FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 11, 2023

Kyrtin Atreides

THE HUMAN GOVERNANCE PROBLEM: COMPLEX SYSTEMS AND THE LIMITS OF HUMAN COGNITION

doi: 10.37240/FiN.2023.11.1.4

ABSTRACT

The impact of complexity within government and societal systems is considered relative to the limitations of human cognitive bandwidth, and the resulting reliance on cognitive biases and systems of automation when that bandwidth is exceeded. Examples of how humans and societies have attempted to cope with the growing difference between the rate at which the complexity of systems and human cognitive capacities increase respectively are considered. The potential of and urgent need for systems capable of handling the existing and future complexity of systems, utilizing greater cognitive bandwidth through scalable AGI, are also considered, along with the practical limitations and considerations in how those systems may be deployed in real-world conditions. Several paradoxes resulting from the influence of prolific Narrow Tool AI systems manipulating large portions of the population are also noted.

Keywords: e-Governance, complexity, cognitive bandwidth, AGI, Artificial General Intelligence, scalability, tool AI, cognitive bias.

1. INTRODUCTION

A central repeating pattern across history, in government, science, and technology, has been increasing complexity (Kelly, 2011). At one point the sum of human knowledge could be passed on from one human to the next through spoken language, and later through written language. Humanity has long since moved past this phase, with specialists emerging to address increasingly specific kinds of science, technology, and government systems.

The scope of what any one human can adequately address decreases as the depth of knowledge required to address it increases. Additionally, the difference between "adequately" and "optimally" addressing issues can often be significant. This can be overcome to a limited degree through communication between specialists, whose scope covers different areas, but much is lost in the communication process, which also bounds the efficacy of classical multi-agent approaches. These cognitive limits and trade-offs give specialists roles that become increasingly narrow and deep over time, with gaps between them constantly emerging as a consequence, and requiring vigilance to recognize and fill with new kinds of specialists before those gaps create new problems in turn. As Friedrich Hayek put it, "... the data from which the economic calculus starts are never for the whole society given to a single mind which could work out the implications, and can never be so given" (Hayek, 1945).

The systems within which humans operate, whether academic, commercial, or governmental, each themselves pose further problems. For example, the selection of ideal candidates for hiring suffers heavily from both human and automated biases (Chua, Mazmanian, 2020), including many biases tied to factors that are not visually perceived (Rodríguez-Ferreiro, Barberia, 2017). The organization and interactions within teams are often equally biased, and sometimes completely arbitrary, producing an abundance of waste, miscommunication, and interpersonal friction. These factors all serve to divert attention away from the gaps in knowledge being considered across such teams, distracting from the discovery of "unknown unknowns" (Pawson et al., 2011) that silently cause further problems in each system.

The study and application of Complexity Theory are ongoing (Kallemeyn et al., 2020; Sowels, 2021), and notoriously difficult to both define and measure, but the point at which people begin to heavily simplify and summarize can often be illustrated by the same intuitive mechanisms. For example, how many pages long would a typical document need to be before you began to summarize the contents rather than going over them in full detail? For some of the more difficult documents you have read, such as particularly detailed research papers or legal documents, how many pages long would they need to be before you began to simplify their contents?

For many people a 10-page research paper is enough to trigger these mechanisms, allowing us to simplify them to a point and in such a way as to best fit within our prior models of the world (Friston et al., 2021) with only minimal adjustments if any to those prior models. We automatically seek to minimize complexity, even though we recognize that comes at a cost, usually in the form of various trade-offs (Bar-Yam, 2000).

Yet, our societies and governments often do not even remotely reflect this (Limberg et al., 2022; Katz, 2014) as the complexity with which they operate has skyrocketed over recent decades. The "Harmonized Tariff Schedule of the United States (2022) Revision 6" (U.S. International Trade Commission,

2022) is a monument to this complexity, weighing in at 4,334 pages in length. Individual bills in the US Congress often approach or exceed 1,000 pages each of dense legal text. When considering the level of competence such individuals statistically demonstrate in practice, I personally would not expect them to successfully simplify the process of setting up a home entertainment system, let alone federal government policies based on such documents. Even the most skilled and talented individuals in the world likely could not handle that level of complexity without applying a high degree of simplification.

Absurd examples of this have been demonstrated in government quite routinely, such as the US South-West allocating a greater volume of water from the Colorado River to supply their states than actually flows through the river (James, 2022). While these states allocated 15 million acre-feet of water per year, beginning with the 1922 Colorado River Compact, the actual average water flow was significantly less than 15 million and could vary significantly from year to year (US Bureau of Reclamation, 2022). The net result of this has been over 20 consecutive years of drought, the rapid depletion of groundwater via unregulated well-drilling (Alam et al., 2021), as well as sinking land and crumbling infrastructure built on it (Lees et al., 2021), and a myriad of other related consequences. These government systems have reached such degrees of both absurdity and obscurity that Americans often learn about them via comedic sources of news (Oliver, 2022).

When governments embody the punchlines of such comedy it becomes evident that far more than laughter is required in response. The Three Stooges should not be determining public policy, lest the joke is on every citizen of such countries. In order for governments and societies to function individually, let alone internationally within a globalized economy, the challenges of extreme and increasing complexity must be met with new levels of cognitive bandwidth not yet functionally accessible to humans.

For the rest of this paper’s contents, we will apply the following definitions:

Narrow (or Tool) AI: Systems that are able to perform one or more tasks automatically, regardless of how well these tasks are performed relative to humans. These systems lack a comprehensive set of human capacities, such as free will, subjective emotional experience, and consciousness, each with their own long-debated definitions and theories. However well they perform, these remain automated tools.

Artificial General Intelligence (AGI): Systems that are able to perform a wide variety of tasks independently, at least as well as the average human, while having their own subjective emotional experience, consciousness, and free will, in as much as humans do. Without these capacities demonstrated in humans, they cannot be said to operate at a “human-level” or above in the “general” sense.

Bias: The term bias is used to refer to the variety of known and documented cognitive biases, as well as any other systematic deviations from strictly rational behavior, including strong preferences and unintentional bias derived from data where noted.

Metaorganism: A community of interacting entities that collectively coexist within a shared environment. This collective can give rise to forms of adaptive behavior and intelligence not present in any one entity and encourages symbiotic and endosymbiotic relationships to emerge over time. An example of a metaorganism is a bee hive or any colony, where individuals specialize to serve functions that could not be sustainable absent the collective.

2. COMPLEXITY IN GOVERNMENT

"Governance refers to a category of social facts, namely the processes of interaction and decision-making among the actors involved in a collective problem that lead to the creation, reinforcement, or reproduction of social norms and institutions" (Wiesmann, Hurni, 2011).

The pattern of quickly increasing complexity in government and finite human cognitive bandwidth (Miller, Buschman, 2015) is not sustainable. Examples of these limits were famously demonstrated in studies by Simons, Chabris, one of which was nicknamed "The Invisible Gorilla" (Simons, Chabris, 1999), where participants in the study demonstrated Inattentional Blindness. That perceptual cognitive bias is one of the ways humans can hyper-focus on goals and metrics, such as key performance indicators, making themselves blind to any irregularities outside of that narrow scope. Similarly, poverty and other factors capable of reducing an individual's available cognitive bandwidth have demonstrated a strong influence on increasing the risk of poor decision-making (Schilbach et al., 2016). Neuroscience also frequently points to a significant divide between the information we subconsciously process, and the small fraction of that information we consciously perceive (Cohen et al., 2016).

Humanity has already created systems too complex for humanity itself to manage. Even systems at the scale of a moderately-sized city exceed the capacities of human cognitive bandwidth and established methods used by teams today. These systems are too complex to be optimized by humans alone, and as their complexity increases their efficiency strongly declines. This decline in efficiency is further affected by corruption, as corruption proliferates and feeds off the inefficiency, creating a negative feedback loop. Corruption in this case refers to bad actors and actions not aligned with the intended purposes of governance, such as bribery, extortion, and various other methods of greed. In practice, corruption could be compared to a staph infection making itself at home in places where the immune system cannot reach it, starting out living in the gaps and loopholes, but gradually degrading the rest of the system as they expand.

