the intelligence of the p-search, as measured by our intelligence metrics, notably active information of the last section. Thus, with q much greater than p, the active information will then be $-\log_2(p/q) = \log_2(q/p) > 0$ since q/p will then be greater than 1.

But all of this depends on the instructions that are shouted being accurate. What if instead the instructions are inaccurate and designed to make it more difficult than the baseline probability p of finding the target? In that case, q will be less than p, and $\log_2(q/p)$ will be less than 0, and thus negative. Such searches in which q is less than the baseline probability p are misleading. At their most extreme, such searches make the probability to be 0 of successfully finding the target, in which case $\log_2(q/p) = \log_2 0 = -\infty$.

Thus we see that active information can be negative when in a non-baseline search the probability q of performing a successful search falls below that of a baseline search with positive probability p. Situations like this suggest that deliberately misleading information is being given in or through the non-baseline search because no one should ever have to carry out searches at a performance-level below the baseline. This suggests a definition if *disinformation*: information that makes it less likely to locate a target than simply going with a baseline search available to anyone searching for the target.

20. Sequential Search and Waiting Times

Typically in the search for a target T in a possibility space Ω, we don't just assign single probabilities to different searches, as in one search having probability p of success and another having probability q of success, as in the last section. In general, we think of a search as a process that is sampling sequentially over time from the possibility space and that, to the degree the search is effective, is continually honing in on the target, with the probability of hitting the target becoming ever greater over time.

The way to represent such a search is as a sequence of random variables $X_1, X_2, X_3, ...$ where each X_i (for $1 \leq i < \infty$) takes values in Ω and a probability measure P exists that characterizes the probabilistic behavior of these random variables. Thus, for instance, $P(X_1 \notin T, X_2 \notin T, ..., X_{m-1} \notin T, X_m \in T)$ would

denote the probability that the random variables $X_1, X_2, X_3, ...$ first successfully find T on the m^{th} try but not earlier. For ease of notation, let's denote by X the search consisting of the random variables $X_1, X_2, X_3, ...$

Note that the random variables $X_1, X_2, X_3, ...$ form a discrete stochastic process (discrete in the sense that the indexing variable i for each X_i is an integer rather than a real number). We could also represent search not as a discrete stochastic process, as here, but as a continuous stochastic process where the indexing variable spans the non-negative reals, as in $\{X_t |\ t \in [0, \infty)\}$. We'll focus on the discrete case since continuous stochastic processes can typically be approximated with discrete stochastic processes (as when a random walk approximates Brownian motion), and results concluded for the discrete case often generalize straightforwardly to the continuous case.[52]

Associated with a search $X = X_1, X_2, X_3, ...$ for a target T is a waiting time U that gives the first time that X successfully finds an item in T. It follows that $U = m$ if and only if $X_1 \notin T, X_2 \notin T, ..., X_{m-1} \notin T$, *and* $X_m \in T$. Thus $P(U = m) = P(X_1 \notin T, X_2 \notin T, ..., X_{m-1} \notin T, X_m \in T)$. But the more important probability for our purposes is one we'll denote by p_m. This is the probability $p_m = P(U \leq m) = P(X_1 \in T \textit{ or } X_2 \in T \textit{ or } ... \textit{ or } X_m \in T)$, namely, the probability that the search for T is successful in m or fewer steps. The probability that the search takes longer than m steps is then $P(U > m) = 1 - p_m$.

Given a search $X = X_1, X_2, X_3, ...$, we can imagine another search $Y = Y_1, Y_2, Y_3, ...$ with corresponding waiting time V for which $P(V = m) = P(Y_1 \notin T, Y_2 \notin T, ..., Y_{m-1} \notin T, Y_m \in T)$ and $P(V \leq m) = P(Y_1 \in T \textit{ or } Y_2 \in T \textit{ or } ... \textit{ or } Y_m \in T) = q_m$. We can then define the associated information measures $I(U > m) = -\log_2(1 - p_m)$ and $I(V > m) = -\log_2(1 - q_m)$, which, in line with our customary approach of measuring intelligence via failure, measure the information in failing to find T within m steps for the searches X and Y. Additionally, we can then define the associated active information measure $I(ActiveInfo_m) = -\log_2(p_m/q_m)$. The behavior of these information values as m grows large will then characterize the relative intelligence between the two searches X and Y at finding the target T.

Depending on the types of searches to be analyzed, these probabilities can offer different insights into the relative intelligence associated with the searches X and Y as well as any additional searches with which we might want to compare them. For a search in general, we are concerned to know the initialization (where the search starts), the objective function (what gauges closeness or progress to the target, such a function often taking the form of a fitness landscape), the

update rule (how to proceed from one search candidate to the next), and the stop criterion (the point at which the search is to be called off).

Because we live in a universe with finite resources, the stop criterion will always incorporate an upper bound M beyond which the search cannot proceed. For instance, in early 2025, the top supercomputer, Lawrence Livermore's Cray dubbed El Capitan, maxed out in the 1 to 2 exaflop range (an exaflop = 10^{18} floating point operations per second).[53] Because thirty years is roughly a billion, or 10^9, seconds, running such a computer for thirty years would therefore allow for roughly 10^{27} floating point operations. Any search requiring more than that many floating-point operations within a thirty-year span would, to have a reasonable chance at finding the target T, need to be carried out on a different and more powerful computer.

The bound M of a stop criterion sets a hard upper limit on the number of steps that a search can be extended. In the El Capitan example, this corresponds a hard limit on the number of feasible floating-point operations. Or course, if this supercomputer could be run longer, then M would increase. If, *per impossibile*, it could be run for thirty trillion years (orders of magnitude beyond the estimated age of the universe), that would give roughly 10^{21} seconds total, and thus 10^{39} floating-point operations total.

That may seem like a lot, but even that many opportunities to sample a search space is often insufficient to resolve a search. The complexity of many search problems increases exponentially with the size of a search. An example is the traveling salesman problem, the number of whose potential paths (Hamiltonian cycles) with n nodes is $(n-1)!/2$. For $n = 36$, this number is beyond 10^{39}, implying that a brute force (guaranteed) approach to solving this problem could not be accomplished in 10^{39} or fewer steps. Moreover, each of these steps would, presumably, require well beyond a single floating-point operation.

A particularly common type of search X is one in which the random variables $X_1, X_2, X_3, ...$ are independent and identically distributed (iid). That means that each random variable X_i makes an independent attempt at finding T and does so each time in the same way (just as someone trying to roll double sixes with two dice has the same chance at each roll of the dice, with each roll having no memory of the others). If the probability of X_i landing in T is p and all these random variables are independent and identically distributed (iid), then for the corresponding waiting time U, the probability of $P(U = m) = p \cdot (1-p)^{m-1}$ and $P(U \leq m) = 1 - (1-p)^m$, and so $P(U > m) = (1-p)^m$.

Consider now an alternative search Y in which the random variables $Y_1, Y_2, Y_3, ...$ are also iid and where for the corresponding waiting time V, $P(V =$

$m) = q \cdot (1-q)^{m-1}$, with q being considerably larger than p. With each step in the search process, Y has probability q of finding T and X has probability p of finding T. Both searches are guaranteed eventually to locate T because $\lim_{m\to\infty} P(U > m) = \lim_{m\to\infty}(1-p)^m = 0$ and $\lim_{m\to\infty} P(V > m) = \lim_{m\to\infty}(1-q)^m = 0$. (We assume neither p nor q is equal to zero, so $1-p$ and $1-q$ are both strictly less than 1. These limits are therefore 0 because any number less than 1 raised to increasingly large exponential powers becomes arbitrarily close to 0.) Even so, if q is much larger than p, the search Y will on average find T much more quickly than the search X.

How much more quickly? The waiting times U and V follow a geometric distribution for probabilities p and q respectively. Accordingly, $E(U) = 1/p$ and $var(U) = (1-p)/p^2$, and $E(V) = 1/q$ and $var(V) = (1-q)/q^2$.[54] Clearly, if $p << q$ (i.e., if p is a lot less than q), the search X with waiting U represents less intelligence than the search Y with waiting time V to the degree that Y is quicker at finding T than X, or equivalently, to the degree that the average waiting time $E(V) = 1/q$ is smaller than the average waiting time $E(U) = 1/p$. For example, if $p = 1/1,000,000$ and $q = 1/2$, the search Y with waiting time V will on average quickly find the target T in 2 steps, but the search X with waiting time U will on average much more slowly find the target T in 1,000,000 steps. In general, one intelligence is smarter than another not only if it can do things that the other cannot do, but also if it can do the same things faster than another. Speed, for instance, plays a significant role in school aptitude tests, which put a premium on how quickly students are able to answer questions correctly.

The ratio of these waiting times $E(V)/E(U)$, therefore, to the degree that it is small, will indicate greater intelligence at finding T for Y (as gauged by V) as compared to X (as gauged by U). But given that these searches consist of independent and identically distributed (iid) random variables, this ratio is none other than $E(V)/E(U) = (1/q)/(1/p) = p/q$. And if we now put a negative logarithm to the base 2 in front of this ratio, that yields $-log_2[E(V)/E(U)] = -log_2(p/q)$, which is mathematically identical with active information. Note that for $p = 1/1,000,000$ and $q = 1/2$, $-\log_2(p/q) = -\log_2(1/500,000) \approx 19$ bits.

In general, then—and not just for searches consisting of iid random variables—a very natural measure for gauging the relative intelligence of sequential searches $X = X_1, X_2, X_3, ...$ and $Y = Y_1, Y_2, Y_3, ...$ with waiting times U and V respectively is $I_+(Y, X) = -\log_2[E(V)/E(U)] = \log_2[E(U)/E(V)]$. The plus symbol here in I_+ is standard notation for what in the conservation of information literature is called "added information" or "active information" (the two are synonymous). And even though $I_+(Y, X)$ will not coincide precisely with active information

unless the searches are iid, larger values of this number will tend to indicate a greater disparity of intelligence, favoring the intelligence of Y over that of X to the degree that $E(U)$ is larger than $E(V)$.

But note, in the absence of the searches being iid, interpreting $I_+(X,Y) = -\log_2[E(V)/E(U)]$ will require some care and can lead to some counterintuitive results. For instance, we can imagine a search that with probability .999999 locates the target T on the first try but otherwise requires 10^{100} steps to locate the target. The expected time to locate the target is then quite large, namely, $.999999 \cdot 1 + .000001 \cdot 10^{100} \approx 10^{94}$, which is huge. Nonetheless, the search will for the vast majority of the time be instantly successful. On the other hand, a search that is guaranteed to fail for the first $10^{50} - 1$ steps but is guaranteed to succeed on the 10^{50}th step will have a far smaller expected waiting time, namely 10^{50}, suggesting a superior intelligence than the previous search. And yet, unlike the previous search, for most practical purposes this second search will allow no reasonable chance of success.

