Working Paper

EDWARD GEIST, ALVIN MOON

What Even Superintelligent Computers Can't Do

A Preliminary Framework for Identifying Fundamental Limits Constraining Artificial General Intelligence

RAND Global and Emerging Risks

WR-A3990-1
June 2025

For more information on this publication, visit **www.rand.org/t/WRA3990-1**.

Published by the RAND Corporation, Santa Monica, Calif.

WR-A3990-1

About This Working Paper

Many commentators anticipate that artificial general intelligence (AGI) will have unprecedented geopolitical impact because it will be able to invent new technologies and radically improve existing ones. However, the laws of the physical universe there impose fundamental limits such that things that we can predict with confidence that even the most powerful forms of AGI will not be able to do. This Working Paper outlines an approach for estimating the likelihood that a particular technology, skill, or capability is attainable in practice, and explores the example of post-quantum cryptography (PQC) to illustrate the application of this approach.

Intended Audience

This Working Paper is intended for policymakers and general audiences interested in understanding the potential limitations of AGI in light of discussions about its capabilities.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

This effort was independently initiated and conducted within the Technology and Security Policy Center using income from operations and gifts from philanthropic supporters, which have been made or recommended by DALHAP Investments Ltd., Effektiv Spenden, Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Jaan Tallinn, Longview, Open Philanthropy, and Waking Up Foundation. A complete list of donors and funders is available at www.rand.org/TASP. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

We would like to thank Jason Matheny and the Technology and Security Policy Center in RAND Global and Emerging Risks for their support in carrying out this study. We would also like to thank our RAND colleagues and those of the RAND Artificial General Intelligence Advisory Council, particularly Miles Brundage, for their thoughtful and timely feedback on an earlier draft of this piece.

Contents

Figures

What Even Superintelligent Computers Can't Do: A Preliminary Framework for Identifying Fundamental Limits Constraining Artificial General Intelligence