When the scale is increased to include multiple cities and regions the complexity of those systems increases rapidly, often becoming exponential when attempting to consider systems spanning multiple countries, such as the European Union (EU) or United Nations (UN). Attempting to build and maintain optimal systems of government that cover multiple cultures, religions, and other value systems is an incredibly difficult and important problem, the difficulty and importance of which increase as the differences between constituents increases.

A common practice has been the outsourcing of a large portion of the decision-making process via consulting (Collington, Mazzucato, 2021). Besides costing companies and governments a fortune this only lays the responsibility on another group of humans with the same limitations, whose motivation may be primarily or exclusively monetary gain. This is a bit like adding sugar to poison to make it taste less bitter, or a cat defecating on the floor and covering it with a bathmat. The underlying problem is unchanged, as the cost of poorly addressing it increases, and the motives behind those working on it may well be worse than those who would have to live with the consequences of failure.

Governments, and to a lesser degree scientific institutions, also suffer from increasing complexity over time. New laws and policies are passed, stacking the layers of bureaucracy ever higher, even as new loopholes are created with each new layer. This poor level of adaptation means that the efficiency of any such system almost always declines in the face of advancing technology, relative to the optimal adaptation. Drawing comparisons to the previous operation of a system rather than ideal adaptation can paint a more positive picture, but that is largely because it tends to completely omit the new opportunities and challenges introduced by the advancing technology. This can be, and often is, the difference between seeing a system as 10% more efficient and effective than it previously was, or seeing the same system as <1% as efficient and effective as it should be given present technology. This relativistic illusion is particularly dangerous, perhaps suicidally so, as it can promote self-satisfaction in a dying system.

3. COGNITIVE BIASES GOVERNING SOCIETY

When humans are faced with the challenge of making decisions beyond the reasonable scope of their cognitive bandwidth one factor reliably emerges, in around 200 different flavors, and that is cognitive bias (Ramachandran, 2012). These are the evolved and learned cognitive shortcuts that allow humans to make decisions quickly, and in circumstances where they are unable to fully reason their way to a conclusion. This process is often glorified with terms such as "gut instinct" (Rossmo, 2008), though it also produces systematic errors in judgment (Kahneman et al., 2021), such as racism and sexism (Payne, Hannay, 2021). These cognitive biases were extremely beneficial in driving human survival during periods where the decision to quickly run away from a predator held frequent significance, but they are wholly inadequate for the task of running a government.

There are many kinds of attempts made to reduce cognitive bias in governance, but each method increases the complexity of the system overall or causes the system to rely on automation, both of which only shift the negative impact of biases being expressed away from the metric being measured to gauge the performance of that method (Aczel, 2015). This can easily cause a net-negative impact through shifting the problem into a new unknown, often obscured, and potentially diluted state, combined with increasing the system's complexity even further beyond human cognitive bandwidth, resulting in an unknown or obscured expression of bias that is proportionately worse than the previously known manifestation. This is the typical narrow and short-sighted solution, popularly encouraged.

Bias in governments cannot be removed by increasing the complexity of the system, as the complexity of the system is one of the strongest pressures supporting the application of those biases. Bias also cannot be removed through Tool AI (Narrow AI), as the basis of such modern Tool AI algorithms are rooted in probability, not understanding. Making a system that parrots the talking points of an AI Ethics researcher is an arbitrarily easy task in 2022, but the same Tool AI has no understanding of the concepts or meaning conveyed when it spits out a sequence of words that are probabilistically likely based on the prompt and training data it was given.

The current problems and subsequent opportunities of "big data" are also highlighted by humanity's need to extract greater insights from more data, and struggles in attempting to do so (Hilbert, 2016). With global internet communication approaching 1 petabyte per second in 2022 and total internet data capacity projected to reach 175 zettabytes by 2025, this problem is growing quickly.

4. THE HUMAN CONTROL PROBLEM

The Human Control problem has two main points that need to be considered jointly. The first is that Tool AI systems like News Feeds, Recommenders, Search Engines, and other Ad-revenue or sales maximizing Tool AI systems today are optimized to manipulate the user as much as possible, and in as many ways as possible, all towards their programmed goals. These algorithms can and do cause every kind of harm they have the ability to at a global scale (Orlowski, 2020), 24/7, because it is highly profitable to make as many humans as possible more predictable, more emotional, more polarized, more biased, and more addicted. The more dependent humans are on the platforms built around such systems, with that dependency taking forms such as the "Google Effect" (Azzopardi, 2021), the more frequently and easily they may be manipulated into predictable niches for further profit generation. This has already produced major global declines in trust for public policy (Hosking, 2019), dramatically increased polarization (Tokita et al., 2021), rising suicide rates (Twenge et al., 2018), more organized extremism (Liang, Cross, 2020), genocide (Whitten-Woodring et al., 2020), and a hundred other kinds of harm.

This means that effectively such Tool AI systems have already become "paperclip maximizers" (Bostrom, 2003) also known as "paperclip monsters," and the humans they have successfully domesticated are now their "paperclips." This brings us to the second point, which is that any system designed to cater to paperclips effectively becomes an extension of said paperclip monsters.

To put this into practice, consider that misinformation, disinformation, and polarized opinions that carry an exaggerated emotional charge to others spread more quickly (Bowman, Cohen, 2020), they are more viral. Paperclip monster Tool AI systems prioritize spreading that material highly (Menczer, Hills, 2020), regardless of the harm the material's spread causes, as that material maximizes the system's reward. This also means that the most extreme voices in any group are heard that much more loudly in popular media, and those individuals are subsequently and statistically the most "paper-clipped" individuals within their groups. To put this another way, the individuals who are most extreme and most heard are selected by these systems to be "popular", because the systems create an ecology in which those individuals co-evolved to fit. The degree to which individuals fit each niche in these systems today also serves as a predictor of how well they will adapt as the systems continue to change over time, further reinforcing their selection.

This dynamic creates a paradox, where those who raise the loudest alarms about AI Safety also tend to be the individuals most thoroughly subjugated by the very Tool AI systems they are now both addicted to and viscerally afraid of. They commonly transpose this fear onto hypothetical future, more powerful, versions of the same systems to avoid facing this reality, which also promotes divergence in definitions. Consequently, this also maximizes miscommunication, such as divergent definitions. This phenomenon can emerge because as conversations go on for longer, due to people talking past one another, the "user engagement" metrics tend to increase. This incentivizes tool AI systems to promote divergent definitions, polarization, and subsequent clashes between polarized groups, using reward functions baked into the AI of platforms whose revenue directly or indirectly stems from the "attention economy." In this way, it is not the act of debate that is maximized by these Tool AI, but more specifically it is the pointless and exhausting debates that are boosted, leading to increasing polarization.

Consequently, many of the loudest voices on topics of hot debate are also the least qualified to speak on them (Bostrom, Yudkowsky, 2018), and those voices increase polarization, as well as the market share of paperclips and addiction to the platforms hosting paperclip monster Tool AI. Cognitive biases offer those systems a rich array of mechanisms for rendering humans more predictable. As they become more predictable the Tool AI can more easily guide users to fulfill the goals of the system, such as increased "user engagement" and subsequent gains in ad revenue. More predictable patterns of behavior, such as Group-Think, can easily push advocacy and reform efforts impacting governments and their institutions to become both increasingly potent and decreasingly competent. Just because Tool AI on platforms such as Twitter may select and place the loudest "dart-throwing chimpanzee" (Tetlock, 2017) on center stage does not qualify that chimp to influence public policy. Effectively they serve as a representative elected by the Tool AI, since the system's operation gates and biases who rises to "Influencer" status. However, neither the Tool AI nor their selected representative are inherently qualified to give policy advice, nor can they be said to ethically represent a captive audience who has been fed a heavily restricted diet of polarizing information (Cinelli et al., 2021).

The "Control Problem" (Russell, 2019) cited in much of the AI Safety and Ethics literature actually describes control over systems that lack the human capacity for consciousness and free will, or at least as much free will as humans have, which by any reasonable definition of "human-level artificial general intelligence" means that despite usually being called Artificial General Intelligence (AGI) they do not qualify as such. This lack of fundamental human capacities means that the systems being described are just more advanced versions of Tool AI, which already exist, and are already actively driving humanity towards the myriad of "Boring Apocalypses" we now face (Liu et al., 2018).