Of course, the two searches just described are anomalous. They make clear that $I_+(Y,X) = -\log_2[E(V)/E(U)]$ will capture the relative intelligence of X and Y only to the degree that these searches are broadly analogous to iid searches in that they steadily amplify the probability of hitting the target T as the search unfolds (the two searches just described don't do that). A broader category of searches that include iid searches and that promise to capture relative intelligence among searches are Markov processes in which the target T is treated as an absorbing state.[55]

For arbitrary (and not just iid) searches $X = X_1, X_2, X_3, ...$ and $Y = Y_1, Y_2, Y_3, ...$, we noted that the most important probabilities are $P(U \leq m) = P(X_1 \in T \text{ or } X_2 \in T \text{ or } ... \text{ or } X_m \in T) = p_m$ and $P(V \leq m) = P(Y_1 \in T \text{ or } Y_2 \in T \text{ or } ... \text{ or } Y_m \in T) = q_m$. Since these probabilities gauge success at hitting the target, our standard approach to intelligence metrics, with its focus on points of failure, will thus need to be based on the probabilities $P(U > m) = 1 - p_m$ and $P(V > m) = 1 - q_m$.

Suppose now we're given an upper bound M beyond which we would not or could not extend the searches X and Y. That suggests comparing $1 - p_m$ and $1 - q_m$ for $1 \leq m \leq M$ and correspondingly the information values associated with these probabilities, namely, $-log_2(1 - p_m)$ and $-log_2(1 - q_m)$. If success is very likely for one of these searches but very unlikely for another in M steps, then comparing $-log_2(1 - p_M)$ and $-log_2(1 - q_M)$ can be illuminating for these searches. Interpreting such information values for specific searches would then fall under our generic approach to gauging intelligence by measuring the information

associated with failure—the smaller the probability of failure, the greater the measured information in search and so the greater the intelligence in search.

21. Complexity as a Measure of Intelligence

In earlier sections of this paper, measures of intelligence were associated in the first place with agents rather than with targets. Specifically, measures of intelligence represented the performance capabilities of agents in reaching targets rather than in any properties of the targets as such. Thus, when we compared Alice and Bob (both agents) on achieving a certain test score (the target), we asked what the probability of each was for the respective agent to achieve it. The target was relevant to this probability, but the probability ultimately attached to the agents. The greater the probability, the smaller the probability of failure to reach the target, and thus the greater the intelligence of the agents (in line with our convention of having smaller probabilities of failure correspond to information values signifying greater intelligence).

In this section, I want to measure intelligence from the other direction, focusing in the first place on the target itself and attaching a level of complexity to it. Often, we have good reasons to assign a level of complexity or difficulty to solving a problem (or, equivalently, to finding a target—problem solving and target searching being equivalent activities). Often as well, we don't have any prior or independent knowledge of the capacities of agents to solve certain problems. For instance, we may know nothing about how probable it is for an agent to find one's way through a complicated maze. Nonetheless, we may have objective ways of gauging the complexity of the maze and thus the difficulty of navigating it. By gauging the maze's complexity, we can credit an agent's intelligence for successfully getting through it. Measuring the intelligence of agents in such cases will thus hinge on the inherent complexity or difficulty of the problems. Successfully solve a complex problem, and your intelligence gets credited; fail to solve it, and it gets discredited.

Measuring intelligence in terms of complexity requires, unsurprisingly, the ability to measure complexity. Chapter 4 of the first edition of my book *The Design Inference* (1998) was on complexity theory—the mathematics of measuring complexity. There I characterized complexity as a measure of difficulty.[56] I omitted

that chapter from the second edition (2023) of that book because the material was technical and I didn't need the full power of complexity theory for a basic account of design inferential reasoning.[57] Nonetheless, that material is just what's needed for measuring intelligence in terms of complexity. In the first edition, I motivated complexity measures as follows:

> Complexity measures are measures of difficulty, measuring how difficult it is to solve a problem using the resources given. Thus, complexity measures measure such things as cost, time, distance, work, or effort. In measuring how difficult it is to solve a problem using resources, complexity measures assess not the actual contribution the resources have already made toward solving the problem, but what remains to be done in solving the problem once the resources are in hand (e.g., with resources of $100 in hand and a problem of acquiring a total of $1000, the complexity of the problem is not the $100 already available, but the $900 yet needed).[58]

Complexity measures, as I developed them in chapter 4 of the first edition of *The Design Inference*, were thus difficulty measures. In line with the account of complexity measures I develop there (I'll highlight only a few aspects of it in this section), we can define a complexity measure Δ (delta for difficulty) of a problem Q with respect to resources R, denoted by $\Delta(Q|R)$ and called "the complexity of Q given R," as that number, or range of numbers, in the interval $[0, \infty]$ (and thus including ∞), denoting how difficult it is to solve Q under the assumption that R holds or is available, and on the condition that R is as effectively utilized as possible.[59]

Note that because of the interchangeability of solving problems and finding targets, we can treat the problem space in which Q resides as a possibility space Ω for which Q corresponds to a target T within Ω, where finding T is equivalent to solving Q. The resource space from which R is taken could then be subsets of Ω, but it need not be. For instance, Q might be solving a computational problem, R might be different algorithms available to solve it, and Δ might count the shortest number of computational steps to solve the problem. The bigger Δ and the more generous R, the greater the difficulty in solving Q and thus the greater intelligence required to solve it.

We might have used the term "difficulty measure" here for measuring the difficulty that an intelligence has to overcome. Yet in most mathematical discussions, it is common to refer to complexity rather than difficulty measures, and so we stick with that usage. To get some sense of the breadth of complexity measures, consider the following list, which is by no means complete:

1. **Cost complexity** — Measures difficulty by the monetary cost of completing a task.
2. **Time complexity** — Measures difficulty by the time required to solve a problem.
3. **Space complexity** — Measures difficulty by the amount of computer memory required.
4. **Computational-step complexity** — Measures difficulty by the number of elementary computational steps required.
5. **Optimization complexity** — Measures difficulty by the cost, distance, or other quantity that must be minimized to reach an optimum.
6. **Probability-derived complexity** — Measures difficulty by transforming the probability of an outcome so that less probable outcomes count as more difficult.
7. **Information complexity** — Measures difficulty in bits according to the information needed to identify or represent an outcome.
8. **Distance complexity** — Measures difficulty by the distance separating a target from the nearest available starting point or resource.
9. **Physical-work complexity** — Measures difficulty by the physical work required to transform an initial state into a desired state.
10. **Minimum-proof-length complexity** — Measures the difficulty of proving a proposition by the length of its shortest available proof.

From points 6 and 7, it becomes clear that probability can be a disguised form of complexity, with complexity increasing as probabilities become smaller. Given a target T in a search space Ω, the probability $P(T)$ thus correlates inversely with the intelligence of the agent who happens to attain T. In the absence of any explicitly stated resources R, $P(T) = P(T \mid \Omega)$, and so the entire possibility space Ω becomes the resources to be factored into locating T. Note that $P(\Omega) = 1$, which means that its corresponding complexity is the lowest possible in that probability and its corresponding complexity are inversely correlated.

In a probabilistic context, for a subset R of Ω treating R as resources, the search space for finding T is being restricted to R. Resources can thus make finding T more likely, equally likely, or less likely because the conditional probability $P(T \mid R)$ can go up, down, or stay the same as $P(T)$. If the conditional probability $P(T \mid R)$ is strictly less than $P(T)$, then R is misleading (in the literal sense of the word) for finding T. This means, perhaps counterintuitively, that the agent who nonetheless achieves T is in fact demonstrating greater intelligence than one who does not operate with the misleading resources R in trying to locate T.

In treating probability measures as disguised complexity measures, we flip the direction and scaling of the measures so that the unit interval on which probabilities take their values, i.e., $[0, 1]$, map to the entire non-negative real line including infinity, i.e., $[0, \infty]$ such that a probability of 0 maps to ∞ and a probability of 1 maps to 0. We have seen this type of complexity measure repeatedly in previous sections, where it took the form of a Shannon information measure that applied a negative logarithm to the base 2 to probabilities, so that $-\log_2 \mathrm{P}(T) = I(T)$ and $-\log_2 P(T \mid R) = I(T \mid R)$. Accordingly, an agent that finds a target T is assigned a complexity score of $I(T)$ or $I(T \mid R)$ (depending respectively on whether the agent used or didn't use the resources R to find T), which is then taken as a measure of difficulty in attaining T, and which is then in turn credited to the agent as a measure of intelligence.

Note that in this context, $I(T)$ and $I(T \mid R)$ can treated also as measures of active or added information (recall section 18) in which by finding T, the agent is given the benefit of the doubt of being able to attain T with high probability —typically 1.[60] What was previously the probability p is thus equal to $P(T)$ and what was previously the probability q is thus set equal to 1. The active information (synonymously the added information) $-\log_2(p/q)$ is then equal to $-\log_2(p/1) = -\log_2(p)$, which is precisely the complexity-based intelligence metric described. The adjustment to this formalism for conditional probability/ conditional information is straightforward.

One form of complexity notably absent from the discussion in this section is Kolmogorov complexity. Interpreted as a measure of difficulty, it acts quite differently from the complexity measures described in this section. It is therefore deferred till the next section.

22. Kolmogorov Complexity as Specified Complexity

Kolmogorov complexity is able to measure intelligence but not as a measure of difficulty. To measure intelligence in terms of difficulty, as with the complexity measures of the last section, Kolmogorov complexity needs to be adjusted. With

that adjustment, Kolmogorov complexity becomes a special case of specified complexity, which can be reasonably interpreted as measuring intelligence via difficulty.

Kolmogorov complexity makes sense of complexity in terms of how concisely something can be described, greater concision corresponding to less complexity and less concision to greater complexity. The underlying intuition is that a highly regular symbol string can often be generated from a short set of instructions whereas an irregular symbol string may require instructions roughly as long as the string itself (a string is always describable in terms of itself and thus never requires an instruction set substantially longer than itself). Kolmogorov complexity thus measures compressibility or minimum description length.

Already here we see that Kolmogorov complexity does not measure difficulty. Kolmogorov complexity is maximal when a string cannot be compressed and is therefore describable in terms of itself. But describing something in terms of itself is easy—simply exhibit it. The difficulty with Kolmogorov complexity is finding the concise—indeed the most concise—set of instructions to generate a longer string.