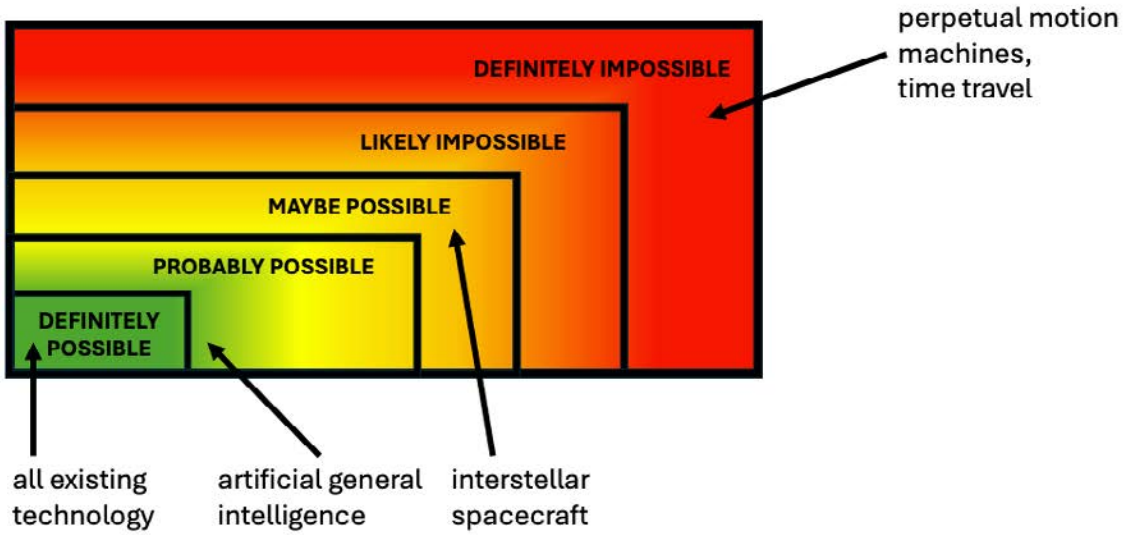
Introduction

Many commentators anticipate that artificial general intelligence (AGI) will have an unprecedented geopolitical impact because of its expected ability to invent new technologies and radically improve existing ones. Preeminent AI expert Stuart Russell went so far as to compare AGI with "the genie in the lamp, or the sorcerer's apprentice, or King Midas" (Russell, 2014). AGI may indeed prove capable of inventing technologies and performing tasks that we cannot even imagine at present. But contrary to Arthur C. Clarke's famous dictum that "any sufficiently advanced technology is indistinguishable from magic" (Clarke, 1973), we can predict with confidence many things that even the most powerful forms of AGI will not be able to do.

To formulate effective policies in anticipation of AGI's emergence, policymakers and analysts need methodologies to predict the way in which fundamental physical limits will likely constrain what AGI might enable. In this Working Paper, we outline an approach for estimating the likelihood that a particular technology, skill, or capability is attainable in practice. A key feature of this framework is its open-ended definition of *artificial general intelligence*. No matter what form AGI takes, the framework should remain applicable.

To set priorities and navigate potential trade-offs, policymakers need to predict the feasibility of various prospective technological capabilities. Figure 1 depicts a spectrum of technological possibilities ranging from the definitely possible (e.g., all existing technologies) in the lower left-hand corner to the definitely impossible at the top right-hand corner. Unless the prevailing scientific understanding of physical reality is very grievously in error, perpetual motion machines and time travel can be relegated to the definitely impossible category (Hawking, 1992).

Figure 1. A Spectrum of Technological Possibility



The intermediate region of technologies that may or may not be practically feasible is of greater interest to analysts and policymakers. An interstellar spacecraft, as long as it travels slower than the speed of light, falls into the middle category of maybe possible. While this starship would not violate any known physical laws, if it were of greater than trivial size and if it were to travel interstellar distances in a period shorter than millennia, it would require immense energy and other resources, as well as solutions to many nettlesome subsidiary engineering challenges (Lubin, 2019). Absent existential proof of such an interstellar spacecraft, we ought not be too confident that these obstacles are insurmountable.

AGI, by contrast, appears to belong in the probably possible category. The example of the human brain supports the contention that machines with general intelligence can be created. Presuming that human intelligence has a physical basis, it should, in principle, be possible to replicate those same processes mechanically.[1]

This spectrum of technological possibility would be of limited interest to policymakers if it were limited to science fiction staples, such as time machines, starships, and human-like robots. Its practical value lies in the fact that many comparatively mundane technologies and applications can be shown to belong in the likely impossible and definitely impossible categories. Conversely, sometimes exotic capabilities can be more feasible than they seem.

---

[1] This presumption is distinct from the question of whether AGI is likely in the near term. While confidence in a technology's feasibility is likely to increase in the run-up to the emergence of that technology, the framework outlined in this Perspective is intended solely to predict the feasibility of a technology, not *when* it will become a reality if it proves feasible.

An example of a comparatively mundane fictional technology that turns out to be definitely impossible is the so-called enhance button regularly featured on TV series, such as *CSI*, *Law and Order*, and *24*. This device (usually a piece of computer software) can take a low-resolution image—for example, a highly pixelated low-resolution still image from a closed-circuit television camera—and enhance it to reveal additional detail when the protagonists need to see additional detail. Regrettably, information theory forbids a version of the enhance button that can reveal information that is not present in the original signal. (Real-world counterparts of this technology, including current artificial intelligence [AI] upscaling, work by interpolating data to make an educated guess of what additional detail would look like—which inevitably risks confabulating something different from reality.)