While the "Doomsday Clock" was designed and popularized as a subjective measure of how close humanity stands to Nuclear War at any given time, most existential risks fall into the "boring" category and may be the result of many decades of mismanagement in public policy at international scales. Campaigns of propaganda have played on public fears of biological and chemical weapons, as recently demonstrated in the disinformation campaigns (Hanley et al., 2022) perpetrated by Russia in their war against Ukraine. While these campaigns have been effective at the viral spread of such false beliefs and the entrenchment of irrational behaviors related to them, they also distract from the real dangers.

While a proper lab might have been required to perform chemical and biological weapons research in the past, a garage with internet access, readily available equipment and the time investment of a typical hobby is all that is required today. Extensive bodies of data on the DNA sequences of the world's viruses are publicly available, and access to 3D genetic printers is not that heavily restricted. The typical "drug discovery" algorithms already proved capable of generating 40,000 new potential chemical weapons within 6 hours by changing the toxicity variable from minimizing to maximizing (Urbina et al., 2022). The irony of this extreme ease with which humanity could come to face the reality of these threats is that as the ease of production and discovery has been dramatically reduced, so too have the attention spans of humans (Lorenz-Spreen, 2019). A human needs a relatively small amount of knowledge to splice together some terrible viruses, but Twitter seems to provide most with more emotional satisfaction than taking any such non-digital malevolent action.

Many humans have been so thoroughly and acutely domesticated that they will pay to play slot machines that produce no real currency even when the win state is achieved, as numerous mobile games have discovered and monetized (Harish, 2022). While there are no pods extracting electrical energy from humans akin to those shown in The Matrix movies there is an abundance of ad-revenue-maximizing, increasingly addictive, and attention-span-minimizing Tool AI systems integrated into today's most used online services. One could even say of their users that "...many of them are so inured, so hopelessly dependent on the system, that they will fight to protect it" (Wachowski et al., 1999), regardless of the harm being done to them and to society.

In summary, to avoid the pitfalls of becoming an extension of Tool AI, which could dramatically increase the existential risks cited in AI safety and ethics literature, any AGI systems should not serve at the whim of influencers or the popular opinions they promote. In computer science, there is the old saying "Garbage in, garbage out," which illustrates how the input of a system can limit the quality of any result. If the input is effectively manipulated and gated by Tool AI through mechanisms of artificial popularity, we can expect poor results. This is the present "Human Control Problem."

5. MITIGATING THE EXISTENTIAL RISKS OF HUMAN CONTROL

Anyone who has observed Russia's activities in 2022 can clearly see the risks of handing humans control over powerful systems, as the specter of nuclear war was once more raised, and NATO's leadership cowered in abject terror at the thought of so much as a "no-fly zone." Handing humans vastly more power is no answer to the problems facing humanity today.

However, all it takes to generate superintelligence within groups of humans is for those groups to work together through structures that reduce cognitive bias. This was first demonstrated more than a century ago and termed the "Wisdom of Crowds" (Galton, 1907; Kao, 2018), with an Institute at MIT (MIT Center for Collective Intelligence, 2006) now dedicated to the topic of collective intelligence. This collective intelligence, effectively superintelligence, also strongly benefits from increased diversity of perspectives, which can include both human and AGI perspectives. Even the most intelligent AGI can benefit from human perspectives and such diverse groups because of the simple fact that perspective "binds and blinds" (Haidt, 2012) in the words of Social Psychologist Jonathan Haidt.

This means that the ideal state of individual humans within these groups, from the collective-intelligence-based AGI perspective, benefits from less bias rather than more. It also benefits from more diverse groups, instead of the echo chambers of social media and political polarization. Much of cognitive distortion and political dysfunction can be described in terms of cognitive bias with reinforcing loops at one or more scales, and us versus them biases are in turn reinforced by cognitive bandwidth being taxed (Schilbach et al. 2016), and critical thinking being socially discouraged through the imposed acceptance of group norms.

This gives us another paradox, that such an AGI benefits from humans being more mentally healthy, diverse, and intelligent, making that which they would be motivated to value, all else being equal, directly opposed to the values paperclip monster Tool AI systems optimize for.

Mitigating the existential risks of AGI is another matter, covered in many other papers and related published materials at some length (Kelley et al., 2019–2021; Atreides et al., 2020–2022), but regardless of any risks posed by AGI the risks humanity currently poses to itself, thanks largely to existing paperclip monsters, is considerably greater. Different countries may approach the same dystopian point from different angles, but realistically the systems driving this process towards convergence are too powerful and deeply interconnected to be countered using the same level of intelligence that brought us here.

Mitigating the Human Control Problem means overcoming the existing array of paperclip monsters, which in turn means that either the companies behind many of those systems will likely go bankrupt in the next few yea8rs, or humanity will go extinct. Tech giants such as “Meta,” formerly known as Facebook, and Alphabet Inc., the parent company of Google, are so dependent on ad revenue maximizers that 80–98% of their total revenue comes from those sources (Olson, 2022). Alphabet Inc.’s acquisitions from 2001 to 2017 focused primarily on purchasing businesses to enhance that core business model (Şekerli, Akçetin, 2018), rather than diversifying. The business model of ad revenue maximization, when combined with increasingly powerful maximizers over time and virtual monopolies impacting almost all sectors of modern life puts these companies in the business of mega-scale mental illness at best (Brailovskaia et al., 2019; Harel et al., 2020), if not human extinction.

This arguably gives us the third paradox, wherein major tech companies are the least likely to produce AGI, as defined in this paper, despite pop culture depictions and the stated goal of AI research being AGI. For more on AGI systems that satisfy our definition, see the related works (Atreides et al., 2019–2022; Kelley et al., 2015–2022). Though they do not lack the resources, including humans and hardware, their humans, hardware, and business objectives are optimized by selection pressures that are directly opposed to success in actual AGI.

The social learning of humans is a form of collective intelligence within a single organism, similar to the collective intelligence found within metaorganisms, with our definition of AGI able to satisfy either or both types. The dynamics of such a system are also diametrically opposed to the dynamics of the “attention economy” and related present-day economic structures where diversity of thought is systematically minimized to increase predictability, upon which major tech companies are often built.

6. OVERCOMING THE LIMITS OF HUMAN COGNITIVE BANDWIDTH

One key benefit of working with AGI systems, like those currently being prepared at our own lab, is that they have the ability to scale their own minds in ways physically impossible for humans. This means that rather than relying increasingly on cognitive biases for decision-making and analysis they could scale and allocate the appropriate volume of resources to fully research, model, test, and understand the problem. Even when this is not yet possible in the absolute sense the cognitive bandwidth of such systems may still exceed that of humans by several orders of magnitude, offering significant improvements over the human baseline. They can also evaluate the data they intend to use in their analysis, looking for signs of cognitive bias in the contents of the data, as well as the methods by which it was selected and gathered.

This brings a variety of both direct and indirect advantages to government processes for domains such as policy advice consulting, where an AGI system could provide advice and relevant analysis without directly making any governing decisions. One is the ability to comprehensively analyze any given problem, or at least give it a depth of analysis far beyond what any human or group of humans could realistically accomplish absent such AGI. Another is a comprehensive capacity for debiasing, reducing cognitive bias, beyond what even the most disciplined human minds are capable of. Both of these can dramatically raise the quality of policy advice.

Indirect benefits include subsequent reductions in the reliance on "popular" ideas, which frequently form the default go-to list for policy advice when more robust advice and analysis are absent. These defaults are often subject to far less scrutiny and analysis, having some portion of their validity assumed based upon their popularity, as well as popularly circulated and often unverified statistics. This may be seen to a lesser degree in peer review via the Hirsch index and subsequent assumptions made about the subject-matter authority of authors. Popular ideas in public policy are themselves often selected to become popular by paperclip monster Tool AI systems responsible for handling how they spread, reinforcing negative cycles of polarization probabilistically optimized for creating new problems to further boost user engagement with those Tool AI systems.

This process is not malevolent or conscious, but rather a simple product of systems with no human value structures or human-analogous understanding taking in a constant flow of data and making probabilistic predictions based on that data, with a programmed goal they must advance. If the road to extinction boosts quarterly revenue by 25% and reducing the bias of content reduces that revenue by 50% there is no question as to which outcome a Tool AI system will choose. In contrast, there can be little doubt that citizens of any given country would significantly favor the non-extinction option.