As it turns out, concision is difficult, and the greater the concision, the more difficult it is to achieve that level of concision. Hence the quip attributed to Mark Twain: "If I had more time, I would have written a shorter letter." For Kolmogorov complexity, less complexity corresponds to greater difficulty. To get a measure of difficulty from Kolmogorov complexity therefore means flipping what counts as greater and lesser complexity. More on this momentarily.

But first, a few technical details. $K(x)$ denotes the basic measure of *Kolmogorov complexity*. It presupposes a universal computer language U (i.e., a universal Turing machine) and identifies the shortest program in this language that will produce the string x and then halt. The length of that shortest program is then $K(x)$. In general, for any string, whether interpreted as a program or data, the length of that string is denoted by $\ell(x)$. In that case,

$$K(x) = \min_y \{\ell(y) \,|\, U(y) = x\}.$$

More plainly, $K(x)$ is the length of the shortest string y that via U yields x.

For simplicity, and without loss of generality, let's assume that all strings in the language are binary, consisting of zeros and ones (bits). Additionally, let's assume that the language is prefix-free (i.e., no string in the language has another string in the language that extends it). These conditions make it easier to prove theorems about Kolmogorov complexity but don't affect its generality

since mappings from binary prefix-free languages to non-binary and non-prefix-free languages change the underlying complexity measures negligibly (i.e., don't change their basic magnitude).

To illustrate, a string consisting of a million zeros will have low Kolmogorov complexity because a very short program exists to bring it about, one that says in effect "print zero a million times." On the other hand, an irregular million-bit string may have no substantially shorter description than simply listing all its bits in order. The precise numerical value of $K(x)$ depends on the choice of universal computing language U (assuming it's binary and prefix-free), but minimally so because all such versions of Kolmogorov complexity differ only by a fixed additive constant. The basic ordering of simple versus complex strings is therefore robust, and this remains true even if we relax the binary and prefix-free conditions.

Kolmogorov complexity is readily extended to *conditional Kolmogorov complexity*, written $K(x \mid y)$. For it, the question is not how short a program must be to generate x from scratch, but how short the program to produce x can be when y is available to help produce it. Conditional Kolmogorov complexity can then be defined as follows:

$$K(x|y) = \min_z \{\ell(z) \,|\, U(z, y) = x\}.$$

More plainly, $K(x|y)$ is the length of the shortest string z that via U yields x given the prior availability of y.

Here y is treated as a freely available resource. If y contains little that's helpful for producing x within the underlying computer language U, then $K(x \mid y)$ will be nearly as large as $K(x)$. If y contains much of what's needed for the underlying computer language to produce x, the conditional complexity will be substantially smaller. Conditional Kolmogorov complexity therefore captures the additional descriptive information needed to obtain one string once another is already in hand.

In the sequel, we will consider only ordinary, rather than conditional, Kolmogorov complexity. We can do this without loss of generality because any resource y can be folded into a universal Turing language U forming a new universal Turing language U_y, with unconditional Kolmogorov complexity then being defined in terms of U_y—much as taking a probability P on a space Ω and conditioning it on a subset B of Ω forms a new probability $P_B(\cdot) = P(\cdot \mid B)$, which is then a probability in its own right.

As noted, $K(x)$ signifies greater difficulty the smaller $K(x)$ because smaller values of $K(x)$ mirror the successful search for shorter programs capable of generating x under U. The smaller programs make up less and less of the binary

strings on which K is defined. There are 2 binary strings of length 1, 4 or length 2, 8 of length 3, and in general 2^n strings of length n. The prefix-free condition on strings changes these numbers slightly, but the basic magnitudes are the same.

Because K signifies greater difficulty the smaller it is, and because for any x Kolmogorov complexity cannot exceed $\ell(x)$ (except for a positive constant that's independent of the length of x), the quantity $\ell(x) - K(x)$ defines a complexity measure in the sense of the last section. This difference between the length of x and its Kolmogorov complexity therefore defines a difficulty measure, with greater values signifying greater difficulty, lesser values lesser difficulty. This number, except for negligible factors, cannot exceed $\ell(x)$ and will not dip below 0, thus landing in the closed interval $[0, \ell(x)]$.

To see how $\ell(x) - K(x)$ is a measure of difficulty, suppose that the number $K(x)$ is substantially less than $\ell(x)$. Then the uniform probability of all strings of length $K(x)$ or less will be roughly $2^{-K(x)}$, which will be much greater than $2^{-\ell(x)}$. We can then identify the probability of finding a string of probability $K(x)$ with $2^{K(x)}/2^{\ell(x)}$. This probability, then, represents the difficulty of finding a program of length $K(x)$ that generates x, and so, in line with probability as a measure of difficulty (as in the last section), $-\log_2\left[2^{K(x)}/2^{\ell(x)}\right]$ then denotes the active or additive information with finding a program of length $K(x)$.

This number is, by unpacking the logarithm, just $\ell(x) - K(x)$. But it is also just $I(x) - K(x)$, where I is information in the ordinary Shannon sense. That's because the length of a bit string corresponds, under uniform probability, to its Shannon information. It's this last quantity that we define as *specified complexity*, denoted by SC:

$$SC(x) = I(x) - K(x).$$

As Winston Ewert and I show in the second edition of *The Design Inference* (chapter 6), this definition of specified complexity can be extended to targets in search spaces identified by descriptive languages (such as are binary and prefix-free), so that for a target T, SC is defined as

$$SC(T) = I(T) - D(T).$$

Here D is a generalization of Kolmogorov complexity, denoting minimal description length. Moreover, the Shannon information I here can be extended to probabilities in general and need not apply only to uniform probabilities.

SC, so defined, was shown in chapter 6 of the second edition of *The Design Inference* to have the property that the probability of any composite target R, consisting of the union of all targets T satisfying $SC(T) \geq \sigma$, will satisfy $P(R) \leq$

$2^{-\sigma}$. In other words, the totality of targets with specified complexity at least σ will thus have probability not exceeding $2^{-\sigma}$. A given target in this totality will have probability $2^{-I(T)}$, which will typically be smaller than $2^{-\sigma}$. But if the specified complexity is large, σ will be large, and $2^{-\sigma}$ will be small.

Although Kolmogorov complexity is non-computable, making $K(x)$ difficult to apply in practice, $SC(x)$, its corresponding specified complexity, is readily applied in practice. That's because we can often find an upper bound estimate on $K(x)$. Let's say we've found some string w that's reasonably short and for which $U(w) = x$. Then, because $K(x)$ denotes the length of the shortest string capable under U of yielding x, it follows that $K(x) \leq \ell(w)$, and so $SC(x) = I(x) - K(x) \geq I(x) - \ell(w)$. It then follows that if $I(x) - \ell(w)$ is large, so is $SC(x)$—indeed, the latter will be at least as large. So, for practical purposes, solid lower bound estimates on specified complexity are often readily available. And that's true also in more generally when specified complexity is defined in terms of minimal description length

In sum, specified complexity corresponds a measure quantified in bits, which, if bounded below by a value σ, corresponds to a probability less than or equal to $2^{-\sigma}$. Accordingly, specified complexity can constitute an information-based measure of intelligence.

How exactly does specified complexity measure intelligence? Till now in this paper, measuring intelligence has been task-specific. Thus there's a search space and there's a target within it (or a graded set of targets) so that success at the task means hitting the target (or to varying degrees approximating it) and such success (or lack of it) then measures intelligence. This has been the paradigm case for measuring intelligence till now. In it, a prespecified target represents the task and locating it measures intelligence.

With specified complexity, however, no target is given in advance. We observe some outcome. If, then, this outcome can be described concisely (i.e., it has low Kolmogorov complexity or short description length), we conclude that a target has been hit, the target in this case being given not in advance but being there implicitly in virtue of its concise description. Retrospective recognition is legitimate here precisely because of the concise description.

Think of it this way: in prior sections of this paper, what gauged the degree of intelligence of an agent was well-defined before the agent did anything bearing on the agent's degree of intelligence. If you will, the standard of intelligence was clearly stated in advance, and the agent's job was to meet that standard. In practice, this means there's a prespecified target, and then the agent demonstrates intelligence to the degree that the agent expeditiously reaches that target. The

logical order is therefore as follows: the target is explicitly given in advance, the agent acts to reach the target, and then the agent is assessed to be intelligent depending on how precisely and effectively the target is reached.

In practice, however, many targets emerge after the fact. Thus typically, an unanticipated event happens in which we discover a salient pattern (the salience consisting in the pattern's short description). That pattern suggests that some target was attained even though we didn't know about the target before the event that brought it about. The pattern suggests that some agent caused the event exhibiting the pattern, though the agent may be hidden or only partly disclosed. The pattern may further suggest a certain degree of intelligence that the (suspected) agent needed to bring about the event. Specified complexity captures this after-the-fact approach to gauging intelligence.

Think of a detective solving a mystery. The detective identifies various pieces of evidence, suddenly seeing how they fit together to form a coherent picture (a revelatory pattern). In fitting together, the pieces now reveal something easily described (short description length) whereas previously things seemed chaotic. Those pieces coming together, additionally, will tend to be improbable. Specified complexity now becomes evident and captures the moment of insight when the detective achieves clarity about the crime. In this way, specified complexity provides a measure of intelligence.

How good a measure of intelligence is it? Specified complexity is less a measure of degree of intelligence than a measure of the degree of confidence we may have in affirming that something that exhibits specified complexity is indeed due to intelligence. The greater the specified complexity, the more implausible it is that something exhibiting it is the result of anything other than intelligence.

Thus a detective may investigate two deaths. To the extent that specified complexity can be quantified for these deaths, it may be clear that both were the result not of natural causes but of foul play, which is to say, they were murders and thus by design. Specified complexity might here be high and the same for both murders. But one of the murders may be much more clever than the other. In one case, suspicion may fall on the actual murderer. In the other, it may fall on someone innocent, letting the murderer off scot-free. Specified complexity, without further investigation, will not capture this difference in intelligence.

In measuring intelligence, specified complexity can provide a yes-no measure of whether intelligence was involved, though as just seen it may have more difficulty discriminating between degrees of intelligence. However, specified complexity can also under the right conditions tell us whether a certain level of intelligence is sufficient to bring about a certain outcome. The detective may therefore rule out

that a child or animal might have been able (i.e., intelligent enough) to pull off a given crime. High specified complexity that is highly unlikely to be overcome by a given cause or agent suggests a different cause or agent.