Equipped with determinations of likely technological feasibility, policymakers can decide how to allocate available resources. The United States does not need to waste valuable effort or resources on attempting to obtain an impossible capability or worry that an adversary might use such a capability against the United States. (However, the United States might try to deceive its rivals into squandering *their* resources chasing a possibility that is known to be illusory.) Intermediate possibilities can be prioritized based on a combination of their anticipated likelihood and prospective impact. A high-consequence capability might still deserve serious consideration despite appearing probably impossible; a near-certain but low-consequence capability might deserve to be ignored.

The laws of physics and theoretical mathematics present many potential barriers to technological feasibility. Some of the more significant of these constraints are as follows:

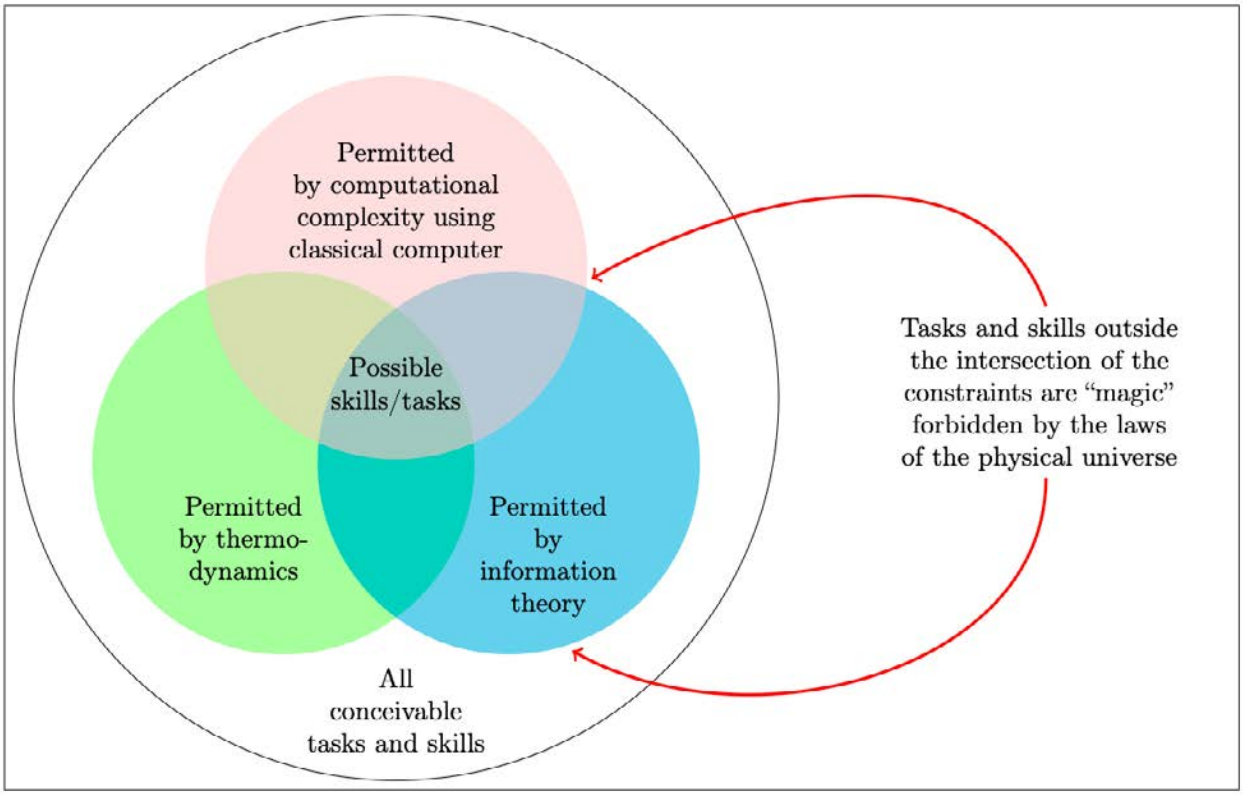
- the laws of thermodynamics, most importantly the second law of thermodynamics (i.e., one cannot do work without using energy, and available energy is finite) (Atkins, 2010)
- information theory and Shannon entropy (i.e., a finite amount of bandwidth can convey only a bounded amount of information without losses) (Khinchin, 2013)
- computational complexity (i.e., one cannot solve arbitrary instances of NP-hard problems without carrying out the associated, potentially intractable, computations) (Karp, 1975)[2]
- inescapable trade-offs that an agent must navigate, such as between exploitation and exploration
- the impossibility of reasoning with knowledge that one does not have
- the law of unintended consequences, which always applies at the margins of an agent's knowledge, no matter how intelligent that agent otherwise is.

