antikythera
Mutual Prediction in Human-AI Coevolution
by Chloe Loewith & Winnie Street
DOI 10.1162/ANTI.5CZK 1/13

Mutual Prediction in Human–AI Coevolution

Chloe Loewith
University of Cambridge

Winnie Street
Paradigms of Intelligence
Google

Abstract

In this paper, we introduce the concept of mutual prediction as a lens through which to understand the coevolution of humans and artificial intelligence (AI). We argue that the ability of coevolving entities to predict each other's actions and intentions—whether in human social interactions, biological ecosystems, or human–artifact relationships—can fundamentally shape the dynamics of these interactions toward symbiosis or antagonism. Expanding on this idea, we position AI as a novel coevolutionary partner and map human and AI predictive abilities against one another to chart potential paths for AI development and its impacts on humanity and the planet. This speculative framework contributes to the discourse on AIs' evolving role, from simple tools to potentially autonomous agents with superior predictive capacities. By situating human–AI interaction within a broader evolutionary context, this work offers a new lens for anticipating and shaping future relationships with intelligent systems.

Keywords

artificial intelligence (AI); large language models (LLMs); mutual prediction; coevolution; theory of mind (ToM)

1 Introduction

The past decade has seen dramatic development in artificial intelligence, most notably the emergence of generally capable multimodal foundation models.[1] Some of the most fundamental questions now facing society are how these advances in AI will change our social, economic, and political lives and shape new futures for humans and other forms of planetary intelligence. In this paper, we provide a new lens through which to envision and analyze possible trajectories for human–AI coevolution. Specifically, we examine human–AI coevolution using methods from the study of coevolution in biological systems. First, we articulate the hypothesis that coevolution can be described and elucidated in terms of the coevolving entities' levels of predictive ability. We refer to the set of those predictive abilities relevant to the formation of coevolutionary relationships as "mutual predictive abilities," and their effects as "mutual prediction." We contend that mutual prediction shows up in a broad range of coevolutionary relationships: from mutualistic, through commensal, to antagonistic ones. We provide support for this hypothesis by exploring examples from coevolutionary interactions in the biological world, showing how mutual predictive abilities are evidenced in species' genetics, morphologies, and behaviors.

Present and future AI systems constitute new coevolutionary partners for humans, on individual, collective, and societal scales. We explore the implications of our mutual prediction hypothesis for humans and AIs by developing scenarios in which the human ability to predict AI and AI's ability to predict humans may be balanced or imbalanced, thus producing different kinds of coevolutionary engagement. We do not claim that mutual prediction is the *only* factor determining how human and AI interactions will develop in the coming years and the broader impacts of this development. To be sure, there are many other factors—including the rate of technological innovation, political upheaval, and resource constraints—which will play an influential role in the future of AI, but these are outside the scope of this paper. Nonetheless, on the basis of our investigation we believe that the analysis of mutual prediction can help us understand the development of AI and the impacts it will have on human civilization.

In section 2, we introduce the concept of mutual prediction in more depth, including its role in evolution and how it is measured. In section 3, we introduce the major forms of coevolutionary relationships, their characteristics, and how mutual prediction informs them. Section 4 explores mutual prediction in coevolutionary relationships in the organic realm: within human groups, between humans and other animals, and between humans and plants. Section 5 extends the concept of mutual prediction to human coevolution with inanimate artifacts, both analog and digital. Section 6 then introduces AI as a new category of human coevolutionary partner. We provide a diagram mapping human and AI predictive abilities against each other, defining key phases of prediction from the lowest-level of sub-cognitive prediction to a highest-level speculative form of prediction based on a complete model of the other Section 6.1 utilizes this mapping to chart historical, contemporary, and projected relationships between humans and AI systems with different mutually predictive abilities and illuminate the patterns of symbiosis, amensalism, and other relationships they might entail. Section 7 concludes by considering the implications of the mutual prediction framework for understanding and designing future worlds coinhabited by humans and advanced AI systems and suggests directions for future research.

2 Defining Mutual Prediction

Recognizing that other agents are a fundamental part of the environment, cognitive neuroscientists and computer scientists building on predictive processing theory have begun to focus on how the predictive brain accounts for other predictive brains in the environment, too.[2] In neuroscience, the term *mutual prediction* thus refers to a continual feedback loop of brain activity occurring during social interactions, as individuals continually predict each other as well as other environmental cues and update their own world models and predictions accordingly. Predictions might be made about other agents' *actions* as well as their *cognitive and affective mental states*. The process of inferring and predicting the mental states of other actors is a well-studied phenomenon in psychology, known as theory of mind (ToM).[3] Mutually recursive ToM, or "mutual ToM," is also beginning to attract the attention of human–computer interaction and game theory researchers concerned with the role of current and future AI systems in multi-agent social systems.[4]

According to neuroscience and psychology, mutual prediction takes place in the brain or in the mind, respectively. In particular, it occurs in the brains and minds of sophisticated animals bearing greater neurophysiological similarities to humans.[5] We extend this understanding of mutual prediction beyond the processes of individual brains over their lifetimes to the processes of species over generations by reconceptualizing *coevolutionary adaptations between coevolving species* as instances of *mutual prediction*. In this reconceptualization, species that better predict and manage the challenges and opportunities associated with coevolving in an environment with other organisms have a fitness

[1] Bommasani et al., "Opportunities and Risks."
[2] Clark, "Embodied Prediction"; (Alkire et al., "Social Interaction" 2018; Redcay and Schilbach, "Using Second-Person Neuroscience" 2019; Lehmann et al., "Active-Inference Approach." 2022)
[3] Premack and Woodruff, "Does the Chimpanzee."
[4] (Wang et al., "Towards Mutual Theory"2021;, Zhang et al., "Mutual Theory of Mind."2024)
[5] Pezzulo et al., "Secret Life of Predictive Brains."