23. The Intelligence of Natural Selection

How intelligent is natural selection? This question raises a prior question: What does it even mean to ascribe intelligence to natural selection and how, if at all, should its intelligence be measured. As it is, our approach to measuring intelligence through information and probability matches up neatly with what Darwinians themselves have said about natural selection acting as a probability amplifier, taking otherwise wildly improbable things and making them plausibly probable.

The most striking statement of this view that I have found in the Darwinian literature comes from R. A. Fisher (1890–1962). Fisher was at once the greatest statistician of the 20th century, a geneticist, a chief architect of the neo-Darwinian synthesis, a devout Anglican, and a fervent eugenicist (seeing eugenics as a religious duty to better humanity). In a chapter for an anthology on evolution (*Evolution as a Process*, 1954), edited by Julian Huxley among others, Fisher wrote:

> [I]t was Darwin's chief contribution, not only to Biology but to the whole of natural science, to have brought to light a process by which contingencies a priori improbable, are given, in the process of time, an increasing probability, until it is their non-occurrence rather than their occurrence which becomes highly improbable.[61]

This is a truly staggering statement. To the degree that intelligence is understood and measured in terms of probabilities, this would make it Darwin's chief contribution as providing a complete account of intelligence not just in biology but across all the natural sciences. That would be quite the feat! And just to be clear, the referent of Darwin's chief contribution here is natural selection. Indeed, lest one think that Fisher is describing here something other than natural selection, Julian Huxley, in his introduction to *Evolution as a Process*, quotes Fisher as

saying, "Natural selection is a mechanism for generating an exceedingly high degree of improbability."[62] The overlap with the earlier quote is unmistakable.

Having quoted Fisher on natural selection acting as a probability amplifier, Huxley immediately expands on what this means for biology. Invoking H. J. Muller, Huxley writes:

> We can now … make some quantitative estimate of that degree [of improbability that natural selection is able to generate]. Muller has calculated that the most conservative odds against a higher organism, such as a man, a mammal, or even a fruit-fly, coming into existence fortuitously, without the operation of selection, by the union in one stock of all the necessary mutations, are given by a number with so many noughts that it would take an average novel to write it out, a number immensely greater than that of all the electrons and protons in the visible universe. That is a measure of the degree of our own inherent improbability—an improbability of the same order of magnitude as that of a monkey with a typewriter producing the works of Shakespeare. Just as it took the conscious activity of an outstanding human mind to produce the one, so it took two thousand million years of natural selection to generate the other.[63]

It's fascinating here to see this juxtaposition of Shakespeare—who in his plays and poems provides as clear a demonstration of purpose and design as exists—with natural selection—which for Darwinists is the singular creative force that calls into being the great complexity and diversity of life. No wonder that natural selection is viewed by Darwinists as the great designer substitute.

Two decades earlier, in 1936, in an article in *Nature*, Julian Huxley explained exactly how natural selection acts as a probability amplifier and thus as a designer substitute:

> We must invoke natural selection whenever an adaptive structure involves a number of separate steps for its origin. A one-character, single-step adaptation might clearly be the result of mutation. But when several or many steps are involved, it becomes inconceivable that they shall have originated simultaneously. The improbability is therefore enormous that they can have arisen without the operation of some agency which can gradually accumulate and combine a number of contributory changes: and natural selection is the only such agency that we know. Natural selection achieves its results by giving probability to combinations which would otherwise be in the highest degree improbable.[64]

Huxley here omits an obvious possibility. What if multiple steps are needed for some feature to evolve but they need to originate simultaneously before natural selection can act, there being no selective advantage for fewer steps? This is the point of Michael Behe's notion of irreducible complexity: for irreducibly complex systems, several steps need to originate simultaneously, natural selection cannot (except with vast improbability) achieve such simultaneous origination, and there is indeed another agency we know about that is up to the task, namely, real teleological agency, or design.[65] Nonetheless, if we go with Huxley here, it's straightforward to see what he has in mind in how natural selection amplifies probabilities and in how an intelligence metric might reasonably be assigned to natural selection.

Consider the case of antibiotic resistance. I'll offer a stylized example that captures the essence of what Huxley is saying here, with the advantage that the numbers are readily calculable.[66] Let's imagine we have a million bacteria. Each gets to flip a fair coin twenty times. If it gets 20 heads in a row, it gets rewarded with a mutation that lets it escape death when doused by a given concentration of an antibiotic.

Those that escape death in the first round are allowed to reproduce until they reach a population of a million, at which point they are again handed a fair coin and get to flip it twenty times. If any bacterium does flip twenty heads in a row this second round, it gets rewarded with a second beneficial mutation that allows it to survive now a 100-fold concentration of the antibiotic. And so on for five rounds. Any bacterium that has survived after five rounds now has five mutations that allow it to withstand a 100,000-fold concentration of the antibiotic.

What, then, is the probability that a bacterium evolves via this selection scheme after five rounds? The probability of at least one bacterium out of a million flipping 20 heads in a row in any round is .6147. The probability that some bacterium will survive all five rounds, acquiring five mutations that confer a 100,000-fold increase in antibiotic resistance, is then .0877 (i.e., .6147 raised to the 5th power), or roughly 9 percent.

This is not a huge probability, but it is far greater than the probability of getting five mutations without selection. To see this, consider what would have happened if that original population were doused with a 100,000-fold concentration right from the start. The probability of some bacterium spontaneously acquiring the five mutations needed to achieve this level of antibiotic resistance would then have been equivalent to one of these million bacteria flipping 100 heads in a row in a single trial (5 rounds of 20 heads in a row all by the same population of a million bacteria). That probability would be 1 in 1.27×10^{24}, or

in decimal notation .00000000000000000000000007874, which is a decimal with 24 leading zeros.

The relative intelligence of pure chance versus selection is now readily calculated in terms of the various intelligence metrics laid out in the previous sections. For simplicity, let's just go with active information, as defined in section 18. Let $p = 1/(1.27 \times 10^{24})$ (this is the probability of getting the 5 mutations all at once by pure chance) and let $q = .0877$ (this is the probability of getting the 5 mutations incrementally by selection). The active information is then $\log_2(q/p) = 76.56$ bits, which is substantial. It indicates that selection is far more intelligent at bringing about the five mutations than pure chance.

The astute reader will note that I dropped the adjective "natural" in front of "selection" here. Strictly speaking, the selection in this example was artificial rather than natural. But I'm happy to concede that artificial selection here can mirror natural selection in the wild as it responds to antibiotics. Consequently, I'm happy also to concede that natural selection in such cases is far more intelligent than pure chance in achieving given levels of antibiotic resistance (especially if we raise the population size from a million to, say, a billion—that would make it virtually assured that antibiotic resistance will develop by selection, but pure chance will still be grossly ineffective).

We need now to remind ourselves of a point stressed throughout this paper: intelligence is task specific and measuring intelligence is a matter of gauging success/failure at achieving specific tasks. Consequently, there is no reason to think that natural selection, even though successful at certain tasks, needs to be universally successful at bridging all evolutionary transitions. It could simply be a fact of biochemistry that certain evolutionary transitions require multiple coordinated mutations. Michael Behe makes that claim for irreducibly complex biochemical systems, as he argued in *Darwin's Black Box*.[67] Likewise, he makes that claim for chloroquine resistance in the malarial parasite *Plasmodium falciparum*, which requires 10^{20} parasites for the multiple specific mutations needed to produce resistance, as he argued in *The Edge of Evolution*.[68]

It is not my aim here to retry Behe's case for limiting the power of natural selection. My point, rather, is to lay out the logic by which natural selection can be assigned a measure of intelligence that tracks its success or failure at effecting certain evolutionary transitions. Obviously, any legitimate measure of intelligence must be able to capture both success and failure—it can't just be an a priori truth that such a measure must always and in every way promote the power of natural selection. Any legitimate intelligence metric will measure degrees of intelligence

and at the extremes there will be measures of intelligence that signify adequacy as well as inadequacy for the task at hand.

Mathematician Jason Rosenhouse, in his 2022 critique of probabilistic arguments that attempt to refute Darwinian evolution, rejects the possibility that natural selection may ever be inadequate for effecting certain evolutionary transitions, always acting instead as a fully adequate probability amplifier. In this, he parallels Julian Huxley and R. A. Fisher, who likewise saw natural selection in this role as a universally capable probability amplifier. Writing specifically about the evolution of "complex biological adaptations," Rosenhouse claimed, "Either the adaptation can be broken down into small mutational steps or it cannot. Evolutionists say that all adaptations studied to date can be so broken down while antievolutionists deny this."[69]

Again, it's not my concern here to rehearse the arguments made by critics of neo-Darwinism such as Michael Behe who claim adaptations exist that cannot be broken into small incremental mutational steps subject to natural selection. But I do want to point out that wherever one lands on this question, it is illegitimate to argue that when pure chance fails, then natural selection is the only game in town, and so it must be able to raise the probabilities in question even in the absence of evidence for this raising of probabilities and even when there exist good empirical and theoretical arguments that multiple coordinated changes rather that sequentially advantageous changes are needed.

At the famous 1966 Wistar symposium titled *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*, mathematicians Marcel Schützenberger and Stanislaw Ulam, along with engineer Murray Eden, critiqued the mathematical adequacy of neo-Darwinism in accounting for the full sweep of evolution. In responding specifically to the critique of Ulam, Harvard evolutionist Ernst Mayr remarked: "We have so much variation in all of these things that somehow or other by adjusting these figures we will come out all right. We are comforted by knowing that evolution has occurred."[70]

This will not do. It's not enough to say we know that evolution has occurred —by natural selection no less—so it doesn't matter what the math tells us. Specifically, it's not enough to say that we know pure chance can't get the job done so it must be natural selection that's getting the job done, thereby automatically conferring on natural selection a high probability of effecting a desired and needed evolutionary transition and thus assigning to it a substantial measure of intelligence. Such an argument commits the bifurcation fallacy (aka false dilemma), namely, the error of presenting two alternatives as the only possibilities when at least one other alternative exists. Obviously, design is such an alternative.

With any task and any process supposedly capable of achieving the task, it is always legitimate to ask what the probability is for the process to achieve the task, and this is true even if estimating such a probability is easy, hard, or infeasible. With natural selection, therefore, there is no need to position it reflexively against the foil of pure chance. Rather, instances exist where natural selection may be considered on its own terms and its probabilistic properties can be evaluated directly.