The more known or suspected constraints to impede the feasibility of a technology, the less likely it is to be a practical possibility. Only those capabilities that are not forestalled by at least one of the potential barriers can be feasible. Figure 2 presents a Venn diagram illustrating this

[2] *NP* refers to nondeterministic polynomial, a class of decision problems that can be solved in polynomial time relative to input length on a nondeterministic Turning machine (a conjectural type of computer that could evaluate multiple program branches simultaneously). NP-hard problems, some of which are themselves in NP but most of which belong to harder complexity classes, are those that can be reduced to any problem in NP in polynomial time. What this means is that a fast solution to any NP-hard problem would be a fast solution to every problem in NP.

concept for three constraints: thermodynamics, information theory, and computational complexity on a classical computer (i.e., the kind of digital computer in widespread use today).

Figure 2. Possible Tasks and Skills Subject to Constraints Imposed by Thermodynamics, Information Theory, and Computational Complexity



While two of the potential constraints shown in Figure 2, thermodynamics and information theory, are relatively well understood, those related to computational complexity are less certain (Aaronson, undated). (For example, whether P = NP remains an open question among theoretical computer scientists.) New discoveries might significantly relax one or more of such constraints, so that new technologies that appeared infeasible before might suddenly become practical possibilities.

These constraints can be expected to apply equally to natural and artificial intelligences. Despite timeworn arguments contending that reasoning machines would be inhibited by constraints from which human minds would be immune (see, for example, Lucas, 1961), there is no compelling reason to assume that humans would possess some kind of durable advantage in this area. Those skills and tasks that are not permitted by the constraints are things that neither humans nor computers can do.

Sample Case: Cryptography and Quantum Computing

The intersection of cryptography and quantum computing illustrates how scientific surprise can beget prospective technological surprise with vast potential geopolitical implications, as well as how the framework outlined above can help policymakers anticipate the likelihood of such surprises. It also exemplifies how theory can potentially identify the means to *forestall* technological surprise, even in the face of prospective AGI smarter than human intelligence.

Public-key cryptography is a foundational technology of contemporary networked communications and the contemporary economy based on them. Developed in the 1970s, this technology makes it possible to transmit secure, encrypted communications via insecure infrastructure that could permit eavesdroppers to intercept encrypted data without having to share the means to decrypt messages with the transmitter. It does this by splitting the cryptographic key into two parts: a public key that can only encrypt messages and a private key that the recipient uses to decrypt messages. The transmitter uses the public key to encrypt the message, transmits the encrypted result, and the recipient uses their private key to decrypt the message. Public-key cryptography is used for everything from securing financial data for internet transactions to transmitting highly sensitive intelligence data.