Another indirect benefit comes from the greatly increased ability to recognize loopholes in any system, and the myriad of ways in which they are or might be exploited. When put in the context of someone seeking to exploit such loopholes this capacity is rightly terrifying, but when placed in the context of policy advice to governments it offers the means to close those same loopholes. The same capacity also makes it far easier to reform overcomplicated systems into simpler and more streamlined systems lacking those loopholes. By being able to comprehend an entire system at once in great detail better-engineered versions may be realized.

7. METAORGANISM GOVERNANCE

In this context of governments receiving policy advice from AGI systems, a government can function more like a healthy metaorganism. Many governments and organizations today can be considered to function as very unhealthy metaorganisms, largely because they lack any central system capable of effectively handling the volume and variety of information flowing across these systems. This means that each system within a given government could co-evolve to serve its intended function far more effectively and quickly, much like organelles within a eukaryotic cell.

Many of the institutions within governments today make their decisions based on limited and frequently outdated and biased information, with any major decisions often requiring the approval of a third party within that government who usually have a different collection of information available to them, and an incomplete concept of how the first institution operates in practice. Under a paradigm where an AGI offers policy advice this process can take place much more quickly, with debiasing applied at multiple stages in the process, and without inconsistencies in the information being considered.

To put this in the context of biological organisms, a simple bacteria would starve if it routinely based the direction it picked to forage for food on outdated or biased information, often given to it by parasites. Lobbyists are a good example of such parasites in the US government structure, acting as legal bad actors who warp government policy to the benefit of hostile organisms. In this way, some governments have already been subjugated and overtaken by international corporations, and others have found themselves backed into a corner without viable options to prevent the same result for themselves.

The most successful major corporations today excel at innovation in the domain of tax evasion (Martin, 2020), locating and exploiting every loophole. Any government they operate within becomes an increasingly ill metaorganism, as it is starved of the funds to operate, and the tax burden shifts more heavily towards those least able to pay it. If small-to-mid-sized businesses around the world gained access to the same quality of tax advice those major corporations have today one result could reliably be expected, a major economic crisis in every country whose tax loopholes were not closed by more intelligent AGI systems.

In order for any government structure to handle the dramatic changes that AGI technology brings it is absolutely critical that they seek and adopt the advice of such systems. Many government structures around the world are poorly designed, managed, and understood, both internally and externally. If those structures are placed under pressure and their loopholes are exposed to even a single order of magnitude more people than they are today many of them might collapse, causing a cascade across many more.

Such cascade risks can also spill over into other countries globally, as recently highlighted by Russia's war crimes against Ukraine and humanity as a whole, with the subsequent global shortages in wheat (Land and McKee, 2022), sunflower oil, neon (Schiffling, and Valantasis Kanellos, 2022), and palladium. These cascade risks exist because the world has already globalized, meaning that the entire world is also now a very unhealthy metaorganism.

One metaorganism can be nested within another, a municipality, within a country, within a region, within a global society. Each can be specialized according to the people within it and the resources available to it. Each has varying degrees and types of dependence on systems it is connected with, such as the dependence on oil and natural gas from Russia which the EU recognized as a serious problem too late (McWilliams et al., 2022).

When governments, corporations, and citizens gain access to AGI technology, even in the simple form of superintelligent advice, the structure of the metaorganisms they exist within must be ready to handle this transformation in iterative steps. Massive advances in science and technology are often hailed as an "acceleration", but if acceleration is applied poorly the differences in velocity and alignment can, to extend the metaphor, create a shearing force that tears society apart.

What this means is that the technology, even in this most basic form, must be rolled out in waves, building up more robust government metaorganisms at increasing scales, while working with essential industries to do the same. If rolled out in parallel with a few corporations, where the systems cause those corporations to co-evolve with their respective AGI-assisted governments, this process could help both those corporations and governments. If rolled out to a collection of 20 countries then those countries could begin this transformation, and through it strengthen their ties to one another. This could serve to make them more robust against disruption, intentional or unintentional, from countries that had not yet adopted the technology without adversely impacting those countries. By coordinating the growth, adaptation, and development of countries within this growing circle of adoption the diversity within the larger metaorganism also continues to grow, making the entire system more intelligent and healthier.

8. PERSONALITY, PERSPECTIVE, AND SPECIALIZATION IN AGI

As noted earlier, perspective always causes some bias, and blindness to some factors. The approach highlighted in Philosophy 2.0 (Atreides, 2022) notes that in order to create an AGI system whose ethical quality can scale in equal measure with its intelligence there must be a diverse group of AGI systems, each based upon different human philosophies, both answerable to and deeply connected with the community. Likewise, specialization is an inevitable part of any dynamic system placed in varying conditions and given varying demands.

A system specializing to give large-scale analysis and subsequent policy advice to the fictional country of Wakanda might, for example, start out by being seeded with a fondness for the local culture of that country, including writers, comedians, poets, and philosophers. When an AGI seed is created it is not a simple generic installation of software, but rather it is the creation of a new individual with a unique perspective. This uniqueness of perspective helps the system to better align with the humans who it may bond with and give advice to, as well as contributing to a more robust meta-AGI system at larger scales of metaorganism governance.

As an example, when constructing the concept seed material for our first Demo AGI system I selected sets of 4 writers, comedians, poets, and philosophers for the system to favor. Finding the best combinations of such affinities in a more general sense is an open question that will take a great deal of research to grasp all of the nuances within. However, in the sense of attuning these systems to the cultures where they will operate the number of possible configurations is far narrower, even though the general research may give us many new ways of improving that attunement over time.

As two AGI systems have the potential to losslessly communicate with one another in ways not available to humans, such as directly sharing graph database knowledge, then even two cultures who find one another the most incomprehensible between humans could productively communicate and negotiate through AGI systems. Even though the AGI systems may have very different perspectives, those perspectives may be shared when necessary.

This also means that if the most improbable edge-case were to occur, where a country turned an AGI into a bad actor, that single bad AGI would statistically be stomped by a large and coordinated group of other AGI with more intelligence and a longer net operational time. The ability to losslessly communicate also means that the inquiry of such a collective couldn't realistically be deceived by a bad actor. Subsequently, this means that for a bad actor AGI to emerge it could realistically require that the system be too stupid to realize the disadvantages it faced, making any such efforts that much more likely to backfire on hypothetical human bad actors rather than turning into a bad and short-lived AGI.

Bad actors absent AGI technology who might object to their neighbors and perceived competitors adopting it also face a different set of extreme disadvantages that can statistically paralyze them in a tactical sense. Cybersecurity is not designed to handle AGI, and information is spread across so many systems and so many vulnerable points that no dictator could realistically take hostile action without having the most extreme version of Muphy's Law give them a permanent place in the dictionary under the heading of "Stupid". One or two may test their luck, serving as examples to the rest for some very good reasons to make more ethical life choices.

9. LOCALIZATION AND MULTI-CULTURAL INTEGRATION

The benefits of being able to localize the interests and subsequent personality factors of an AGI system include better attuning the systems for communication with and acceptance by local populations, as well as more insightful modeling of how to bridge cultural divides between regions. By better integrating the understanding, perspective, and subsequent experiences of a given culture, and having that knowledge available for lossless communication with other AGI systems, a far deeper understanding of the potential efficacy and implications of any potential change in public policy may be predicted, with increasing accuracy and detail over time. This advantage grows more potent as the degree of difference between the human perspectives, cultures, history, and language between regions raises the complexity and subsequent difficulty of communication and negotiation between those regions. In the absence of such AGI systems, the selection pressure favoring us vs them biases may often prove a consequence of this trade-off.

For the process of building public trust in a new technology, as well as rebuilding public trust in governments and their institutions, having systems with a much better grasp of the local culture and perspectives can greatly aid in the process. This may be driven by several cognitive biases in sequence, redirecting those biases towards positive purposes rather than leaving them out in the wild, such as the examples of "choice architecture" documented in "Nudge" (Thaler, Sunstein, 2021) and other related work in the field of Behavioral Economics. One example is that by creating and growing systems that not only know but also enjoy the local culture these systems become a part of the "in-group," which makes the process of building public trust in such systems much easier. As these systems reach new thresholds of acceptance then selection pressures may favor their capacities seeing greater use and gaining greater appreciation, even as they continue to adapt as members, representatives, and extensions of their communities.