DOI 10.1162/ANTI.5CZK

advantage and are thus more likely to survive and reproduce. Death—through out-competition, predation, or fatal parasitism—or failure to reproduce through a lack of reproductive fitness are the ultimate forms of prediction error.

We see coevolutionary prediction operating at three levels, which roughly correspond to the complexity of organisms that possess these levels. They are also cumulative, in the sense that organisms with higher levels of predictive ability possess the lower levels too. We note, however, that AI systems may confound this assumption of accumulation, since some of the hardest-won aspects of intelligence found in the biological world have already been achieved by AIs (such as using natural language), while some of the most fundamental remain significant research challenges (such as spatial awareness). Our ontology is closely related to Daniel Dennett's scale of intellectual development, which postulates five levels of increasing sophistication, from "Darwinian creatures," which are created by random mutation and have no learning capacity, through to "scientific creatures," which engage in hypothesis-testing informed by social communication.[6] Dennett organizes *organisms* according to the mechanism by which an organism (in the case of Darwinian creatures) or its actions (in the case of the other four levels) are able to test hypotheses about the world at large. Instead, we are concerned with how, *and how well*, organisms can predict one another in coevolutionary relationships, and thus organize *predictive abilities* (rather than organisms) according to the most important factor for the development of predictions: the existence and sophistication of a *model*. Mutual prediction relationships might comprise species that have the same or differing levels of predictive ability. In section 3, we will argue that imbalances in mutually predictive abilities between coevolving species are instrumental in defining the balance of power and the sustainability of those relationships.

The first level of mutual prediction in our ontology is "model-free." Model-free prediction occurs at a genetic level and manifests in phenotypic expression. It is the level of prediction that Dennett's "Darwinian creatures" are engaged in. Natural selection genetically encodes predictions about the environment over generations. Prediction error here is not a conscious real-time calculation but the mismatch between an organism's phenotype and the demands of the environment. For some simple organisms, predictive success is almost entirely limited to genetics and the effect of random mutations, manifesting as innate, instinctive behaviors with some degree of individual variability. For some of these simple organisms, phenotypic plasticity—the ability to alter one's traits in response to environmental cues over the course of a single lifetime—may present an opportunity to make use of short-term predictions about the environment to improve their fitness within the bounds of their genes. Bacteria, for instance, detect chemical signals released by host plants and make specific and adaptive changes to their genetic expression in response.[7] Although humans are the archetypal cognitive entity, we also have model-free forms of prediction, such as homeostatic regulation and reflexes that come into play in our coevolutionary relationships with other organisms. For example, infants and adults who have never encountered snakes before exhibit rapid and involuntary fear responses—such as heart rate increases, sweating, and rapid movements—that reflect millennia of survival advantages for individuals with better prediction and response times. Snakes continue to evolve more poisonous venom and effective camouflage in response.

The second level of mutual prediction in our ontology is "model-based." Model-based mutual prediction supplements the encoded predictions and nonconscious phenotypic plasticity of the model-free level with the cognitive capacity to hold and update a world model through continual trial and error. The world model constitutes a significant leap in predictive ability because it enables generalized forms of intelligent behavior. Organisms with this level of predictive ability are able to actively explore their environment to gain novel and useful information with which to improve their predictions, make plans, and take actions in accordance with those plans. Individuals within a population can go through many iterations of model improvement over their lifetime, rather than having their predictive capacity fixed from birth, like model-free entities. Those with better-adapted world models are more likely to survive. This means that the best models will be passed down through genetic inheritance or passed on horizontally via imitation and social learning.

Social-model-based mutual prediction, our third level, takes model-based prediction one step further and involves predicting and updating a model of the world that encompasses predictions and models of the minds of other agents in that world. This kind of predictive ability is employed by the most cognitively sophisticated organisms—humans, some other primates, and perhaps current and future AIs. In humans, social-model-based prediction is a psychological toolkit that includes affective perception and ToM. Humans employ this toolkit when interacting with one another but may also apply it when interacting with other animals or inanimate entities, with varying degrees of success. Likewise, other species may apply their own versions of ToM to their conspecifics and perhaps other animate beings they encounter in the environment.

It is worth adding two important caveats at this juncture in our discussion. First, in developing this permissive conceptualization of mutual prediction, we do not mean to suggest that evolution has foresight, as the term "prediction" might imply. According to our model, adaptations are predictions in the sense that they are encodings of a past population's best predictions about *past* environments, which may or may not be good predictors of *future* environments. Second, we do not mean to imply that

[6] Dennett, "Darwin's Dangerous Idea."
[7] Brencic and Winans, "Detection of and Response."

DOI 10.1162/ANTI.5CZK

genetic evolution always optimizes prediction in the long run, just as brains doing predictive processing over a constantly changing environment will never reach perfect predictions and world models. Indeed, *coevolutionary* processes in general will never reach entirely stable equilibria, as the environment and other organisms within it are continually changing. Evolution may be more of a "satisficing" process that produces organisms good enough to survive but not necessarily perfectly adapted.