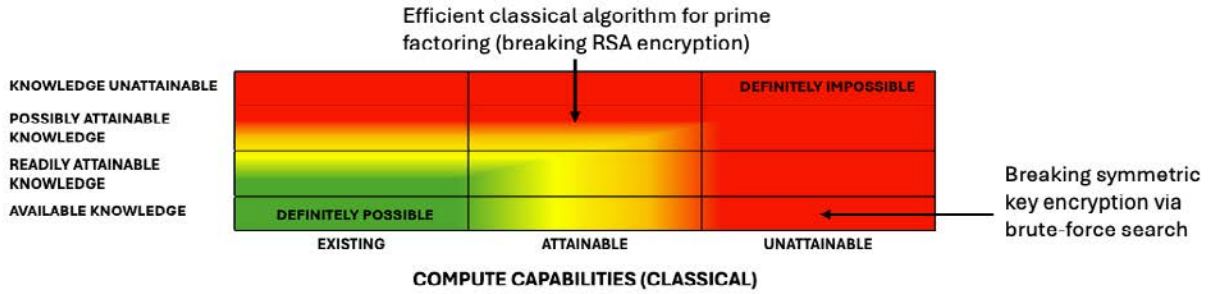
The most prominent public-key cryptosystem is RSA (Rivest–Shamir–Adleman). Named after the three authors who published the algorithm in 1977, RSA employs a public key based on two large prime numbers along with an auxiliary value (Rivest, Shamir, and Adelman, 1977). The user (recipient) generates and shares the public key and keeps the two prime numbers secret. Senders can encrypt messages via the public key but cannot themselves decrypt the resulting ciphertext without the private key (pair of primes). RSA appeared secure at the time it was designed and for many years afterward because it relies on the practical difficulty of factoring the product of two large prime numbers using a classical computer.

Decrypting RSA ciphertext without the public key is *theoretically* possible, but a *practical* algorithm for doing so on a classical computer has not yet been identified. A naive approach is to do a brute-force search for the private key, but this would require far too much time to be practical. Doing so would require a literally astronomical amount of time for real-world RSA implementations. A more sophisticated approach is to use a factoring algorithm, such as the general number field sieve, to try and identify the secret prime numbers using the public key. This still requires far too much computation to be practical to decrypt typical encrypted traffic, but it works on a small scale for weaker forms of RSA. It is conceivable that an efficient classical algorithm could be discovered for prime factoring that would make it possible to decrypt the types of RSA encryption presently in general use (more on this below).

Like Figure 1, Figure 3 presents a spectrum of the limits on breaking RSA encryption using classical computers given our current state of knowledge. The availability of knowledge is presented along the y-axis. (By *knowledge*, we refer here to algorithm design, not to knowledge of the private key or encrypted message.) The x-axis represents the amount of computational

power available to run those algorithms, ranging from the amount of compute available today (shown at left) to the amount of compute that might be available in the future (shown in the center) to quantities of compute that cannot be realized physically (shown at right). Because we know how to design a naive brute-force algorithm that no attainable computer can run, that algorithm is located in the lower right-hand corner of Figure 3. The hypothetical efficient classical factoring algorithm, meanwhile, is located at the intersection of possibly attainable knowledge and attainable compute. This algorithm might or might not exist to be found; hence, its discovery is possible. If this algorithm were found, it would move to the intersection of available knowledge and attainable or available compute.[3]

Figure 3. Limits on Breaking RSA Cryptographic Protocol with Classical Computer



NOTE: The y-axis represents algorithmic knowledge, not any insight into the private encryption key or plaintext content of an encrypted message.

Classical computers are not the sole possible kind. In the 1980s, physicists proposed the idea of a quantum computer—a computing device that represents information using quantum qubits instead of the discrete states employed by a classical (digital) computer (Aaronson, 2013). Contrary to popular misconceptions, a quantum computer is not a vastly faster version of a digital computer but rather an embodiment of an alternate computational paradigm. For many tasks, quantum computers offer no theoretical advantage whatsoever over classical computers. But there are certain applications for which quantum computers may offer an immense qualitative advantage over their classical counterparts.