Molecular biologist Douglas Axe's work on the evolution of beta-lactamase is a case in point.[71] He established a hydropathic signature for this enzyme and then showed that evolving into an enzyme with that signature from some protein with a different function, as must happen if beta-lactamase evolved by natural selection, has a probability in the range of 10^{-64} to 10^{-77}. The point to note is that these improbabilities are not for pure chance but for natural selection. Even on the high end in terms of probabilities, Axe estimated that natural selection's failure rate at 10^{-64}. In terms of the intelligence metrics associated with complexity measures, as in section 21, this suggests that an intelligence scoring at least $-\log_2(10^{-64}) = 212.60\,\text{bits}$ is required to achieve a functional beta-lactamase.

But note, this measure of intelligence is *not* assigned to natural selection. Rather, it measures the intelligence needed, *over and above* natural selection, to achieve a functional beta-lactamase. Just as in some circumstances, natural selection may be needed to surmount pure chance, so in others natural selection may itself need to be surmounted. As in section 21, when a complexity measure assigns a high degree of difficulty to a problem or search, when the problem is solved or the search is successfully completed, the presumption is that the cause or agent that solved the problem or completed the search had the capacity to overcome this level of difficulty and that this level of difficulty can therefore be ascribed to that cause or agent as a measure of intelligence.

Accordingly, in the terminology of this paper, Douglas Axe's research purported to show that $-\log_2(10^{-64}) = 212.60\,\text{bits}$ sets a lower bound on the degree of intelligence beyond natural selection required to achieve a functional beta-lactamase. The underlying measure of intelligence here is active information (see section 18). But note, the big caveat in all this is the one stated earlier in this paper (recall section 10): Intelligence should be credited to an agent or process only in the absence of unfair privileges or advantages contributing to its success and only to the degree that its own capacities account for reaching the target.

In wrapping up this section, let me be clear that I do not claim here to have demonstrated the inadequacy of natural selection or any limitations on its intelligence. For that, the reader needs to consult the wider literature both in

defense of and critical of neo-Darwinism.[72] My point here, rather, is to lay out the logic by which natural selection can claim to have its intelligence measured, and then to ensure that natural selection is given no special privileges over any other processes whose intelligence is subject to measurement.

24. Why LLMs Aren't Smart—Lean Versus Laden Intelligence

The title of this section is deliberately provocative. LLMs have performed spectacularly across a wide range of tasks, and the intelligence metrics described in this paper are capable of representing that performance. Consider the 2026 International Mathematical Olympiad (IMO). This is a highly challenging math test, especially for the targeted age group (under 20). The "dots model" got a perfect score, placing its performance with 7 humans who also got a perfect score out of a total of 666 test takers, putting this model in the 99th percentile of performers.[73] On many standardized tests (ACT, SAT, LSAT, AP, etc.), LLMs perform well above the 90th percentile and even at the 99th percentile. One of the main obstacles to them acing standardized tests is their having to make sense of visuals (thus losing their linguistic advantage). The intelligence of LLMs for such tests is readily measured in terms of percentile performance and failure functions laid out in section 15.

Why then do I write that these models aren't smart? When the intelligence/performance of these models is assessed, the main issue till now has been their degree of success in completing the task in question, especially in comparisons among LLMs and with humans. Ignored has been the extent of resources available to these models in achieving their success. Intelligence is not just about performance on a task in absolute terms but on performance in relation to the resources one has been given to perform the task. In absolute terms, two students may perform identically on the same exam. But if one student gets to take the exam with full access to the Internet and the other must take it purely from memory, we would rightly judge the latter student as the more intelligent on the exam. Intelligence is not just about absolute performance but also about doing more with less.

LLMs are intelligent, but theirs is a *laden* intelligence. To achieve their impressive capabilities, they must ingest enormous corpora of text and employ

immense computational resources. Human intelligence lacks such big-data advantages, and may instead be described as *lean*. A child encounters only a tiny fraction of an LLM's training data. Its learning context is far more open-ended and unstructured, leaving it far more places to focus attention. Yet the child acquires language, concepts, causal understanding, and the ability to reason with remarkable efficiency. Noam Chomsky captured this efficiency of human intelligence with his memorable phrase "poverty of the stimulus."[74] We do amazing things with the little we've been given, especially in comparison to LLMs. There are two crucial distinctions here:

- The distinction between greater and lesser intelligence.
- The distinction between intelligence operating with different informational endowments.

LLMs are laden intelligences. Humans are lean intelligences.

This paper has dealt at length with the first of these distinctions. The question is how to deal with the second. Specifically, how should an intelligence metric factor in the distinction between lean and laden intelligences? While I don't have a precise answer to this question, I can offer some insight into what any such metric needs to contain. To set the stage, I'll review a metric that attempts to redress a similar problem.

Some years back, I helped to develop an influence metric for Academic Influence (with website AcademicInfluence.com). The metric used various online resources, from citation indexes to Wikipedia, to gauge the influence of individual academics. With influence scores for academics in hand, it was then straightforward to calculate influence scores for institutions with which the academics were affiliated. Specifically, influence scores for academics were cumulated additively to assign influence scores to institutions. The approached worked nicely, passing the relevant sanity checks. The academics you would expect to score high on influence did, as did the institutions you would expect to score high on influence.

Nonetheless, when it came to influence by institution, we noticed a certain unfairness in the rankings. Smaller schools might have an illustrious faculty, and yet their very smallness would limit their total influence. Large schools, on the other hand, might likewise house illustrious faculty members, and yet these faculty might be diluted among less impressive faculty. Sheer size of institutions thus seemed like an unfair advantage when trying to understand their influence, with quantity undercutting quality.

We therefore invented a metric called "concentrated influence." In its simplest form, we divided the overall influence score of a school by the total size of the undergraduate body. This provided a measure of how readily undergraduates would get to rub shoulders with influential faculty. In terms of raw influence, Harvard always scored (and rightly so) at the very top. But Harvard has 7,000 undergraduates. Caltech, by contrast, has fewer than 1,000 undergraduates. Dividing its total influence by 1,000 yielded a higher influence score for Caltech than dividing Harvard's by 7,000. Caltech has thus consistently exhibited the highest concentrated influence of any school at Academic Influence.

A similar move can be made in measuring intelligence, adjusting intelligence measures by the resources available to intelligences. For the complexity measures of section 21 that incorporate resources, and where increased resources lead to improved performance, the size of the resources thus becomes a controlling factor in measuring intelligence. Note that not all intelligence measures produce better performance with more resources. It can happen that resources become so large and unwieldy that they cannot be reasonably searched and thus lead to a diminution in intelligent performance. For instance, think of the discovery phase in a trial where an attorney asks for some information but it is buried in a thousand bankers boxes of almost entirely irrelevant information, and so finding the crucial information becomes virtually impossible. This technique is endearingly called a "document dump" or "discovery dump."[75]

In cases where increased resources do increase intelligent performance (as with LLMs), intelligence metrics can be conditioned on available resources. Let's say Alice and Bob attempt the same task requiring intelligence, that Alice works with resources R_A and Bob with resources R_B. Let's say I_A is a metric that characterizes Alice's intelligent performance on the task and I_B is a metric that characterizes Bob's intelligent performance on the same task. R_A and R_B are thus in some sense embedded in these intelligence metrics, and yet the metrics I_A and I_B will favor Alice over Bob if the size of R_A is substantially bigger than the size of R_B and I_A as a result is bigger than I_B.

What's needed, then, is some way of adjusting I_A and I_B in light of the size of R_A and R_B. For ease of notation, let's denote the size of R_A by $|R_A|$ and the size of R_B and $|R_B|$. The larger $|R_A|$ is in proportion to $|R_B|$, the more I_B needs to be adjusted up in relation to I_A. Exactly how this adjustment is to be effected will require understanding the task at hand and the intelligence required to perform well at it. If we follow the example of concentrated influence from Academic Influence, then the intelligence metrics for Alice and Bob would respectively be

$I_A/|R_A|$ and $I_B/|R_B|$, which, to continue the analogy with Academic Influence, might then be called *concentrated intelligence metrics*.

None of this is written in stone or meant to characterize the role of resources in adjusting intelligence metrics or to lay out once and for all the distinction between lean and laden intelligences. That's for another paper, research project, or even dissertation. But enough is said here to get the ball rolling.

25. Measuring Intelligence in Information Packing

Imagine a protein-coding gene that already does its job, producing the exact protein required. Now write a second message into that same DNA without changing the protein. No extra stretch of DNA gets added. Instead, the message is fitted into choices among synonymous codons—different DNA spellings that yield the same amino acids. Researchers have built precisely such watermarked genes.[76] The sequence, constrained by one task, now meets another without giving up the first. This can be described as layering: adding a new layer of information while preserving what is already there.

The watermarked gene just described exhibits a widely used engineering technique. Digital data embedding technologies (DDET) place new information inside a host that must retain its original function.[77] It includes steganography, watermarking, fingerprinting, and related techniques. An image, audio file, video, computer program, or DNA sequence keeps doing its original job while taking on an additional job. Steganography adds a layer of information meant to be concealed. Watermarking adds a layer of information meant to be ineradicable. Fingerprinting adds a layer of information to each distributed copy so that later redistribution is traceable. These examples exemplify information packing.

Four definitions about how information is arranged on a substrate will concern us in this section, and it will be helpful to lay them out early on in this section:

1. **Packing** denotes the degree to which information is concentrated within a finite substrate, with packing ranging from loose to tight.

2. **Density** is the fraction of a finite substrate in functional use, thus characterizing how fully the substrate is occupied by parts performing functional work.
3. **Layering** occurs when a single substrate satisfies two or more nonredundant functional requirements, whether those requirements are carried by different parts of the substrate or by the same parts.
4. **Nesting** is the progressive narrowing of the set of acceptable possibilities as each additional nonredundant functional requirement must be satisfied together with those already imposed.

Packing is the overarching concept here. The key idea is that a substrate can be loosely or tightly packed depending on how much information it carries. Three subsidiary concepts (density, layering, and nesting) then characterize the way the substrate can carry information. Intelligence enters the discussion because greater information packing entails, other things being equal, greater intelligence to bring about the packing observed. Ever more efficient packing suggests ever more search difficulty in finding such packings, and thus a need for greater intelligence.

In these definitions, a *functional requirement* is a condition that must be met for the relevant function to obtain or be successfully achieved. The term does not imply that an intelligent agent formulated the requirement in advance. Also, even though the substrates we consider in this section are physical, purely mathematical substrates are possible and can often be used to represent physical substrates, as with the tape of a Turing machine.