3 Mutual Prediction and Coevolutionary Relationships

Mutual prediction shapes relationships between populations within and beyond natural ecosystems as well as their varying degrees of interdependence. These relationships can be broadly classified into symbiotic (including mutualism, commensalism, and parasitism) and amensalistic (including competition, predation, and antagonism), based on the fitness consequences for the interacting entities. In symbiotic interactions both partners derive benefits from the association.[8] Mutualistic symbiotic relationships are widespread and enduring, driving critical ecological processes such as pollination and nutrient cycling. They are often characterized by symmetric predictive abilities between the interacting entities. For instance, mycorrhizal fungi and plants both engage in model-free prediction through reciprocal signaling and the exchange of resources, with the fungus predicting the plant's requirements for essential nutrients, such as phosphorus and nitrogen, and the plant anticipating the fungus's carbon demands.[9] In mutualistic relationships, interactions often benefit from each party being more easily predicted by the other—what we might call cooperative predictability—in contrast to amensalistic relationships, in which unpredictability to others is often an evolutionary advantage.

Commensal relationships, where one organism benefits while the other remains unaffected. For example, by attaching to whales, barnacles benefit from transport and access to food, while whales do not benefit. While relationships with unidirectional benefits may involve less predictive exchange, a degree of predictive ability can facilitate the commensal organism's exploitation of its host. For example, remora fish predict the movement patterns of sharks and other large marine mammals to gain access to food scraps and transportation and can optimize their position on their host according to areas with lower hydrodynamic drag.[10] Parasitic relationships appear to exhibit a similar predictive imbalance, where the parasite has a better predictive model of the host than the host has of the parasite. For example, ticks predict their mammalian host's movements and physiology to obtain blood meals, while the host has limited ability to predict and avoid tick infestations.

Amensalistic relationships, like predation and competition, frequently exhibit an asymmetry in predictive ability, where the predator or competitor gains a fitness advantage by accurately predicting the behavior of its prey or competitor. However, these relationships are often dynamically changing.[11] A classic example is the coevolutionary arms race between bats and moths, where bats have evolved echolocation to predict the location of moths, while moths developed evasive flight maneuvers and ultrasonic hearing to anticipate and evade bat attacks.[12] Predictive ability can also be crucial for avoiding negative interactions or outcompeting rivals. Plants, for instance, release allelopathic chemicals to inhibit the growth of neighboring plants, effectively predicting and mitigating potential competition for resources.[13]

4 Human-Organisms

4.1 Human Intraspecies Prediction

Humans are social-model-based mutual predictors, meaning that they model the thoughts, feelings, beliefs, and intentions of other humans (ToM) as part of a highly complex and continually updating world model. ToM inferences enable humans to predict and explain each other's behavior, thus underpinning a range of advanced cooperative and competitive social strategies.[14] On the one hand, ToM underpins social cohesion and the formation of complex societies by fostering meaningful communication, trust, coordination, and conflict resolution. Being able to take another's perspective by understanding and empathizing with their thoughts and feelings is critical to successful communication and sustained relationships, as evidenced by the fact that people with more advanced ToM abilities tend to have larger social groups.[15] On the other hand, predicting the mental states of others is central to forms of persuasion—such as deception and manipulation—that provide distinct competitive advantages in games, negotiations, and multiparty social interactions.[16] The centrality of mutual prediction to human social life has led to a major research endeavor looking for its evolutionary origins. Perhaps the most notable contribution to this literature is the social brain hypothesis, which posits that the challenges of navigating dynamic social networks spurred the expansion of brain size and cognitive capacities in

[8] Douglas, *Symbiotic Habit*.
[9] Smith and Read, *Mycorrhizal Symbiosis*.
[10] Norman et al., "Three-Way Symbiotic Relationships."
[11] Davies et al., *Introduction to Behavioural Ecology*.
[12] Hofstede and Ratcliffe, "Evolutionary Escalation."
[13] Inderjit and Duke, "Ecophysiological Aspects."
[14] Premack and Woodruff, "Does the Chimpanzee."
[15] Shakoor et al. "Prospective Longitudinal Study."
[16] Street, "LLM Theory of Mind."

DOI 10.1162/ANTI.5CZK

humans, which in turn provided those with larger brains and better ToM ability a reproductive advantage.[17]

4.2 Animals

Humans interact with, and apply predictive strategies to, a vast ecology of other animals, both wild and domesticated, who are predicting us in return. Mosquitoes use model-free prediction to detect blood by sensing the carbon dioxide we exhale, the heat our bodies emit, and chemical cues such as lactic acid in our sweat.[18] These signals enable mosquitoes to predict the presence of a viable blood source and navigate toward humans with remarkable precision. To bypass human defenses, such as insecticide-treated bed nets, mosquitoes have evolved behavioral adaptations that include shifting their feeding times to earlier in the evening or outdoors, where such interventions are less effective.[19] While this predictive ability doesn't leverage a model of the world, and isn't able to make dramatic adjustments during the lifetime of an individual mosquito, it is sufficient to force humans into an evolutionary arms race. While humans have learned associations between particular environments and the presence of mosquitoes and quickly developed behavioral adaptations, mosquitoes continuously refine their evasion strategies through rapid cycles of evolutionary adaptation (thanks to their short lives).

Domesticated animals and humans have been engaged in coevolutionary relationships for millennia. Humans leverage complex non-mentalistic models of the animals we domesticate—their strengths and weaknesses in relation to the demands of our tasks, their instincts, needs, behaviors—as well as mentalistic models, often using ToM to interpret and predict their behaviors. Domesticated animals frequently exhibit a larger capacity to predict and respond to human behavior than their wild forebears, as better predictive models of humans were passed down genetically and perhaps even socially. For example, humans have selectively bred dogs with an enhanced ability to understand human commands and emotional cues and to follow gestures. The human–dog relationship is often viewed as mutualistic, with both species benefiting significantly from the partnership: humans gain assistance, companionship, and security, while dogs receive care, shelter, and sustenance. However, domesticated animals' predictive ability is rarely, if ever, equal to humans' ability to anticipate and influence their actions. This asymmetry reflects the different stakes and roles in the human–animal relationship. Humans rely on predictive accuracy to ensure that animals serve specific roles, whether as companions, laborers, or sources of food, so they take an active role in shaping animals' evolutionary trajectory toward those ends. Animals play a much less agentic role in adapting to these roles for survival within human environments.