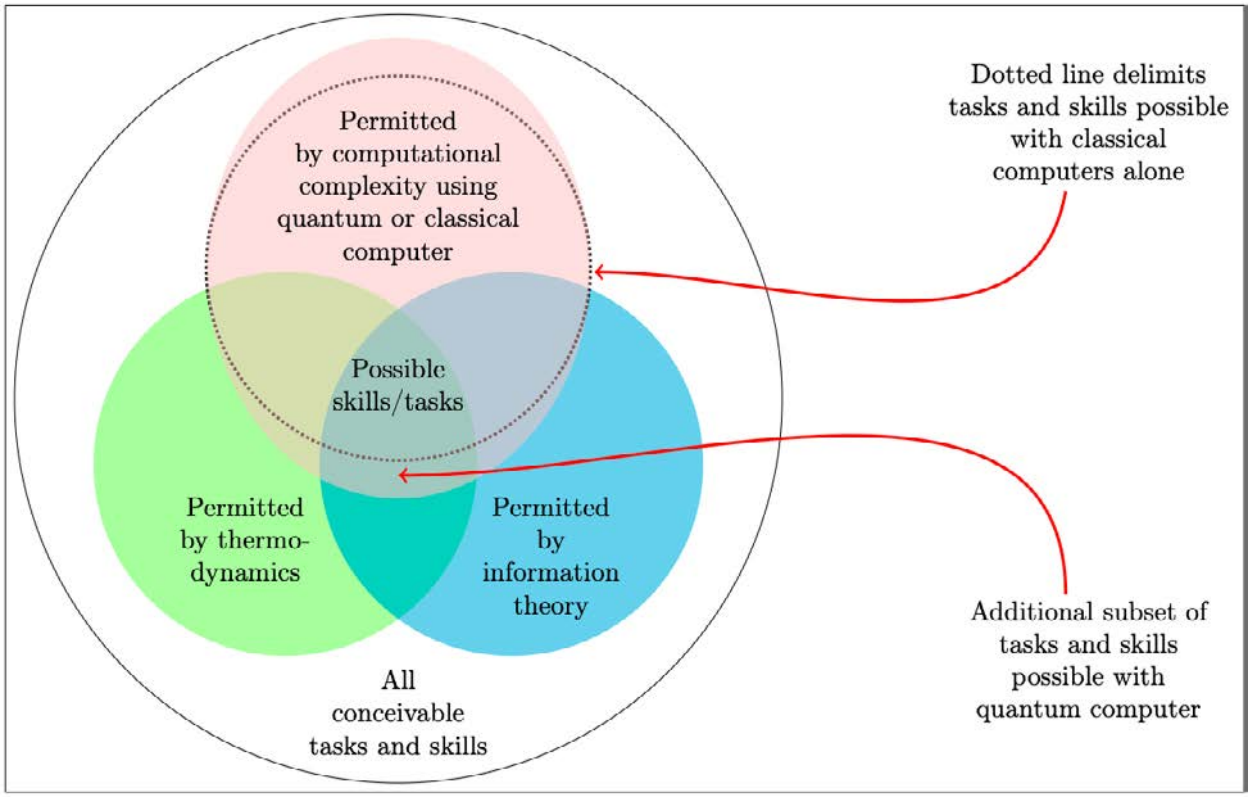
In this respect, quantum computers can be characterized as relaxing the constraints imposed by computational complexity on the set of possible skills and tasks. This is illustrated by the Venn diagram in Figure 4. Thanks to the additional computational tasks rendered feasible by the

[3] An efficient classical factoring algorithm or other means of breaking RSA encryption using attainable or existing classical computers is an example of a highly consequential technological innovation that might be enabled by even a primitive, nascent variant of AGI. Because this task can be analyzed from the standpoint of number theory and other theoretical mathematics concepts, it does not require the solution of such challenging problems as physical embodiment. Alternatively, if this algorithm exists, it might be identified in the near term either by a dedicated narrow AI (e.g., a mathematical concept discovery system) or by flesh-and-blood human mathematicians.

availability of quantum computers, the set of possible skills and tasks has expanded compared with the version of the Venn diagram in Figure 2, which considered classical computational capabilities only.