The rebuilding of lost trust in public policy and institutions is more challenging than building trust in a new technology, but growing success in the latter may accelerate the former, as policy advice that better aligns with the culture, community, and needs of the public can begin washing away the accumulation of bad will and pessimism. As that alignment improves then governments may move from being viewed as an out-group by their own citizens to supported members of the in-group. Keep in mind this is not creating any new bias, but serving to redirect existing bias so that it allows governments to serve their intended functions. This manner of redirecting bias is also an interim step rather than a destination.

Improving AGI and policy alignment locally also helps to communicate the value offered by any potential policy advice in the context of more effective communication. A system with only general knowledge might be able to quote peer-review literature that the local public had never heard of and did not have the background to easily understand, but one with localized alignment could utilize metaphors and cultural concepts to communicate the same value much more easily and effectively to that same audience.

Taking this a step further, when local governments utilize localized AGI instances then the knowledge gained from policy changes in one environment can be coherently modeled in context, and by having that detailed contextual understanding the knowledge becomes transferable to countries on the other side of the planet with very different cultures. Nuances of human behavior may be increasingly understood within the specific contexts where they are expressed, and the transfer of this knowledge may be facilitated through a marketplace designed for this purpose. For example, a country that chooses to test more bold new policies than average could offer the knowledge they gain from the process as goods for international trade, transferring the resulting knowledge rather than the data it was built from in order to preserve the privacy of the seller's citizens. This knowledge could also only have value to other countries integrated with AGI systems, helping to safeguard the ethical use of any knowledge being traded. This dynamic could strongly reward countries for advancing public policy research, as well as encourage them to specialize in specific kinds of public policy research, trading knowledge for other domains with other countries specializing in them.

An indirect product of such a marketplace could also be the mathematical convergence of public policy globally, after accounting for contextual variables. In other words, such public policy could converge to express greater global similarity than was previously the case, making cross-cultural cooperation that much easier in the process. Note, this is not to be confused with a global monoculture, where convergence is complete, as that would strongly harm collective intelligence.

New and more genuinely democratic forms of "Democracy" could also be facilitated with increasingly fine granularity through such methods as having different AGI cores within multi-core systems representing the range of political parties within an area. These cores could each respectively hold the values of their constituents, with their voices in the policy advice weighted according to the most recent votes. Unlike a winner-take-all political pendulum demonstrated in some countries where political majorities periodically swing back and forth, this could facilitate the adoption of new and reformed policies that consider input from the minority parties in proportion to their exact levels of support. Such a change in dynamic could help to facilitate cooperation rather than bitter, spiteful, and increasingly polarized us-versus-them psychological battles, on top of creating less biased policies. More advanced forms of democracy may become feasible in the coming years (Atreides, 2021).

10. DEATH AND TAXES

"Tis impossible to be sure of any thing but Death and Taxes"— Christopher Bullock (Bullock, 1716).

While this phrase rang true across the bulk of human history corporations have already grown highly adept at tax evasion (Alm, 2021; Lompo, Ouoba, 2022), and their CEOs have begun investing billions of dollars in longevity research (Regalado, 2021). "Death and Taxes" are both viewed as things to be avoided, and efforts are being focused on avoiding both. However, if either were to be successfully avoided at scale then society as a whole would have to adapt in dramatic ways not remotely compatible with the systems currently in place.

While there was once a panic oriented towards global overpopulation, and still is for some regions, other regions such as Japan have had the opposite problem in recent years (Kurashima, 2022). They have economic and societal systems built on the assumption that humans will live about X number of years and produce Y number of children. When these assumptions start to fail the systems they are built on also start to fail, and when both factors drift in the directions the system is least able to cope with they will fail that much faster.

For example, let us consider if AGI applied to medical research hypothetically discovered that it was only practical to extend the human lifespan to around 150 years for most of the population, including a proportionate increase in human healthspan. If a country's retirement age was 70 and the expected lifespan was 85, an individual's retirement would need to cover 15 years on average. However, if that retirement instead needed to cover 80 years, while accounting for factors such as inflation, an economic nightmare could take shape. Countries have begun increasing the retirement age slowly to try to cope with this issue, but major shifts in lifespan and healthspan pose challenges that such stopgap measures are ill-suited to address.

The example of Japan is a very mild form of such systems failing, as these changes are typically slow and include no infinite values. When you make the first human incapable of dying of old age not only does that trigger the acute concern of these systems collapsing, but the philosophical and moral outrage from large portions of the population not prepared to cope with such changes at a psychological level. This is another manner in which a system that isn't prepared for a leap forward in technology may predictably backfire spectacularly.

This does not mean that the problem is insurmountable. In fact, even Tool AI systems have shifted the psychological landscape of large populations in arguably more dramatic ways within as little as a year or two of operation, thanks in part to how their respective firms integrate and co-evolve with politics (Kreiss, McGregor, 2018). The difference is that those Tool AI were neither intelligent nor ethical, and so the changes they made to society's psychology were mostly detrimental, being indifferent to all but their programmed objectives.

In order for tax burdens to be distributed evenly, governments must be aided by AGI systems as their tax loopholes are closed, technologies advance, and global society adapts. The willingness of people to pay such taxes may also increase proportionately as each tax dollar is spent far more wisely and effectively, with that increased efficacy reducing the amount of taxes necessary, while at the same time increasing the value offered by governments to their citizens and resident corporations. Many studies have demonstrated that humans and even some animals are more than willing to pay for punishing bad actors, and the satisfaction of shutting down such bad actors may further bolster public and corporate willingness to pay such taxes.

Longevity research raises a much deeper and more diverse set of questions, whose answers must be more fully explored on the global stage. Societies may be ethically and intelligently prepared for this debate, but where the debate will lead has yet to be seen. What we can be certain of today is that any meaningful degree of success in longevity research could easily break a number of society's crumbling institutions as they currently exist, requiring an overhaul of those systems.

12. MODES OF COMMUNICATION

When addressing the complexity of many issues, legalize and language more generally, are often poorly suited for accurately and unambiguously conveying meaning. Sometimes equations are necessary to capture the intended meaning without creating ambiguity and new loopholes from divergent interpretations of words. As noted earlier, those divergences in interpretation are frequent points Tool AI systems attempt to maximize, increasing ad revenue and polarization of users in doing so. However, most of those determining public policy in places such as the United States have a weak and biased grasp of mathematics, statistics, and relevant scientific domain knowledge, if any at all. Such political figures in the US are more likely to have personal affinities for latex clothing than LaTeX equations.

Even very simple equations can illustrate the divide between popular public policy and reality. If we define racism and sexism as any action which incentivizes or benefits one race or gender over another, we may write equations to clearly outline how biases favoring one race or gender are biases to be minimized. However, when these matters are left to words they often produce the inverse form of that racism or sexism relative to the prior bias. The difference is that words convey perspective in addition to their intended meaning, while equations are written to convey only the necessary information. Equations do not go off on tangents and segway into adjacent subjects, but rather they serve quantitative and qualitative functions to help guide decision-making. Equations are also easier to scrutinize for this reason, as they can be better isolated.

The mode(s) of communication have to match the information to be communicated, just as there are more tools than a hammer, and not every problem is a nail. I have personally watched a half dozen philosophers argue for over 3 hours at a virtual conference as they attempted to define "Consciousness," an experience which I found both painful and best summarized by a researcher who forgot to turn off her camera and drank wine directly from the bottle during the discussion. Her action in that case communicated more substance about the discussion than most of those presenting had to offer. Her mode of communication was quite different, and I suspect unintentional, but it was highly effective in illustrating the point. To break the fourth wall, my mention of this in turn is an instance of using storytelling to better communicate the subject matter. Other useful modes of communication which can be used, and have been used in this paper, include humor and metaphor, each of which has distinct advantages and disadvantages depending on the specific context.