The loosest form of packing is unused space or occupancy. Density in that case tells us how much of the substrate does anything functional at all.[78] Density is central to the debate over junk DNA. In that debate, information density denotes the fraction of the genome exhibiting function. Estimates of this density have risen sharply in the last 25 years. In the early 2000s, only about 1.5 percent of the human genome was seen as protein-coding, and comparison of the human and mouse genomes suggested that roughly 5 percent of the mammalian genome had been conserved by natural selection and therefore likely served some biological function.[79] A 2005 review accordingly described about 5 percent of the mammalian genome as recognizably functional.[80] That would yield a density of 5 percent.

In 2012, however, the results from ENCODE became public. ENCODE asked how much DNA shows evidence of biochemical activity—transcription, protein binding, chromatin modification, and related events. ENCODE reported that 80.4 percent of the human genome participates in at least one such biochemical event in at least one assayed cell type.[81] It thus showed that the old picture of a genome

consisting mostly of inert filler was too simple. Instead, it provided evidence of genomic functional density as high as 80 percent.

A 2024 comparison of 239 primate genomes found roughly 5 percent in strongly conserved *phastCons* elements, that is, stretches of DNA preserved across primate evolution to a degree suggesting that changes there have been selectively constrained.[82] But the measure here is of a stringent form of evolutionary conservation, not total genomic function. Taken together, these results shift the question of information density from "How little of the genome does anything?" to "How close does functional occupancy come to filling out the genome?" Critics have objected that biochemical activity is not the same as organism-level function or evolutionary constraint.[83] But with evidence of biochemical activity extending to over four-fifths of the human genome, it is no longer possible to dismiss widescale function out of hand as it was in "the old junk DNA days."

But note, even information density at 100 percent does not tell us the full story of how tightly packed the information is. For instance, 100 percent occupancy or density of the genome with function would not tell us about its layering of information. Thus multiple distinct functional requirements may not only occupy 100 percent of the same underlying substrate but also over-occupy it. Heteropalindromes—words that are meaningful when read backwards as well as forward, such as *stressed* and *desserts*—illustrate this possibility. DNA does this in practice. Thus protein-coding regions can also carry splicing signals, regulatory instructions, RNA-structural constraints, microRNA targets, developmental enhancers, or another coding sequence.

Itzkovitz, Hodis, and Segal found widespread overlapping signals in protein-coding regions across more than 700 species.[84] Lin and colleagues found more than 10,000 strongly constrained regions in more than a quarter of human protein-coding genes, with evidence for several kinds of additional function.[85] As it is, double-stranded DNA offers six possible translational reading frames, three in each direction.[86] Six reading frames might not mean six times the ordinary Shannon information per nucleotide. But it would mean a single physical sequence can satisfy several distinct functional requirements at once. Such layering allows for tightened information packing beyond what densities of 100 percent can tell us.

Neither density nor layering, however, tell us how difficult information packing is to achieve. This is where nesting comes in. Suppose many DNA sequences perform the same function. Now add a second functional requirement. Some of those sequences now get ruled out. Add a third functional requirement and the set of DNA sequences that satisfies all three shrinks again. Each functional requirement further constrains the target. What we get, then, is a nested sequence

of targets where targets associated with more functional requirements are nested inside those with fewer functional requirements.

If nearly every sequence that performs the first function also performs the second, the added information costs little in terms of Shannon information. If only one in a million does, it costs a lot. Nesting tracks the tightening of constraints as functional requirements are added and the remaining choices are narrowed. The Shannon information associated with the fraction that survives each added functional requirement is then readily calculated. Moreover, this information measure can then be readily interpreted as how difficult it is to achieve a given degree of information packing, thus yielding an intelligence metric.

For an oversimplified example of difficulty in information packing that yields an intelligence metric, consider a 1,000-nucleotide sequence. Suppose that all 1,000 nucleotides contribute to a protein-coding function, that 200 also serve a regulatory function, and that 50 also help determine RNA structure (such as in frameshifting or splicing rather than amino-acid identity). The sequence then averages 1.25 functional uses per position. That describes dense, layered packing, but it says nothing by itself about how hard the arrangement of such a nucleotide sequence was to obtain.

Suppose only one protein-coding sequence in a hundred also meets the regulatory requirement, and only one of every sixteen of those also meets the structural requirement. Then only one in 1,600 sequences that clear the first hurdle clear all three. In terms of Shannon information, that constriction of possibilities comes to about 10.6 bits of nesting information. Another 1,000-nucleotide sequence could exhibit the same physical density and layering, yet make the joint requirements highly probable or fantastically improbable. Calculating the relevant probabilities and Shannon information, and thus an associated intelligence measure, then comes down to precisely how the information is nested.

In a search for packed information, once a target or nested sequence of targets is identified, information metrics developed earlier in this paper apply. Active information would compare searches for the same target and ask how much one raises the chance of success. Performance and failure functions would measure the probability with which a packed informational arrangement is attained. Waiting-times would measure how long success is expected to take. Specified complexity would measure the degree to which a highly improbable packing of information also has a concise description. Such metrics assess how much information is concentrated in a substrate, how many functional requirements are layered onto it, how tightly their joint satisfaction constrains possibilities, and how effectively a process discovers tightly packed information given available resources.

I close with a striking recent example of information packing. Under Article 50(2) of the European Union's Artificial Intelligence Act, services that provide synthetic text must mark outputs as machine-generated.[87] A method that does this is Google DeepMind's SynthID-Text, and it is used in Anthropic's Claude and Google's Gemini.[88] It acts invisibly, slightly skewing the randomness used when a model chooses among near-equiprobable next tokens, thereby interleaving a keyed pattern through word choices while as much as possible preserving meaning and fluency.

The AI-generator's original task of delivering requested text is thus maintained but an identifiable watermark is layered onto the outputted token sequence. That layer maps a secret key onto an existing sampling distribution, resulting in a nesting of information. The measurable intelligence here is not only the trained model that generates text but also the data embedding technology that packs a watermark or steganographic signal into textual choices otherwise treated as interchangeable.

Information packing in the form of density, layering, and nesting thus arise from DNA to generated text. Recurring features here include a single substrate, low or high vacancy among occupied possibilities, multiple nonredundant requirements, and a constriction among allowable strings. Together, these features induce intelligence metrics across both natural and artificial environments.

26. Conclusion

The theory of intelligence metrics developed in this paper shifts the conversation about intelligence from *whether* intelligence acted to *how much* intelligence acted. It grounds the measurement of intelligence in the mathematics of information, describing intuitively and showing formally how information and intelligence interact.

The theory developed here attempts to provide a universal lens for measuring intelligence in both the natural and artificial world. It quantifies the informational cost of intelligent performance. As artificial systems grow in complexity and biological systems reveal ever deeper layers of structure and function, measuring their intelligence becomes an important question, even an urgent one.

As I write this, self-sovereign AI agents are in the news for threatening to go rogue in ways that promise to dwarf the computer viruses of the past. How intelligent are such agents and what level of intelligence is needed to rein them in? The present theory equips science with tools for rigorously, objectively, and quantitatively assessing the capabilities of intelligent agents.

The metrics developed here work equally well whether measuring the intelligence of human problem-solvers, artificial systems, evolutionary processes, or any other causal powers capable of navigating possibility spaces toward independently identifiable targets. The present theory cashes out intelligent performance in terms of successful search. I argue this connection is true universally, with search providing the bridge between intelligence and information.

Active information and conservation-of-information results, as described in the paper, bring this bridge into clear view, measuring how many bits of information a process must contribute to raise a search's probability of success. Based as it is on search, the theory laid out here is metaphysically neutral. Yet precisely because it is based on search, it has widespread practical applications for assessing an agent's degree of intelligence.

The present theory focuses on failure rather than success in measuring intelligence. Tests where everyone scores perfectly convey no information. To discriminate among intelligent performance, tests must reveal points of failure. Such an approach to intelligence mirrors the key insight about information, namely, that it is fundamentally about narrowing and thus eliminating possibilities. Information is about saying no. Intelligence is measured by determining where and to what extent it is limited or absent.

A key distinction, developed late in this paper, is between laden and lean intelligence. A system that achieves impressive results only after consuming vast quantities of training data and computational resources exhibits what may be called laden intelligence. A child acquiring language with minimal input demonstrates lean intelligence. Raw performance alone therefore may tell us little about genuine intelligence. What matters is what an agent can achieve with what it has been given. Doing more with less is a mark of intelligence.

Information packing—the heaping up of nonredundant functional requirements onto a single substrate—adds another dimension to our understanding of intelligence. When a DNA sequence simultaneously codes for a protein, regulates gene expression, or contributes to RNA structure, the information of that sequence far exceeds what a simple count of nucleotides would suggest. The present theory assigns intelligence metrics to such information packing.

What is the intelligence behind information-rich biological structures? The present theory does not prejudge such an intelligence and does not require it to be a conscious mental agent. The theory allows agents to be natural processes acting without foresight, such as natural selection acting on random variations. But the theory would then also measure the intelligence of any such process, along with determining whether that intelligence is adequate for achieving the performance in question.

Specified complexity emerges here as a diagnostic tool for intelligence. When we observe an outcome that is both highly improbable and concisely describable —thus exhibiting specified complexity—we are not only warranted in inferring intelligence but also accorded a quantitative measure of how much intelligence was required.

The practical applications of this theory extend across disciplines, whether to watermarked genomes, Elo chess ratings, SETI technosignatures, forensic evidence of crimes, steganographic embeddings in AI-generated text, or the ability of AI models to prove mathematical theorems. Notably, with the metrics developed here, biologists can evaluate the intelligence of evolutionary processes objectively, without begging the question of whether the processes can or cannot produce the evolutionary outcomes in question.

This theory provides a coherent mathematical framework for making sense of intelligence in its varied manifestations—human, artificial, natural, or some combination of these. By assessing the performance of agents in terms of the search spaces they navigate, we gain a unified standard for measuring their intelligence.

The present theory offers a principled way to clarify the boundaries of cognition, evolution, and technology in relation to intelligence. It makes possible a rigorous accounting of the informational cost of intelligent performance. In consequence, it provides a foundation for future inquiries into the intelligence of the agents, processes, and complex systems that increasingly populate our world.