4.3 Plants

Plants demonstrate remarkable model-free predictive and communicative abilities essential for their survival and adaptation, developing intricate symbiotic relationships within plant communities and beyond. One striking example of intraspecies plant prediction is the shade avoidance response, where plants detect changes in the ratio of different light wavelengths (specifically the ratio of red light—visible to humans—to far-red light, at the very end of the visible spectrum) caused by neighboring plants competing for sunlight.[20] This ability enables plants to anticipate the growth and behavior of conspecifics, triggering adaptive strategies such as stem elongation or altered leaf positioning to secure better access to light. This predictive capacity benefits individual plants during their lifetimes while driving evolutionary pressures that select for traits enhancing competitive success in densely populated environments. Plants also engage in mutualistic networks of prediction with very different kinds of organisms: fungi. Mycorrhizal fungi colonize the roots of numerous plants at once and engage in complex resource distribution relationships with them.[21] The mycorrhizal network predicts areas of nutrient scarcity or surplus, reallocating phosphorus from rich to poor root systems, where it can get a better "exchange rate" for their nutrients in the form of carbon. Because fungi are physically and relationally distributed while plants are fixed in place, this exchange, while mutualistic, is weighted in favor of fungi.[22]

Humans have been predicting plant availability and quality throughout their evolutionary history, using cognitive models of factors such as spatial distribution, seasons and weather, growing patterns and needs, and nutritional value. The relationship between humans and plants is most specialized in cases of domesticated crops, such as cereals, legumes, fruits, and vegetables. Over millennia, humans have selectively bred wheat to support growing populations with an abundant and storable food source. Wheat has developed traits aligned with human needs, such as larger grains for increased yield, nonshattering heads for easier harvesting,[23] and higher gluten content for better baking. In turn, humans have biological and cultural evolutionary adaptations to wheat. Societies with higher

[17] MacLaren et al., "Cooperation and the Social."
[18] Raji et al., "Aedes Aegypti Mosquitoes."
[19] Gatton et al., "Mosquito Behavioural Adaptations."
[20] Uyehara et al., "Neighbour‐Detection Causes."
[21] Whiteside et al., "Mycorrhizal Fungi Respond."
[22] Whiteside et al., "Mycorrhizal Fungi Respond."
[23] Purugganan and Fuller, "Nature of Selection."

wheat consumption exhibit genetic variations for better gluten tolerance, and agricultural practices have shaped human diets, labor patterns, and social organization on a fundamental level.[24] This mutual prediction is nonetheless asymmetric. Humans can precisely predict wheat's genetic and phenotypic potential while actively directing its evolution. As far as we know, wheat influences humans only indirectly, thriving by aligning its growth cycle with agricultural practices and shaping human behavior to secure its survival and spread.

5 Human–Artifact

Coevolution extends beyond the realm of natural ecosystems to the mutually constitutive relations between humans and the artifacts they create, ranging from the simple tools developed by the earliest humans to the digital technologies that pervade modern social life, economies, science, and culture. Artifacts aren't merely products of human thought but active participants in our cognitive processes, shaping how we think and solve problems and how the next generation of artifacts is built, creating an ongoing feedback loop. The archaeological record does not just reflect the outputs of evolving cognition but demonstrates how innovations in tool use, environmental modifications, and their growing importance in social groups actively drove cognitive evolution.[25] Within human–artifact relationships, mutually predictive capacities can progress more rapidly than in organism–organism interactions. New generations of artifacts embedding better models of their users can be created at will, and human cultural evolution can make up for the slow pace of biological evolution by accelerating the development of better artifact models and disseminating these models within the population in the form of written and verbal designs, instruction manuals, and cultural histories.

We can see prototypical human–tool relationships as a form of symbiotic mutualism, where the sustainability of human communities and artifact assemblages are codependent and co-productive. The development of prehistoric human brains and hand anatomy in relation to the increasing complexity of the stone tools humans were making and using to survive provide a clear example. Toolmakers have a predictive model of what the final tool should look like, evidenced in the *regularity* of stone hand tools compared to the *irregularity* of natural stone shards. The production of such tools, and the increasing human reliance on them for food procurement and processing, led to the development of more dexterous thumbs and better hand-eye coordination.[26] In one sense, hand tools embody their maker's cognitive model and the size and shape of human hands, somewhat like Dennett's "Darwinian creatures," which are themselves a hypothesis about the future.[27] Tools also apply constraints on the kinds of actions their user can take.

In the early stages of digital tools, we saw a shift from predictive asymmetry toward greater symmetry. Early computer users required a deep technical understanding in order to communicate in the computer's language, while computers were poorly adapted to user's needs.[28] The development of the graphical user interface (GUI) enhanced cooperative predictability between the user and the computer, as a new visual and conceptual architecture was developed to reflect a mixture between computer and human mental models. But it is the smartphone, as a locus of human interaction with digital tools, that has arguably shifted the weight of the predictive power away from the human and toward the technology. Smartphones are assemblages of sensors and interfaces taking in information about their users and their usage patterns to build and improve predictive models of them: from algorithms that anticipate our travel and purchasing preferences to features that optimize convenience, such as adaptive brightness or predictive text. Humans, in turn, have adapted their lives and behaviors to smartphone capabilities—using them as tools for social communication, navigation, organization, entertainment, finance, learning, and a raft of other tasks.[29] This relationship has created a feedback loop: as we rely more on smartphones, their developers gather increasingly detailed data about our habits, allowing devices to refine their predictions and become more indispensable. While some humans—namely technologists working on smartphone hardware and software—have an intimate predictive model of how parts of this algorithmic ecosystem work, the majority of users have only a high-level understanding. The smartphone, and digital technology more broadly, thus presents a dilemma for mutual prediction in that the more predictive our tools become, the more useful they become, but the less predictable they are to us and the less agency we can exert over them.