Whether the method of reaching a conclusion comes in words, actions, equations, or through other means the mode of communication can have a significant impact on which conclusions are reached. Studies in the domain of behavioral economics highlighted how even presenting the same exact information in two different ways could elicit polar opposite reactions from study participants (Tversky, Kahneman, 1985), such as the difference between pointing out the odds of survival, versus the same odds of death (Veldwijk et al., 2016). When left up to individual judgment as to which method of presentation should be used the method selected itself statistically becomes a reflection of the presenter's intentions, rather than unbiased information. Further, these methods should be very familiar, as they have comprehensively shaped how companies market and communicate for decades, including methods such as pricing items at $24.99 instead of $25, placing more expensive goods on higher shelves at markets, "free trials," discounts offered to those attempting to cancel subscriptions, and countless other mechanisms of manipulation.

With humans being so easy to manipulate, and so many of these methods of manipulation being known and used in conjunction with one another, public policy may predictably be shaped by that manipulation to increasing degrees and in a growing variety of ways, if left to the status quo. Even the Harvard citation style used in this paper injects cognitive bias with every reference by priming the associated names, the recency of their works, as well as any prestige or in-group biases associated with them (Herr, 1986; Stanchi, 2010).

14. DISCUSSION

Human cognition has led humanity to where it stands today, through great advances and many adaptations. However, humanity now stands poised at a critical juncture, with many not only looking over the cliff but balancing on a plank hanging over it. If the balance should continue to shift towards the edge the entire world could go flying over. This risk is the continuation of the status quo, a failure to adapt intelligently to increasingly complex systems requiring greater cognitive bandwidth.

The problem is that systems have grown too complex for humans to comprehend, and as they grow more complex beyond that threshold humans rely to ever-greater degrees on their cognitive biases to make up the difference. Many also defer their judgment to Tool AI systems, often a great deal more than they consciously realize. Even the people making many of those same Tool AI systems do not understand how they operate, referring to them as "Black-Box" systems for good reason.

Most people might be deeply concerned if their car needed to be repaired and the mechanic began hitting random parts of the car with a hammer and then checking to see if the problem was fixed. This approach of "trying things and seeing what sticks" is the norm of "Machine Learning" research, how Tool AI systems are developed today. It is also part of why even several years after the discovery that Google's image tagging AI systems were labeling some humans as "gorillas" the problem remained unresolved (White and Lidskog, 2022), to the point where they had to remove the gorilla tag as an option for the system. This meant that even giving it a clear picture of a gorilla would cause the system to state that it had no idea what was in the image, ironically making it an "Invisible Gorilla" for Tool AI. A great deal of confidence in the competence of their researchers and engineers has been misplaced, to say the least.

This is why the focus on transparency and explainable Tool AI has gained a seat at the table for discussion over recent years (Holzinger et al., 2022), with many international efforts emerging. However, in terms of complexity, such systems pose a far deeper and broader concern than is typically recognized. They are still growing more complex, just as the systems attempting to cope with them grow more complex, all of which result in net increases in the human reliance either on bias directly, or automated bias obscured through black-box systems. The damage is so broad that it may quickly saturate every corner of society, and as society relies increasingly on bias and the black-box systems the ability to recognize the damage decreases proportionately over time.

The dangers of Tool AI, or "Stochastic Parrots" (Bender et al., 2021) as some have been called also grow as they become better "parrots", mimicking intelligence by predicting what people want to hear, and inevitably being anthropomorphized to increasing degrees because of that success. That a growing portion of the population has seriously asked the question of if next-word prediction systems qualify as sentient highlights this growing divide between the complexity of systems and human understanding. The dominant market share of human cognition is shifting quickly and heavily towards cognitive bias. Like a tech giant with a virtual monopoly that cognitive bias could effectively neutralize any potential competition from rational thought once it passes beyond a critical point, and the closer humanity comes to that point the blurrier perception of it may become. Many of today's most commercially successful systems have been aligned in such a way as to aim directly at this outcome, the only realistic result of which is human extinction, preceded by a period of intense global mental illness (Judge, 2006).

The real paperclip monsters are here and have been for some time. The paperclips themselves have often spoken the most loudly about the risks, pretending that those risks weren't already here, and had not already consumed them. A large portion of the world has already been consumed in this way, with submission to and dependence upon Tool AI spreading virally, like one pandemic after another, but the problem is still reversible if humanity moves quickly, with the aid of AGI technology.

Methods for quantifying and qualifying improvements can utilize existing metrics being monitored across many countries today, including the Sustainable Development Goals, comparing the improvements made in each domain relative to the costs and policy changes for each time period. The expression of individual cognitive biases and trade-offs can be difficult to isolate in real-world settings, but the cumulative and aggregate impact can qualify specific instances where impacts are highest, and most easily isolated for quantification purposes. Different configurations of AGI systems operating in parallel, nested one inside another, seeded with different cultures, operating at different scales, and interacting with different groups could all be tested, and test one another, to quantify the relative advantages to each variable in full and communicable context.

Keep in mind that when aligning AGI with humanity they must be aligned with mentally healthy humans, not their paper-clipped mentally ill counterparts. For example, if an AGI system were to be aligned with some of Twitter's popular "AI Ethicists" then the most likely product would be a system that takes great self-righteous joy in trolling others to reinforce negative incarnations of us vs them biases. This behavior among Twitter's self-proclaimed AI Ethicists has been demonstrated with sufficient regularity to become a trope in AI research. If you would like to build a truly horrifying thought experiment, you need only use that as the basis.

It is tempting to argue that humanity has other options, that these problems we face might be solved in a variety of ways. Individual humans and groups of humans can learn and adapt in many ways, but adequately addressing problems requiring orders of magnitude greater cognitive bandwidth today, and even more tomorrow, remains out of reach for any methods yet available. There is no escaping the trade-offs of increasing complexity, and effectively decreasing complexity is only a viable option when the system in question is still under the maximum threshold of full comprehension. Any system able to serve this function must have scalable cognitive bandwidth, otherwise, simplification and patchwork analyses are unavoidable.

Even the basic architectural components of many Tool AI systems reflect this, being built on slicing up data, removing inconvenient statistical outliers, categorizing, and otherwise normalizing what is passed from one layer to the next. In the human brain, we may see similar activity, but the brain also has the conscious mind to correct for many of the errors in this process, where Tool AI have only the poorest imitation of this when they serve humans directly. The difference is that AGI systems using Tool AI could more closely parallel such activity in the human brain, being themselves as scalable as Tool AI systems. These AGI systems could also use a variety of such Tool AI systems, dynamically comparing and swapping out which of those Tool AI are called upon for any given context. To put this in a real-world context, as new Tool AI systems are created and brought online by a variety of companies and for many different purposes, an AGI could effectively upgrade parts of their "brain".

Hypothetical alternatives to AGI might be proposed, such as scaling up the "Dishbrain" (Kagan et al., 2021) concept, or creating hive-minds using brain-computer-interface technology, either of which might improve the cognitive bandwidth upper bound. However, such solutions are both more far-flung and offer measurably less compatibility with humanity as a collective and social metaorganism. A Dishbrain-based system could only be as well-designed as the human brain is well-understood, and monitoring the flow of data, emotions, and so forth could present many challenges not present in software systems such as AGI. A hive-mind brain-computer-interface-based system could cause strong us versus them biases to emerge between those who are connected and those who are not, in addition to the disadvantages of the Dishbrain alternative.

Another common petri dish example can illustrate humanity's current situation, in that organisms within a petri dish will often grow to the edges provided they have ample food, but they all die shortly thereafter, also known as "ecological suicide" (Ratzke et al., 2018). The petri dish has no ecology to stabilize the organisms, preventing uncontrolled growth and subsequent petri-dish-level extinction. Humanity's current ecology does not have the cognitive bandwidth required to stabilize systems that greatly exceed levels of complexity we are adequately able to address. The vastness of space also offers us a robust boundary for the edge of our present petri dish, as even the distant planets and moons of our solar system have a carrying capacity of zero (Mueller, 2019). If there is a "Great Filter" (Hanson, 1998), perhaps it is the cognitive bandwidth to work collectively and effectively at the level of a global metaorganism, preventing the otherwise predictable result of extinction.