As it happens, the ability to break RSA encryption may be part of the additional sliver of possible skills and tasks. In 1994, mathematician Peter Shor published the eponymous Shor's algorithm (Shor, 1994). This is a quantum algorithm for prime factoring that, given a sufficiently capable quantum computer on which to run it, should be able to break the kind of RSA encryption in use today in a practical amount of time. This possibility has led both to enthusiasm and anxiety about the future applications of quantum computers. The current state of knowledge and compute for breaking RSA encryption with a quantum computer is the reverse of that for a classical computer, as illustrated by Figure 5. We already have the algorithmic knowledge in the form of Shor's algorithm, but we lack a practical quantum computer on which to run it.

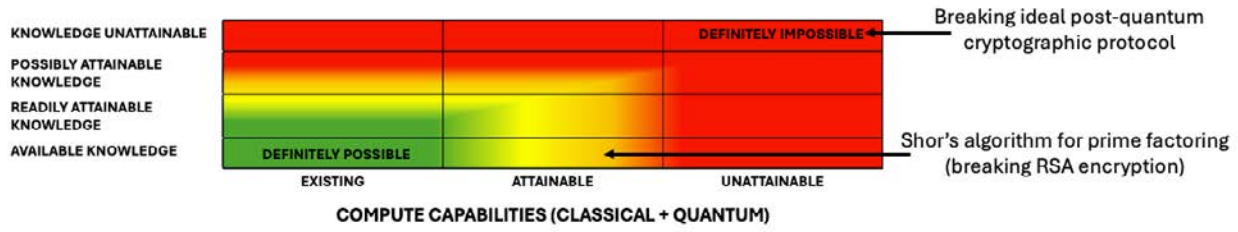
Figure 4. Comparison of Skills and Tasks Possible Using a Classical or Quantum Computer



The possibility that quantum computers could be invented that render existing cryptographic protocols obsolete has inspired a new field: post-quantum cryptography (PQC) (Bernstein and Lange, 2017). PQC aims to create cryptographic protocols that will remain practically unbreakable even using arbitrarily powerful future quantum computers. Ideally, it would be

provably impossible to decrypt the message without the private key. This aspiration is associated with the upper-right-hand corner of Figure 5.

Figure 5. Limits on Breaking Cryptographic Protocols with a Quantum Computer



NOTE: The y-axis represents algorithmic knowledge, not any insight into a private encryption key or plaintext content of an encrypted message. The x-axis represents computational resources from classical, quantum, and hybrid quantum-classical systems.

In hindsight, the security limitations of the RSA cryptographic protocol stem from the fact that its creators believed prime factoring to be computationally intractable without any formal evidence that this intuition was correct (Boneh, 1999). PQC seeks to leverage theoretical computer science to avoid repeating this mistake by anticipating the capabilities of conjectural future quantum computers and identifying tasks that are provably hard (in the sense of requiring an infeasible amount of time for a computer to solve).[4] If PQC succeeds, the resulting cryptographic protocol should by practically infeasible to break with physically realizable quantum and classical computers. One implication of this would be that future AGI could not break such encryption or discover a means of breaking such encryption.

While PQC is an unusually mature example of an attempt to identify tasks that future computers will be unable to do, thanks to its grounding in theoretical mathematics and physics, the same principles can be applied to many other capabilities that AGI could enable. Similar analyses should be possible for many applications of policy and military interest, such as deception and counterdeception, information fusion for situational awareness, and automated strategy formulation. Via a systemic consideration of possible constraints and the degree of uncertainty regarding the applicability of those constraints, analysts should often be able to identify where a prospective capability lies on the spectrum of technological feasibility.

[4] See Brakerski et al. (2013) for a technical overview of what "provably hard" means in the context of PQC. In Figure 5, we represent the hardness of breaking an idealized PQC protocol as *definitely impossible*. In practice, PQC protocols are designed to be effectively impossible to break through arguments that establish provable hardness. Some NP-hard problems are easy in most instances, with only pathological cases posing a significant computational challenge. A canonical example of this is the *knapsack problem*, which is about selecting from a set of items with different values and weights to fill a knapsack of finite capacity to maximize the total value carried in the knapsack. A naive, greedy algorithm finds a good-quality or even optimal solution to this problem most of the time. For a discussion see (Pisinger, 2005).