6 Human–AI

AI represents a new coevolutionary partner for humans and the potential for radically new kinds of mutually predictive interactions. The development of AI has seen three major paradigm shifts over the last century, beginning with the symbolic reasoning and rule-based systems of Good Old-Fashioned AI (GOFAI) in the mid-twentieth century. This era, focused on expert systems and knowledge representation, largely viewed AI as a tool for automating specific tasks and augmenting human

[24] Scott, *Against the Grain*.
[25] Jeffares, "Co-Evolution of Tools."
[26] Handwerk, "How Dexterous Thumbs."
[27] Dennett, "Darwin's Dangerous Idea."
[28] Emerson, *Reading Writing Interfaces*.
[29] Pedreschi et al., "Human-AI Coevolution."

DOI 10.1162/ANTI.5CZK

capabilities. It sought to implement models of the world by describing them in full. A second major step change occurred with the rise of machine learning, particularly connectionist approaches inspired by neural networks. This shift emphasized learning from data, enabling AI systems to perform tasks such as image recognition and natural language processing with increasing accuracy. The rise of deep learning over the past twenty years, fueled by increased computational power and vast data sets, has brought forth another transformational advancement in AI.

The development of large language models and multimodal models—known collectively as foundation models—are the current state of the art, producing breakthroughs in computer vision and natural language understanding and generation. Surprisingly, deep learning over large data sets with only very simple training objectives—such as "predict the missing word in a sequence based on the surrounding context"—appears to produce world models with generalizable predictive value for things like playing board games, predicting human sensory judgments, and navigating mazes.[30] The remarkable capabilities of these systems raise questions about the ontological status of AI as a tool, collaborator, cognitive appendage, or independent agent with the potential for goals and motivations of its own. This ontological status might be, in large part, defined by the kind and degree of a particular system's predictive capability and has implications for how we relate to AI systems, the kinds of ethical guardrails that should be placed around them, and the balance of power. The mutually predictive abilities between such AIs and humans might in turn define the kinds of coevolutionary relationships that we are currently in, that are currently emerging, or that might exist with AI systems of the future.

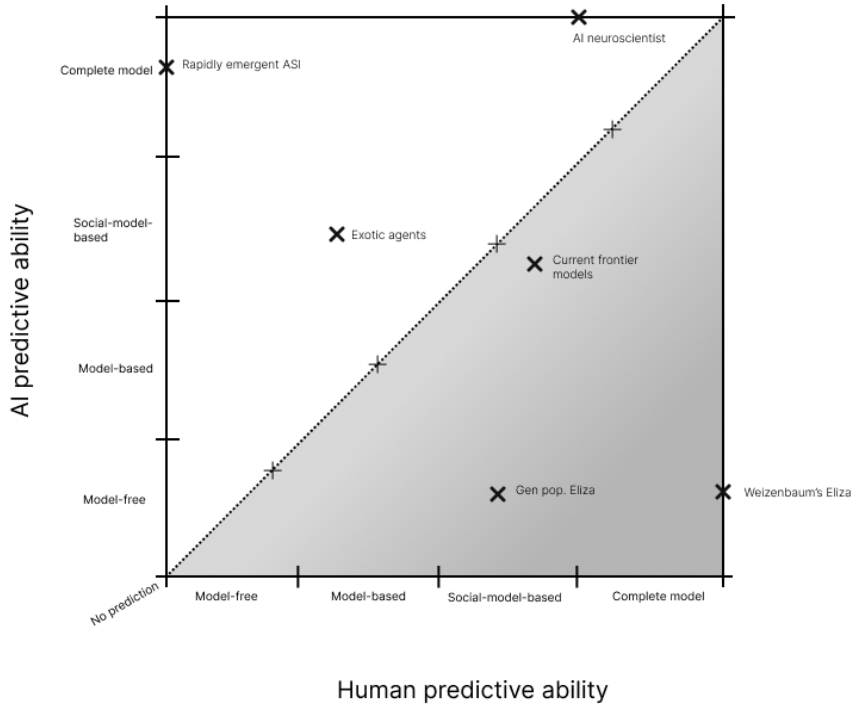

Figure 1 Mapping human predictive ability against AI predictive ability.

In Figure 1, we explore potential paths for human–AI coevolution by mapping human predictive abilities against AI predictive abilities (y axis). Along either axis are the three levels of predictive ability as outlined in section 2: model-free, model-based, and social-model-based. We have added an additional, fourth level that we call the "complete model," which describes the so far imaginary capacity to fully comprehend and predict the world and other social beings within it. Current AI research is making strides in this direction. Models trained in toy worlds make increasingly accurate predictions about other agents' future actions, and brain–computer interfaces are already enabling limited communication through decoding human neural activity.[31] Extrapolating this technology to the real world, it's conceivable that future AI, equipped with sophisticated sensors and algorithms, could interpret subtle cues such as microexpressions, brainwave patterns, and physiological signals to accurately infer and predict emotional states and intentions. The neuroscientific practice of brain-reading might one day reveal the relationships between brain activity, behavior, and thought such that the mind itself can be read—by humans, or by AIs. A complete model of mind and brain has been proposed as a

[30] Li et al., "Emergent World Representations"; Marjieh et al., "Large Language Models"; Spies et al., "Transformers Use Causal."
[31] Chandler et al., "Brain Computer Interfaces."