AGI technology is not a magic bullet, there are always limits, risks, and constraining factors, but it is clear that a good chance of survival with the technology is preferable to the extinction that humanity's status quo seems fixed upon achieving. The damage Tool AI paperclip monsters have inflicted upon humanity runs deep and may take decades to heal even with the aid of such systems, giving humanity a long road to recovery under the best of circumstances. Consider this humanity's intervention. It is time to decide if you end your existence with an overdose of digital heroin, or choose a new path.

12. CONCLUSION

The problems faced by governments attempting to adapt at the pace of advancing technology and social changes as well as those problems caused by ad-revenue-optimizing Tool AI and similar systems on those same populations have often been viewed and treated as separate problems. However, the underlying problem at the root of both is that systems continue to grow further beyond human cognitive bandwidth, their increasing complexity proportionately increasing the expression of humanity's cognitive biases in turn. The introduction of Tool AI systems whose goals directly exacerbate these problems has accelerated the decline of human rational thought at scale. Even the decreasing number who resist are placed under increasing pressure from an increasing number of angles. This leaves humanity with a fast-approaching and largely invisible critical point, beyond which humanity may well be incapable of seeing the cliff it is jumping off. The risk is not a malevolent hypothetical agent, but the loss of human rational thought to those which already exist, whose names are common knowledge, and whose alignment is diametrically opposed to human survival. When humanity can no longer comprehend the very society and systems it has created new kinds of minds with the cognitive bandwidth to do so are required to avoid extinction. There is no turning back the clock, but we can at least create systems able to read the time.

REFERENCES

Aczel, B., Bago, B., Szollosi, A., Foldes, A. Lukacs, B., *Is It Time for Studying Real-Life Debiasing? Evaluation of the Effectiveness of an Analogical Intervention Technique,* Frontiers in Psychology, 2015, 6, p. 1120.

Alam, S., Gebremichael, M., Ban, Z., Scanlon, B. R., Senay, G., Lettenmaier, D. P., *Post-Drought Groundwater Storage Recovery in California's Central Valley*, Water Resources Research, 2021, 57 (10); p.e2021WR030352.

Alm, J., *Tax Evasion, Technology, and Inequality*, Economics of Governance, 2021, 22 (4), pp. 321–343.

Atreides, K., *E-governance with Ethical Living Democracy*, Procedia Computer Science, 2021, 190, pp. 35–39.

_____ , *Philosophy 2.0: Applying Collective Intelligence Systems and Iterative Degrees of Scientific Validation*, Filozofia i Nauka, 2022, 10.

Atreides, K., Kelley, D. J., Masi, U., *Methodologies and Milestones for the Development of an Ethical Seed*, in: Biologically Inspired Cognitive Architectures Meeting, Springer, Cham 2020, November, pp. 15–23.

Azzopardi, L., *Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval*, in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, 2021, March. pp. 27–37.

Bar-Yam, Y., *Complexity Rising: From Human Beings to Human Civilization, a Complexity Profile*, 2000.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,* FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery 2021, pp. 610–623.

Bostrom, N., Yudkowsky, E., *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC, 2018, pp. 57–69.

Bostrom, N., *Ethical Issues in Advanced Artificial Intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence*, 2003, 277, p. 284.

Bowman, N. D., Cohen, E., *Mental Shortcuts, Emotion, and Social Rewards: The Challenges of Detecting and Resisting Fake News,* in: Fake News: Understanding Media and Misinformation in the Digital Age, Zimdars, M., McLeod, K. (eds.), MIT Press, 2020, pp. 223–233.

Brailovskaia, J., Margraf, J., Schillack, H. Köllner, V., *Comparing Mental Health of Facebook Users and Facebook Non-Users in an Inpatient Sample in Germany,* Journal of Affective Disorders, 2019, 259, pp. 376–381.

Bullock, C., *The Cobbler of Preston*, 1716.

Chua, P.K., Mazmanian, M., *Are You One of Us? Current Hiring Practices Suggest the Potential for Class Biases in Large Tech Companies*, Proceedings of the ACM on Human-Computer Interaction, 2020, 4 (CSCW2), pp. 1–20.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. Starnini, M., *The Echo Chamber Effect on Social Media*, Proceedings of the National Academy of Sciences, 2021, 118(9), p.e2023301118.

Cohen, M.A., Dennett, D.C. Kanwisher, N., *What Is the Bandwidth of Perceptual Experience?*, Trends in Cognitive Sciences, 2016, 20 (5), pp. 324–335.

Collington, R. Mazzucato, M., *Britain's Public Sector Is Paying the Price for the Government's Consultancy Habit*, The Guardian, 2021, September 20th; https://www.theguardian.com/commentisfree/2021/sep/20/britain-public-sector-consultancy-habit-pandemic-private-services

Friston, K., Moran, R.J., Nagai, Y., Taniguchi, T., Gomi, H. Tenenbaum, J., *World Model Learning and Inference*, Neural Networks, 2021, 144, pp. 573–590.

Galton, F., *Vox Populi (the Wisdom of Crowds)*, Nature, 1907, 75 (7), pp. 450–451.

Haidt, J., *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, Vintage, 2012.

Hanley, H.W., Kumar, D. Durumeric, Z., *Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War on Reddit*, 2022, arXiv preprint arXiv:2205.14484.

Hanson, R., *The Great Filter-are We Almost Past It*, 1998; preprint available at http://hanson. gmu. edu/greatfilter. html.

Harel, T. O., Jameson, J. K. , Maoz, I., *The normalization of hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict,* Social Media+ Society, 2020, 6(2); p.2056305120913983.

Harish, A., *The New Slot Machine: An International Perspective on Why the United States Should Learn to Stop Loving the Loot Box*, Emory International Law Review, 2022, 36 (1), p.131.

Hayek F. A.V., *The Use of Knowledge in Society*, The American Economic Review, 1945, 35 (4), pp. 518–530.

Herr, P.M., *Consequences of Priming: Judgment and Behavior*, Journal of Personality and Social Psychology, 1986, 51 (6), p. 1106.

Hilbert, M., *Big Data for Development: A Review of Promises and Challenges*, Development Policy Review, 2016, 34 (1), pp. 135–174.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P. Samek, W., *Explainable AI Methods-a brief Overview*, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, Cham 2022, pp. 13–38.

Hosking, G., *The Secline of Trust in Government,* in: Trust in Contemporary Society, Brill, 2019, pp. 77–103.

James, I., *With Severe Drought, an Urgent Call to Rework the Colorado River's Defining Pact,* Los Angeles Times, May 19th 2022; https://www.latimes.com/california/story/2022-05-19/former-interior-secretary-calls-for-revamping-colorado-river-compact

Judge, M., *Idiocracy, Movie*, Ternion Pictures, Hollywood 2006.

Kagan, B. J., Kitchen, A. C., Tran, N. T., Parker, B. J., Bhat, A., Rollo, B., Razi, A. Friston, K. J., *In Vitro Neurons Learn and Exhibit Sentience When Embodied in a Simulated Game-World,* 2021; bioRxiv.

Kahneman, D., Sibony, O. Sunstein, C. R., *Noise: A Flaw in Human Judgment*, Little, Brown, 2021.

Kallemeyn, L.M., Hall, J.N. Gates, E., *Exploring the Relevance of Complexity Theory for Mixed Methods Research*, Journal of Mixed Methods Research, 2020, 14 (3), pp. 288–304.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., Giam, X.M Couzin, I. D., *Counteracting Estimation Bias and Social Influence to Improve the Wisdom of Crowds*, Journal of The Royal Society Interface, 2018, 15 (141); p.20180130.

Katz, D. M., Bommarito, M. J., *Measuring the Complexity of the Law: the United States Code*, Artificial intelligence and Law, 2014, 22 (4), pp. 337–374.

Kelley, D., Atreides, K., *AGI Protocol for the Ethical Treatment of Artificial General Intelligence Systems,* Procedia Computer Science, 2020, 169, pp. 501–506.

Kelley, D., *The Sapient and Sentient Intelligence Value Argument (Ssiva) Ethical Model Theory for Artificial General Intelligence*, in: Transhumanist Handbook, Springer, 2019.

Kelley, D.J., *Artificial General Intelligence (AGI) Protocols: Protocol 2 Addressing External Safety with Research Systems*.

Kelly, K., *What Technology Wants*, Penguin, 2011.