Conclusion

The goal of trying to design systems and procedures that are robust against scientific and technological surprise would be of paramount importance in a world in which AGI existed. The laws of nature impose fundamental limits that will constrain even the most intelligent possible machines. No matter how powerful AGI proves to be, it will not be magic, and there will be things that it will not be able to do. The things that AGI cannot do could be leveraged to build a future in which reinforcing constraints reduce uncertainty and provide security. Analysts need to identify these infeasible tasks to protect national security and human agency. The framework outlined in this Working Paper represents an initial step toward this goal.

Abbreviations

AGI	artificial general intelligence
AI	artificial intelligence
PQC	post-quantum cryptography
RSA	Rivest–Shamir–Adleman

References

Aaronson, Scott, "P $\stackrel{?}{=}$ NP," undated. As of January 10, 2025: https://www.scottaaronson.com/papers/pvsnp.pdf

Aaronson, Scott, *Quantum Computing Since Democritus*, Cambridge University Press, 2013.

Atkins, Peter, *The Laws of Thermodynamics: A Very Short Introduction*, Oxford University Press, 2010.

Bernstein, Daniel J., and Tanja Lange, "Post-Quantum Cryptography," *Nature*, Vol. 549, No. 7671, 2017.

Boneh, Dan, "Twenty Years of Attacks on the RSA Cryptosystem," *Notices of the AMS*, Vol. 46, No. 2, February 1999.

Brakerski, Zvika, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé, "Classical Hardness of Learning with Errors," in *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing (STOC '13)*, Association for Computing Machinery, June 2013.

Clarke, Arthur C., *Profiles of the Future: An Enquiry into the Limits of the Possible*, 2nd ed., Harper and Row, 1973.

Hawking, Stephen W., "Chronology Protection Conjecture," *Physical Review D*, Vol. 46, No. 2, July 1992.

Karp, Richard M., "Reducibility Among Combinatorial Problems. *Complexity of Computer Computations, Proceedings of a Symposium on the Complexity of Computer Computations, Held March 20–22, 1972, at the IBM Thomas J. Watson Center, Yorktown Heights, New York*, edited by Raymond E. Miller and James W. Thatcher, Plenum Press, New York and London 1972, pp. 85–103," *Journal of Symbolic Logic*, Vol. 40, No. 4, December 1975.

Khinchin, A. Ya., *Mathematical Foundations of Information Theory*, trans. by R. A. Silverman and M. D. Friedman, Courier Corporation, 2013.

Lubin, Philip, *A Roadmap to Interstellar Flight*, NASA Innovative Advanced Concepts, National Aeronautics and Space Administration, Report No. HQ-E-DAA-TN75825, 2019.

Lucas, John R., "Minds, Machines and Gödel," *Philosophy*, Vol. 36, No. 137, 1961.

Rivest, Ronald L., Adi Shamir, and Len Adelman, "On Digital Signatures and Public-Key Cryptosystems," Massachusetts Institute of Technology Laboratory for Computer Science, Technical Memorandum 82, April 1977.

Russell, Stuart, "Of Myths and Moonshine," *Edge*, November 14, 2014.

Shor, Peter W., "Algorithms for Quantum Computation: Discrete Logarithms and Factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, IEEE, 1994.

About the Authors

Edward Geist is a senior policy researcher at RAND. His research interests include the former Soviet Union, nuclear weapons, emergency management, and AI. He has a Ph.D. in history.

Alvin Moon is a mathematician at RAND. His research focuses on modeling and mathematical analysis across several topics, including AI, cryptography, supply chains, and workforce development. He has a Ph.D. in mathematics.