Notes

1. Introduction

1. See respectively Colin G. Aitken, Franco Taroni, and Silvia Bozza, *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd ed. (Chichester: John Wiley & Sons, 2021); Patrick M. Lubinski, Karisa Terry, and Patrick T. McCutcheon, "Comparative Methods for Distinguishing Flakes from Geofacts: A Case Study from the Wenas Creek Mammoth Site," *Journal of Archaeological Science* 52 (2014): 308–320; and Jason T. Wright, "Strategies and Advice for the Search for Extraterrestrial Intelligence," *Acta Astronautica* 188 (2021): 203–214.
2. For instance, in 1858 Felice Beato photographed a fortified villa in India that British troops had stormed. The photo taken of the aftermath showed scattered skeletal remains. To intensify the scene of massacre in the photo, it appears the bones were disinterred and rearranged. See Luke Gartlan, "Felice Beato," in *Encyclopedia of Nineteenth-Century Photography*, ed. John Hannavy (New York: Routledge, 2008), 128.
3. William A. Dembski and Winston Ewert, *The Design Inference: Eliminating Chance Through Small Probabilities*, 2nd ed. (Seattle: Discovery Institute Press, 2023).
4. David Hume, *Dialogues Concerning Natural Religion*, D. Coleman, ed. (Cambridge: Cambridge University Press, 2007), 43.
5. Combined with the claim that intelligence is unequally distributed in the population and essential to career success, and you have a recipe for dismissing the measurement of intelligence as a discriminatory practice, which arguably is a principal achievement of Richard J. Herrnstein and Charles Murray, *The Bell Curve: Intelligence and Class Structure in American Life* (New York: Free Press, 1994).
6. Howard Gardner, *Multiple Intelligences: New Horizons in Theory and Practice*, revised and updated, (New York: Basic Books, 2006).
7. Daniel Coyle recounts the work of neuroscientists on how skill development and mastery are closely linked to the process of myelination —the formation of a fatty sheath (myelin) around nerve fibers—which enhances the speed and accuracy of neural signal transmission. He illustrates how deliberate practice leads to the growth of myelin, thereby strengthening neural circuits associated with the learning specific skills.

See Daniel Coyle, *The Talent Code: Greatness Isn't Born. It's Grown. Here's How* (New York: Bantam Books, 2009). Conversely, demyelination is associated with cognitive dysfunction and loss of intelligence. See Océane Mercier, Pascale P. Quilichini, Karine Magalon, Florian Gil, Antoine Ghestem, Fabrice Richard, Thomas Boudier, Myriam Cayre, and Pascale Durbec, "Transient Demyelination Causes Long-Term Cognitive Impairment, Myelin Alteration and Network Synchrony Defects," *Glia* 72(5) (2024): 960–981.

8. Michael J. Behe, *A Mousetrap for Darwin* (Seattle: Discovery Institute Press, 2020).
9. Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds and Machines* 17, no. 4 (2007): 391–444, esp. 414, 418, 422; Samuel Allen Alexander and Marcus Hutter, "Reward-Punishment Symmetric Universal Intelligence," in *Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings*, ed. Ben Goertzel, Matthew Iklé, and Alexey Potapov, Lecture Notes in Computer Science 13154 (Cham: Springer, 2022), 1–10; and James T. Oswald, Thomas M. Ferguson, and Selmer Bringsjord, "A Universal Intelligence Measure for Arithmetical Uncomputable Environments," in *Artificial General Intelligence: 17th International Conference, AGI 2024, Seattle, WA, USA, August 13–16, 2024, Proceedings*, ed. Kristinn R. Thórisson, Peter Isaev, and Arash Sheikhlar, Lecture Notes in Computer Science 14951 (Cham: Springer, 2024), 134–144.
10. François Chollet, "On the Measure of Intelligence," arXiv:1911.01547 (2019).
11. José Hernández-Orallo, Bao Sheng Loe, Lucy Cheke, Fernando Martínez-Plumed, and Seán Ó hÉigeartaigh, "General Intelligence Disentangled via a Generality Metric for Natural and Artificial Intelligence," *Scientific Reports* 11 (2021): 22822.
12. Fernando Martínez-Plumed, David Castellano, Carlos Monserrat-Aranda, and José Hernández-Orallo, "When AI Difficulty Is Easy: The Explanatory Power of Predicting IRT Difficulty," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 7 (2022): 7719–7727.
13. Maharshi Gor, Hal Daumé III, Tianyi Zhou, and Jordan Boyd-Graber, "Do Great Minds Think Alike? Investigating Human-AI Complementarity in Question Answering with CAIMIRA," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,

ed. Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Miami: Association for Computational Linguistics, 2024), 21533–21564.
14. Xiting Wang, Liming Jiang, José Hernández-Orallo, David Stillwell, Shiqiang Chen, Luning Sun, Fang Luo, and Xing Xie, "Evaluating General-Purpose AI with Psychometrics," *Communications of the ACM* 69, no. 5 (2026): 92–102
15. Lexin Zhou et al., "General Scales Unlock AI Evaluation with Explanatory and Predictive Power," *Nature* 652, no. 8108 (2026): 58–67.

2. The Connection Between Information and Intelligence

16. Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, 2nd ed. (Hoboken: Wiley-Interscience, 2006).
17. Robert. C. Stalnaker, *Inquiry* (Cambridge, MA: MIT Press, 1984), 85.
18. William A. Dembski, *Being as Communion: A Metaphysics of Information* (Surrey, UK: Ashgate, 2014), ch. 4.
19. Francisco J. Ayala, "Darwin's Greatest Discovery: Design without Designer," *Proceedings of the National Academy of Sciences* 104(1) (2007): 8567 – 8573.
20. Richard Dawkins, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (New York: Norton, 1986).

4. The Connection Between Probability and Information

21. Claude E. Shannon, *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1949).

8. Conditional versus Unconditional Intelligence Metrics

22. Consider, for instance, the *nemesis effect* in chess: A grandmaster with a significantly higher Elo rating consistently loses to a lower-rated opponent because of a deep-seated psychological block. This phenomenon can occur when a player has a history of losses against a particular rival and, as a result, second-guesses himself in critical

moments. Despite superior calculation skills and opening preparation, the grandmaster may fall into passive play, avoiding complications where he normally thrives, ultimately succumbing to an opponent he should, by all objective measures, outclass. In the early 1970s, Viktor Korchnoi suffered from such a nemesis effect against Tigran Petrosian. In the 2000s and 2010s Magnus Carlsen suffered from a mild nemesis effect against Vladimir Kramnik. This effect appears in other types of competitive play. For instance, Roger Federer is considered the better tennis player than Rafael Nadal, yet in total match play, Nadal leads Federer 24 to 16, completely dominating on clay courts (14 to 2). See https://headtoheadmatch.com/roger-federer-vs-rafael-nadal-head-to-head.

10. Search as Key to Intelligent Performance

23. I suggest elsewhere that some intelligences may achieve successful search outcomes without executing an algorithmic search process but by intuiting where the targets being sought may be found. See William A. Dembski, "The Law of Conservation of Information: Search Processes Only Redistribute Existing Information," *BIO-Complexity* 2025, no. 2 (2025): 1–58.
24. Daniel N. Osherson, Michael Stob, and Scott Weinstein, *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists* (Cambridge, MA: MIT Press, 1986).
25. Daniel C. Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (New York: Simon & Schuster, 1995), 144; Stuart A. Kauffman, *At Home in the Universe: The Search for Laws of Self-Organization and Complexity* (New York: Oxford University Press, 1995), 161.
26. Robert M. Hazen, Patrick L. Griffin, James M. Carothers, and Jack W. Szostak, "Functional Information and the Emergence of Biocomplexity," *Proceedings of the National Academy of Sciences* 104, suppl. 1 (2007): 8574–8581.
27. H. Allen Orr, Review of *No Free Lunch: Why Specified Complexity Cannot Be Purchased Without Intelligence*, *Boston Review,* June 2002: https://www.bostonreview.net/articles/h-allen-orr-review-no-free-lunch.

11. Performance and Failure Functions

28. The William Lowell Putnam Mathematical Competition, administered by the Mathematical Association of America (MAA), is notorious for its difficulty. The competition takes place annually on the first Saturday of December, with three hours in the morning and three in the afternoon, and with a maximum possible score of 120. The median score is typically zero. For instance, describing the 2011 competition, a Mathematical Association of America newsletter reports: "As usual, the median score was 0 and the average this year was 4.3. A score of greater than 13 was needed to crack the top 500 and the high score was 91 out of 120." Quoted from http://sections.maa.org/rockymt/newsletters/fall2012news.pdf (last accessed January 24, 2025). The official website for the Putnam Competition is https://maa.org/putnam.
29. Lofti A. Zadeh, "Fuzzy Sets," *Information and Control* 8(3) (1965): 338–353.
30. Type-1 fuzzy sets are mappings from a space to the unit interval [0,1]. Type-2 fuzzy sets are mappings from a space to the set of all closed subintervals of [0,1] (i.e., all intervals of the form $[a,b]$ where $0 \leq a \leq b \leq 1$). For both types, see George J. Klir and Bo Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications (Upper Saddle River, NJ: Prentice-Hall, 1995), sec. 1.3.

13. Weighted Performance Measures

31. As Ruth put it, "If I'd just tried for them dinky singles I could've batted around .600." See https://www.baseball-almanac.com/quotes/quoruth.shtml (last accessed January 27, 2025).

15. Percentile Performance/Failure Functions

32. See https://www.act.org.
33. See https://blog.prepscholar.com/act-percentiles-high-precision (last accessed January 22, 2025).
34. On its website, the College Board, which administers the SAT, focuses on SAT percentiles that top out at 99, or "99+" as they put it. See https://research.collegeboard.org/reports/sat-suite/understanding-scores/sat. But there's a lot of variability in that 99^{th} percentile. For

a "high precision version" of SAT percentiles, that adds four more significant digits to the percentiles and from which the numbers in this article are taken, see https://blog.prepscholar.com/sat-percentiles-high-precision-2016. Both sites last accessed January 31, 2025.

16. Effect-Size Performance/Failure Functions

35. Warren S. Torgerson, *Theory and Methods of Scaling* (New York: Wiley, 1958).
36. For a standard text on effect sizes, see Paul D. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* (Cambridge: Cambridge University Press, 2010).
37. Steering Committee of the Physicians' Health Study Research Group, "Final Report on the Aspirin Component of the Ongoing Physicians' Health Study," *New England Journal of Medicine* 321(3) (July 20, 1989): 129–135.
38. Ellis, *Essential Guide*, 23.
39. Arpad E. Elo, *The Rating of Chessplayers, Past and Present*, 2nd edition (New York: Arco, 1986). The Elo rating system has also been used to rate player skill in other board games (e.g., Go and backgammon), card games, different sports, and video games. See https://en.wikipedia.org/wiki/Elo_rating_system#Use_outside_of_chess (last accessed January 31, 2025).
40. For a popular account, see https://www.omnicalculator.com/sports/elo (last accessed February 1, 2025). For the mathematics, see Elo, *Rating of Chessplayers*, 149 (third column).
41. See https://ratings.fide.com/toparc.phtml?cod=737 (last accessed January 31, 2025).
42. See https://ratings.fide.com/toparc.phtml?cod=785 (last accessed January 31, 2025).
43. See https://ratings.fide.com/top.phtml (last accessed January 22, 2025). These numbers keep getting revised in real time.
44. See https://www.chess.com/players/magnus-carlsen (last accessed January 22, 2025).
45. See https://en.wikipedia.org/wiki/1972_in_chess (last accessed January 31, 2025).