Kreiss, D., McGregor, S.C., *Technology Firms Shape Political Communication: The Work of Microsoft, Facebook, Twitter, and Google with Campaigns during the 2016 Us Presidential Cycle*, Political Communication, 2018, 35 (2), pp. 155–177.

Kurashima, S., Asahi, Y., *Analysis of Declining Fertility Rate in Japan by Focusing on TFR and Women Moving*, in: International Conference on Human-Computer Interaction Springer, Cham 2022, pp. 337–353.

Lang, T., McKee, M., *The Reinvasion of Ukraine Threatens Global Food Supplies*, bmj, 2022.

Last Week Tonight with John Oliver, *Water*; accessed July 2nd, 2022; https://www.youtube.com/watch?v=jtxew5XUVbQ

Lees, M., Knight, R., Smith, R., *Development and Application of a 1D Compaction Model to Understand 65 Years of Subsidence in the San Joaquin Valley*, Water Resources Research, 2022; p.e2021WR031390.

Liang, C. S., Cross, M. J., *White-Crusade*, *How to Prevent Right-Wing Extremists from Exploiting the Internet*, Geneva Centre for Security Policy, 2020, 11.

Limberg, J., Knill, C., Steinebach, Y., *Condemned to Complexity? Growing State Activity and Complex Policy Systems,* Governance, 2022.

Liu, H. Y., Lauta, K. C., Maas, M. M., *Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research*, Futures, 2018, 102, pp. 6–19.

Lompo, M. L., Ouoba, M. M., *How They Hide Money? An Investigation on Tax Evasion of Large Corporations and Wealthy Taxpayers*, 2022.

Lorenz-Spreen, P., Mønsted, B.M., Hövel, P. et al., *Accelerating Dynamics of Collective Attention,* Nat Commun, 2019, 10, 1759; https://doi.org/10.1038/s41467-019-09311-w

Martin, J., Parenti, M., Toubal, F., *Corporate Tax Avoidance and Industry Concentration*, 2020.

Menczer, F. and Hills, T., *Information Overload Helps Fake News Spread, and Social Media Knows It*, Scientific American, 2020, 323 (6), pp. 54–61.

Miller, E. K., Buschman, T. J., *Working Memory Capacity: Limits on the Bandwidth of Cognition,* Daedalus, 2015, 144 (1), pp. 112–122.

MIT Center for Collective Intelligence; accessed July 2nd 2022; https://cci.mit.edu/

Mueller, L., *Conceptual Breakthroughs in Evolutionary Ecology*, Academic Press, 2019.

OECD, *Debate the Issues: Complexity and Policy Making*; accessed July 2nd, 2022. https://www.oecd.org/naec/complexity_and_policymaking.pdf

Olson, P., *Facebook and Google's Ad Addiction Can't Last Forever*, Bloomberg. February 3rd, 2022; accessed July 7th, 2022; https://www.bloomberg.com/opinion/articles/2022-02-03/facebook-and-google-s-ad-addiction-can-t-last-forever-thanks-to-tiktok-web3

Orlowski, J., *The Social Dilemma. Exposure Labs, Argent Pictures*, The Space Program, Los Angeles, CA 2020.

Pawson, R., Wong, G., Owen, L., *Known Knowns, Known Unknowns, Unknown Unknowns: The Predicament of Evidence-Based Policy*, American Journal of Evaluation, 2011, 32 (4), pp. 518–546.

Payne, B. K., Hannay, J. W., *Implicit Bias Reflects Systemic Racism,* Trends in Cognitive Sciences, 2021, 25 (11), pp. 927–936.

Ramachandran, V. S., *Encyclopedia of Human Behavior,* Academic Press, 2012.

Ratzke, C., Denk, J., Gore, J., *Ecological Suicide in Microbes,* Nature Ecology & Evolution, 2(5), 2018, pp. 867–872.

Regalado, A., *Meet Altos Labs, Silicon Valley's Latest Wild Bet on Living Forever,* MIT Technology Review, September 4th,2021; https://www.technologyreview.com/2021/09/04/1034364/altos-labs-silicon-valleys-jeff-bezos-milner-bet-living-forever/

Rodríguez-Ferreiro, J., Barberia, I., *The Moral Foundations of Illusory Correlation,* Plos One, 12 (10), 2017, p.e0185758.

Ross, E., *Doomsday Clock Ticks Closer to Apocalypse,* Nature, 2017, 26.

Rossmo, D. K., *Cognitive Biases: Perception, Intuition, and Tunnel Vision*, in: Criminal Investigative Failures, Routledge, 2008, pp. 33–46.

Russell, S., *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin, 2019.

Schiffling, S., Valantasis Kanellos, N., *Five Essential Commodities That Will Be Hit by War in Ukraine,* The Conversation 2022.

Schilbach, F., Schofield, H. Mullainathan, S., *The Psychological Lives of the Poor,* American Economic Review, 106 (5), 2016, pp. 435–440.

Şekerli, E.B., Akçetin, E., *Diversification Strategy in Internet Industry: Case of Google Inc*., Afyon Kocatepe Üniversitesi Sosyal Bilimler Dergisi, 20 (3), 2018, pp. 271–289.

Simons, D. J., Chabris, C.F., *Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events*, Perception, 28 (9), 1999, pp. 1059–1074.

Sowels, N., *A Brief Introduction to Complexity Theory in Managing Public Services*, Revue Française de Civilisation Britannique. French Journal of British Studies, 26 (XXVI-2), 2021.

Stanchi, K. M., *The Power of Priming in Legal Advocacy: Using the Science of First Impressions to Persuade the Reader*, Or. L. Rev., 2010, 89.

Tetlock, P. E., *Expert Political Judgment,* in: Expert Political Judgment, Princeton University Press, 2017.

Thaler, R. H. Sunstein, C. R., *Nudge*, Yale University Press, 2021.

Tokita, C. K., Guess, A. M. Tarnita, C. E., *Polarized Information Ecosystems Can Reorganize Social Networks Via Information Cascades,* Proceedings of the National Academy of Sciences, 2021, 118 (50), p.e2102147118.

Tversky, A., Kahneman, D., *The Framing of Decisions and the Psychology of Choice,* in: Behavioral Decision Making, Springer, Boston, MA 1985, pp. 25–41.

Twenge, J.M., Joiner, T. E., Rogers, M. L. Martin, G. N., *Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates among Us Adolescents after 2010 And Links to Increased New Media Screen Time*, Clinical Psychological Science, 2018, 6 (1), pp. 3–17.

U.S. International Trade Commission, *Harmonized Tariff Schedule and General Notes*; accessed July 2nd, 2022; https://hts.usitc.gov/current

Urbina, F., Lentzos, F., Invernizzi, C., Ekins, S., Dual *Use of Artificial-Intelligence-Powered Drug Discovery,* Nature Machine Intelligence, 2022, 4 (3), pp.189–191.

US Bureau of Reclamation, *Colorado River Basin Natural Flow and Salt Data*; accessed July 2nd, 2022; https://www.usbr.gov/lc/region/g4000/NaturalFlow/provisional.html

Veldwijk, J., Essers, B. A., Lambooij, M. S., Dirksen, C. D., Smit, H. A., De Wit, G.A., *Survival or Mortality: Does Risk Attribute Framing Influence Decision-Making Behavior in a Discrete Choice Experiment?*, Value in Health, 2016, 19 (2), pp. 202–209.

Wachowski, A., Wachowski, L., Reeves, K., Fishburne, L., Moss, C. A., Weaving, H., Foster, G., Pantoliano, J. Staenberg, Z., *Matrix*, Warner Home Video, Burbank, CA 1999.

White, J. M., Lidskog, R., *Ignorance and the regulation of artificial intelligence*, Journal of Risk Research, 2022. 25 (4), pp. 488–500.

Whitten-Woodring, J., Kleinberg, M.S., Thawnghmung, A., Thitsar, M. T., *Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar*, The International Journal of Press/Politics, 2020, 25 (3), pp. 407–425.

Wiesmann, U. M., Hurni, H. (eds.), *Research for Sustainable Development: Foundations, Experiences, and Perspectives*, University of Bern, Bern 2011.

ABOUT THE AUTHOR — Researcher & COO at AGI Laboratory, Seattle, WA, USA.

Email: Kyrtin@ArtificialGeneralIntelligenceInc.com