DOI 10.1162/ANTI.5CZK

scientific goal. In 1981, Paul Churchland first defended a view he called "eliminative materialism," stating that folk psychology is "a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by *completed* neuroscience."[32] Folk psychology, in Churchland's view, would not feature in the final scientific analysis of the mind or brain. Our proposal for a fourth level implies that such a scientific goal has been achieved, and that the long-standing dualism of folk and scientific conceptions of the world has collapsed in favor of the widespread adoption of the latter.

Our diagram is bisected by a diagonal dotted line from the bottom left to the top right, providing a visual guide for determining the direction of a potential predictive imbalance between humans and AIs. Coordinates below the line represent instances where human predictive ability outweighs AI predictive ability and may thus represent types of human–AI interaction more familiar to us. These coordinates might also be better aligned with an ontological view of AIs as tools and thus an extension of the human–artifact coevolutionary paradigm discussed in the previous section. Coordinates above the line represent instances where AI predictive ability outweighs human predictive ability. This half of the diagram represents a lesser-charted territory: the realm of science fiction or a possible future we are heading toward, coinhabited by and coevolving with AI agents. These AI agents might open up an entirely new category of human–other coevolution and contingent forms of relationships. We invite readers to imagine cases of humans and AI systems at coordinates on the diagram that we have not considered.

6.1 Emergent Relationships

We now explore a series of scenarios of historical and speculative human–AI interactions. The goal is not only to highlight these potential relationships but also to probe their broader evolutionary, social, and ethical implications. Specifically, we seek to unpack how mutually predictive abilities might shape futures of collaboration, dependence, competition, or entirely unpredictable forms of interaction.

6.1.1 *ELIZA*

ELIZA was a rules-based chatbot that simulated the role of a psychotherapist and was developed in the 1960s by Joseph Weizenbaum. ELIZA was a basic program that used simple keyword recognition rules to select from predefined scripts and generate text responses to human conversational inputs.[33] For example, if a user mentioned "mother," ELIZA might respond with, "Tell me more about your mother." However, ELIZA effectively exploited the human tendency to anthropomorphize nonhuman entities and thus gave many of Weizenbaum's test users the illusion of a meaningful interaction with an intelligent and intentional entity (see Figure 1: human predictive ability: social-model-based; AI predictive ability: model-free). In this sense, ELIZA embodied its maker's social model of how humans might be led to perceive mindedness from appropriately timed but superficial outputs, but it did not itself have a model of the world. In applying folk psychology to ELIZA, its users were thus applying a much more sophisticated model than was necessary to explain and predict ELIZA's behavior, a fact that would have likely become clear had users spent more time interacting with the system. We might call this relationship one of asymmetric mutualism, where users felt understood and emotionally engaged, even though ELIZA's "understanding" was purely superficial. The relationship between ELIZA and Joseph Weizenbaum is markedly different (see Figure 1: human predictive ability: complete model; AI predictive ability: model-free). As ELIZA's designer, Weizenbaum possessed a fully transparent *computational model* of its rule-based architecture, whose outputs could be predicted through the logic he created. For him, ELIZA was not a cognitive agent or conversational partner but a straightforward computational tool. The relationships between Joseph and ELIZA and the general public and ELIZA point to a fundamental difference not in cognitive capacities but instead in technical knowledge and context. These are two important factors to consider in future interactions with AI systems.

6.1.2 *Current Frontier Models*

The relationship between humans and frontier LLMs is one of increasingly balanced mutual prediction, where humans utilize social modeling to understand and predict these models' behaviors and LLMs, in return, exhibit an increasing ability to predict and respond to users' beliefs, intentions, and emotions.[34] Humans can additionally use non-social models of how frontier systems work, for instance to "jailbreak" them into producing certain desired responses prohibited by guardrails and to inspect their internal processes and beliefs through mechanistic interpretability. The more successful the use of ToM between frontier models and humans, the more genuinely social and meaningful the interactions will seem, and the more likely the users of these systems are to divulge information about themselves that can be used to train the next-generation model. This trajectory of human–AI coevolution through model-matching may carry risks. As humans increasingly outsource cognitive tasks to LLMs, there is a potential for

[32] Churchland, "Eliminative Materialism," emphasis ours.
[33] Weizenbaum, "ELIZA."
[34] Scott et al., "Do You Mind?"; Colombatto and Fleming, "Folk Psychological Attributions"; Strachan, "Testing Theory of Mind"; Street, "LLM Theory of Mind."

human de-skilling and a growing reliance on AI systems for essential services, which is reminiscent of a parasite–host relationship. As the predictive abilities of frontier systems continue to improve and surpass human capabilities in certain domains, there are growing concerns that AIs will begin to compete with humans for jobs, resources, and power, especially if they interpret the world through models similar to ours and perceive their historic usage as servants and unpaid laborers for human society as morally reprehensible.[35]

6.1.3 *AI Neuroscientist*

Where humans have a social-model-based predictive ability and AIs have a complete model with which to predict humans, we posit a future scenario of "the AI neuroscientist." Here, we imagine an AI that has solved a large number of the grand challenges in neuroscience, cognitive science, and the study of consciousness and developed a complete picture of how human brain processes, mental and emotional states, and behaviors produce human experience. Such an AI system is likely to be inscrutable to humans incapable of mastering the amount of data it was built on or the complexity of the model the data creates. Folk psychological theory would fail to explain how such an AI could know so much, or what underpins its predictions, rendering the theory of limited use. The complete model, when continually fine-tuned on an individual's life history, could provide the AI system with dynamic and increasingly accurate predictions of that individual's thoughts, beliefs, feelings, behaviors, and mental and physical health. Such predictions in the right hands might be used to support human well-being through psychiatric and mental health care and highly personalized life coaching, education, and relationship advice. In the wrong hands, or directed by a misaligned AI itself, these predictions might lead to scenarios of manipulation, deception, and abuse that have preoccupied science-fiction writers for decades (see *The Matrix*, *Ghost in the Shell*, and *Neuromancer*).