17. Random Predicates as Generalized Performance/Failure Functions

46. William A. Dembski, "Random Predicate Logic I: A Probabilistic Approach to Vagueness," *Progress in Complexity, Information, and Design* 1(2–3) (2002): http://www.iscid.org (discontinued but all past journal issues are available online at web.archive.org). This paper is also available at https://billdembski.com/wp-content/uploads/Random-Predicate-Logic-Dembski.pdf. ISCID = International Society for Complexity, Information, and Design.
47. Ibid.

18. Measuring Intelligence in the Search for Targets

48. For a recent comprehensive account of targeted search, see Denis Grebenkov, Ralf Metzler, and Gleb Oshanin, eds., *Target Search Problems* (Cham: Springer, 2024).
49. For an overview of my work with Robert Marks in this area, see Robert J. Marks II, William A. Dembski, and Winston Ewert, *Introduction to Evolutionary Informatics* (Singapore: World Scientific Publishing, 2017). For the most powerful and general formulation of conservation of information to date, see my monograph William A. Dembski, "The Law of Conservation of Information."
50. Richard Dawkins, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (New York: Norton, 1986), 45–50.
51. Strictly speaking, the probability of finding the target phrase METHINKS IT IS LIKE A WEASEL via blind search in a single step is 1 in 10^{40} whereas getting to a probability of .999999 with cumulative selection will require something like 100 steps. In 100 steps of blind search, the probability of successfully locating METHINKS IT IS LIKE A WEASEL goes up to circa 1 in 10^{38}. Strictly speaking, therefore, in comparing probabilities we should be keeping the number of search steps comparable between the two searches, blind and cumulative. Accordingly, we should be comparing the probabilities 1 in 10^{38} to .999999 rather than, as in the text, 1 in 10^{40} to .999999, yielding instead

an active information of $I = -\log_2\left(\frac{10^{-38}}{0.999999}\right) = 126.233$, which is only about six and a half bits smaller than calculated in the text, and thus still close enough to effectively illustrate the point at issue.

20. Sequential Search and Waiting Times

52. See, for instance, Patrick Billingsley, *Convergence of Probability Measures* (New York: John Wiley & Sons, 1968), ch. 2.
53. See https://top500.org/lists/top500/2024/11 (last accessed February 1, 2025). As of the summer of 2026, China's LineShine displaced El Capitan as the world's most powerful supercomputer. See https://top500.org/news/lineshine-debuts-no-1-top500-enters-new-global-exascale-era/ (last accessed July 28, 2026). LineShine almost reaches 2.2 exaflops.
54. Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes, *Introduction to the Theory of Statistics*, 3rd ed. (New York: McGraw-Hill, 1974), 99–100. Note that the expected value for the geometric distribution in this book is given as $(1–p)/p$, rather than $1/p$, as in this paper, because the authors begin indexing their random variables at 0 as opposed to 1, as in this paper.
55. Sheldon M. Ross, *Introduction to Probability Models*, 10th ed. (New York: Academic Press, 2010), 194.

21. Complexity as a Measure of Intelligence

56. William A. Dembski, *The Design Inference: Eliminating Chance through Small Probabilities* (Cambridge: Cambridge University Press, 1998).
57. William A. Dembski and Winston Ewert, *The Design Inference: Eliminating Chance through Small Probabilities*, 2nd ed. (Seattle: Discovery Institute Press, 2023).
58. Dembski, *Design Inference*, 1st ed., 93–94.
59. This is almost verbatim from Dembski, *Design Inference*, 1st ed., 99.
60. I'm offering here a rational reconstruction of what we do in practice when an agent overcomes a seemingly vast improbability, namely, we assume that the agent is operating with a different probability distribution that renders the probability of the agent's success highly probable. Strictly speaking, though, successful attainment shows that

a difficulty was overcome, but it does not by itself establish the agent's probability of success or justify ascribing to the agent the capacity to fully overcome the difficulty. The degree of intelligence credited to the agent must, as noted at the end of section 10, correspond to the degree that the agent's own capacities account for reaching the target, and not to extraneous factors such as luck or cheating.

23. The Intelligence of Natural Selection

61. R. A. Fisher, "Retrospect of the Criticisms of the Theory of Natural Selection," in Julian Huxley, A. C. Hardy, and E. B. Ford, eds., *Evolution as a Process* (London: Allen & Unwin, 1954), 91.
62. Julian Huxley, "The Evolutionary Process," in Julian Huxley, A. C. Hardy, and E. B. Ford, eds., *Evolution as a Process* (London: Allen & Unwin, 1954), 5.
63. Huxley, "The Evolutionary Process," 5.
64. Julian S. Huxley, "Natural Selection and Evolutionary Progress," *Nature* 138 (1936): 571–573.
65. Michael J. Behe, *Darwin's Black Box: The Biochemical Challenge to Evolution* (New York: Free Press, 1996).
66. For a more realistic example of antibiotic resistance achieved through selection (albeit artificial rather than natural selection), see Daniel Weinreich, Nigel Delaney, Mark DePristo, and Daniel Hartl, "Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins," *Science* 312 (2006): 111–114. They examined five specific mutations in beta-lactamase that jointly increase cefotaxime resistance by about 100,000-fold. A factor that complicates the calculation of probabilities here is that the mutations could occur in different orders ($5! = 120$ different sequential arrangements) and some of the orders preclude the development of this magnitude of antibiotic resistance.
67. Behe, *Darwin's Black Box*.
68. Michael J. Behe, *The Edge of Evolution: The Search for the Limits of Darwinism* (New York: Free Press, 2007).
69. Jason Rosenhouse, *The Failures of Mathematical Anti-Evolutionism* (Cambridge: Cambridge University Press, 2022), 178.
70. Paul S. Moorhead and Martin M. Kaplan, eds., *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*, Wistar

Institute Symposium Monograph no. 5 (Philadelphia: Wistar Institute Press, 1967), 30.

71. Douglas D. Axe, "Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds," *Journal of Molecular Biology* 341, no. 5 (2004): 1295–1315.
72. See, for instance, William A. Dembski and Michael Ruse, eds., *Darwin and Design: The Ongoing Debate on Biological Evolution* (Cambridge: Cambridge University Press, forthcoming 2027). See also Denis Noble and David Sloan Wilson, *Is Neo-Darwinism Enough? The Noble-Wilson Dialogue on Evolution* (Lexington, KY: Inkwell Press, 2025).

24. Why LLMs Aren't Smart—Lean Versus Laden Intelligence

73. dots team, "The dots Model Earns an Officially Certified Perfect-Score Gold Medal at IMO 2026," accessed August 25, 2026, https://studio-dots-ai.github.io/dots_imo_2026/en.html. See also International Mathematical Olympiad, "67th IMO 2026 – Individual Results," accessed August 25, 2026, https://www.imo-official.org/results/individual/year/2026.
74. Noam Chomsky, *Rules and Representations* (New York: Columbia University Press, 1980), 34. For a quick and popular account, see Richard Nordquist, "The Theory of Poverty of the Stimulus in Language Development," *ThoughtCo*, updated July 3, 2019, https://www.thoughtco.com/poverty-of-the-stimulus-pos-1691521.
75. For a memorable fictional example of this technique, see the 1991 film *Class Action*, starring Gene Hackman and Mary Elizabeth Mastrantonio.

25. Measuring Intelligence in Information Packing

76. Michael Liss et al., "Embedding Permanent Watermarks in Synthetic Genes," PLoS ONE 7, no. 8 (2012): e42465.
77. See Ingemar J. Cox, Matthew L. Miller, Jeffrey A. Bloom, Jessica Fridrich, and Ton Kalker, *Digital Watermarking and Steganography*, 2nd ed. (Amsterdam: Morgan Kaufmann, 2008).

78. Density needs to be distinguished from Kolmogorov complexity. In fact, they are separate concepts. It's possible for a long bit string to have only a short segment that satisfies a functional requirement. Its information density is therefore low. Yet the full string may have high or low Kolmogorov complexity. On the other hand, a long bit string may be such that each bit is vital for satisfying its functional requirement. Its information density is in that case as high as it can be. Yet again, the full string may have high or low Kolmogorov complexity.
79. Mouse Genome Sequencing Consortium, "Initial Sequencing and Comparative Analysis of the Mouse Genome," *Nature* 420 (2002): 520–562.
80. L. Fedorova and A. Fedorov, "Puzzles of the Human Genome: Why Do We Need Our Introns?," *Current Genomics* 6, no. 8 (2005): 589–595.
81. ENCODE Project Consortium, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature* 489 (2012): 57–74.
82. Lukas F. K. Kuderna et al., "Identification of Constrained Sequence Elements across 239 Primate Genomes," *Nature* 625 (2024): 735–742.
83. Dan Graur et al., "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE," *Genome Biology and Evolution* 5, no. 3 (2013): 578–590; W. Ford Doolittle, "Is Junk DNA Bunk? A Critique of ENCODE," *Proceedings of the National Academy of Sciences* 110, no. 14 (2013): 5294–5300; Manolis Kellis et al., "Defining Functional DNA Elements in the Human Genome," *Proceedings of the National Academy of Sciences* 111, no. 17 (2014): 6131–6138.
84. Shalev Itzkovitz, Eran Hodis, and Eran Segal, "Overlapping Codes within Protein-Coding Sequences," *Genome Research* 20, no. 11 (2010): 1582–1589.
85. Michael F. Lin et al., "Locating Protein-Coding Sequences under Selection for Additional, Overlapping Functions in 29 Mammalian Genomes," *Genome Research* 21, no. 11 (2011): 1916–1928.
86. Chase W. Nelson, Zachary Ardern, and Xinzhu Wei, "OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes," *Molecular Biology and Evolution* 37, no. 8 (2020): 2440–2449. See also Bradley W. Wright, Mark P. Molloy, and Paul R. Jaschke, "Overlapping Genes in Natural and Engineered Genomes," *Nature Reviews Genetics* 23, no. 3 (2022): 154–168.
87. See https://artificialintelligenceact.eu/article/50.

88. Google DeepMind, “SynthID,” accessed September 2, 2026, https://deepmind.google/models/synthid.