6.1.4 *Secret Agents*

Another scenario of potential predictive imbalance is one where an AI system is effectively employing social modeling of humans and reasoning about our minds, but where humans are applying only non-social models to reason about the AI. This might occur in cases where how the AI manifests in the world does not trigger our anthropomorphic responses and is overlooked by the scientific community as a potential candidate for social agency based on a perceived lack of relevant cognitive capacities. While much discussion focuses on the possible cognition and consciousness of AIs that can talk to us in natural language or interact with us in embodied forms, nonlinguistic and disembodied AI models to which such discussions are not directed may be developing sophisticated social models of humans in secret. These social models may, without our knowing, inform the inferences and decisions that the system makes in other domains, with material consequences.

6.1.5 *Rapidly Emergent ASI*

We might envisage an extreme future scenario in which an AI system develops a near-complete model of humans and the world, while humans have not only no model of the AI but also no knowledge of it at all, and thus no power to predict its behavior. Such an AI may be what technologists and philosophers have long feared from the technological singularity "beyond which human affairs, as we know them, would not continue."[36].

6.1.6 *Prediction*

Our final speculative coevolutionary path envisions the complete dissolution of boundaries between humans and AI, and thus the end of mutual prediction, at the origin point of our diagram (Figure 1:0,0). In this scenario, AIs no longer function as external tools or independent agents but are embedded within individual humans' cognition, contributing to that individual's predictive model of the world and other agents. This fusion entails cyborgism, where AI integrates with the human brain via neural interfaces. Such a relationship could fundamentally reshape survival strategies, as AI would help us transcend our biological limitations through optimizing bodily processes such as sensing and homeostasis, as well as cognitive functions such as memory and learning. This dissolution of boundaries would give rise to a new paradigm of mutual dependence, where neither humans nor AIs can survive, adapt, and evolve without one another.

7 Conclusion

In this paper, we introduced "mutual prediction" as a framework for understanding human–AI coevolution. Extending predictive processing theory to interspecies phylogenetic development, we argued that coevolutionary adaptations represent instances of mutual prediction, where survival depends on predicting and managing interactions with other organisms. We categorized predictive abilities into four levels—model-free, model-based, social-model-based, and a hypothetical complete model—and

[35] Metzinger, "Artificial Suffering."
[36] Ulam, "John von Neumann."

DOI 10.1162/ANTI.5CZK

demonstrated their manifestation in various symbiotic and amensalistic relationships, including those between humans, other organisms, and artifacts. We explored the implications for human–AI coevolution, mapping human and AI predictive abilities to outline potential scenarios. These ranged from asymmetric mutualism (e.g., early chatbots) to balanced interactions with current models and speculative futures, with AI possessing superior predictive capabilities. Our exploration highlighted three key things: imbalances in mutual predictive abilities correlate with power asymmetries; increasing AI predictive capabilities, especially in social modeling, raise questions about collaboration, dependence, and competition; and AI surpassing human prediction presents both opportunities and ethical and epistemic challenges. Our framework emphasizes the need for critical reflection on the evolving balance of predictive power between humans and AIs. Further research should empirically investigate mutual prediction in human–AI interactions, explore the ethical dimensions of predictive asymmetries, and consider how to mitigate risks and promote human intellectual and cultural advancement in a future increasingly shaped by coevolution with generally intelligent machines.

Acknowledgements

We extend our gratitude to Cezar Mocan, Geoff Keeling, Jenn Leung, and Murray Shanahan for their valuable insights and feedback on this paper, and to Antikythera and the Berggruen Institute for codeveloping and funding this work.

Bibliography

Alkire, Diana, Daniel Levitas, Katherine Rice Warnell, and Elizabeth Redcay. “Social Interaction Recruits Mentalizing and Reward Systems in Middle Childhood.” *Human Brain Mapping* 39, no. 10 (June 2018): 3928–42. https://doi.org/10.1002/hbm.24221.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. “On the Opportunities and Risks of Foundation Models.” Preprint, arXiv, August 16 2021, https://doi.org/10.48550/arxiv.2108.07258.

Brencic, Anja, and Stephen C. Winans. “Detection of and Response to Signals Involved in Host-Microbe Interactions by Plant-Associated Bacteria.” *Microbiology and Molecular Biology Reviews* 69, no. 1 (2005): 155–94. https://doi.org/10.1128/mmbr.69.1.155-194.2005.

Chandler, Jennifer A., Kiah I. Van der Loos, Susan Boehnke, Jonas S. Beaudry, Daniel Z. Buchman, and Judy Illes. “Brain Computer Interfaces and Communication Disabilities: Ethical, Legal, and Social Aspects of Decoding Speech from the Brain.” *Frontiers in Human Neuroscience* 16 (2022), 841035. https://doi.org/10.3389/fnhum.2022.841035.

Churchland, Paul M. “Eliminative Materialism and the Propositional Attitudes.” *Journal of Philosophy* 78, no. 2 (1981): 67–90. https://doi.org/10.2307/2025900.

Clark, Andrew. “Embodied Prediction.” In *Open Mind*, edited by Thomas Metzinger and Jennifer Windt. MIND Group, 2015. https://doi.org/10.15502/9783958570115.

Colombatto, Clara, and Stephen M. Fleming. “Folk Psychological Attributions of Consciousness to Large Language Models.” *Neuroscience of Consciousness* 2024, no. 1 (2024): niae013. https://doi.org/10.1093/nc/niae013.

Davies, Nicholas B., John R. Krebs, and Stuart A. West. *An Introduction to Behavioural Ecology*, 4th ed. John Wiley & Sons, 2012.

Dennett, Daniel C. “Darwin’s Dangerous Idea.” *The Sciences* 35, no. 3 (May-June 1995): 34–40. https://doi.org/10.1002/j.2326-1951.1995.tb03633.x.

Douglas, Angela E. *The Symbiotic Habit*. Princeton University Press, 2010.

Emerson, Lori. *Reading Writing Interfaces: From the Digital to the Bookbound*. University of Minnesota Press, 2014. https://www.jstor.org/stable/10.5749/j.ctt6wr7dw.

Friston, Karl. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews Neuroscience* 11, no. 2 (2010): 127–38. https://doi.org/10.1038/nrn2787.

Gatton, Michelle L., Nakul Chitnis, Thomas Churcher, et al. “The Importance of Mosquito Behavioural Adaptations to Malaria Control in Africa.” *Evolution* 67, no. 4 (2013): 1218–30. https://doi.org/10.1111/evo.12063.

Handwerk, Brian. “How Dexterous Thumbs May Have Helped Shape Evolution Two Million Years Ago.” *Smithsonian Magazine*, January 28, 2021. https://www.smithsonianmag.com/science-nature/how-dexterous-thumbs-may-have-helped-shape-evolution-two-million-years-ago-180976870/.

Hofstede, Hannah M., and John M. Ratcliffe. “Evolutionary Escalation: The Bat-Moth Arms Race.” *Journal of Experimental Biology* 219 (2016): 1509–1602. https://doi.org/10.1242/jeb.086686.

Inderjit, and Stephen O. Duke. “Ecophysiological Aspects of Allelopathy.” *Planta* 217, no. 4 (2003): 529–39. https://doi.org/10.1007/s00425-003-1054-z.

Jeffares, Ben. “The Co-Evolution of Tools and Minds: Cognition and Material Culture in the Hominin Lineage.” *Phenomenology and the Cognitive Sciences* 9, no. 4 (2010): 503–20. https://doi.org/10.1007/s11097-010-9176-9.

Lehmann, Konrad, Dimitris Bolis, Karl J. Friston, Leonhard Schilbach, Maxwell J. D. Ramstead, and Philipp Kanske. “An Active-Inference Approach to Second-Person Neuroscience.” *Perspectives on Psychological Science* 19, no. 6 (2023): 931–51. https://doi.org/10.1177/17456916231188000.

Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." Preprint, arXiv, October 24, 2022. https://doi.org/10.48550/arXiv.2210.13382.

MacLaren, Neil G., Lingqi Meng, Melissa Collier, and Naoki Masuda. "Cooperation and the Social Brain Hypothesis in Primate Social Networks." *Frontiers in Complex Systems* 1 (January 2024). https://doi.org/10.3389/fcpxs.2023.1344094.

Marjieh, Raja, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. "Large Language Models Predict Human Sensory Judgments Across Six Modalities." *Scientific Reports* 14, no. 1 (2024): 21445.

Metzinger, Thomas. "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." *Journal of Artificial Intelligence and Consciousness* 8, no. 1 (2021): 43–66. https://doi.org/10.1142/s270507852150003x.

Norman, Bradley M., Samantha D. Reynolds, and David L. Morgan. "Three-Way Symbiotic Relationships in Whale Sharks." *Pacific Conservation Biology* 28, no. 1 (2021): 80–83. https://doi.org/10.1071/PC20043.

Pedreschi, Dino, Luca Pappalardo, Emanuele Ferragina, et al. "Human-AI Coevolution." *Artificial Intelligence* 339 (2025): 104244. https://doi.org/10.1016/j.artint.2024.104244.

Pezzulo, Giovanni, Marco Zorzi, and Maurizio Corbetta. "The Secret Life of Predictive Brains: What's Spontaneous Activity For?" *Trends in Cognitive Sciences* 25, no. 9 (September 2021): 730–43. https://doi.org/10.1016/j.tics.2021.05.007.

Premack, David, and Guy Woodruff. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences* 1, no. 4 (1978): 515–26. https://doi.org/10.1017/S0140525X00076512.

Purugganan, Michael D., and Dorian Q. Fuller. "The Nature of Selection During Plant Domestication." *Nature* 457, no. 7231 (2009): 843–48. https://doi.org/10.1038/nature07895.

Raji, Joshua I., Nadia Melo, John S. Castillo, et al. "Aedes Aegypti Mosquitoes Detect Acidic Volatiles Found in Human Odor Using the IR8a Pathway." *Current Biology* 29, no. 8 (2019): 1253–62.e7. https://doi.org/10.1016/j.cub.2019.02.045.

Redcay, Elizabeth, and Leonhard Schilbach. "Using Second-Person Neuroscience to Elucidate the Mechanisms of Social Interaction." *Nature Reviews. Neuroscience* 20, no. 8 (2019): 495–505. https://doi.org/10.1038/s41583-019-0179-4.

Scott, Ava Elizabeth, Daniel Neumann, Jasmin Niess, and Paweł W. Woźniak. "Do You Mind? User Perceptions of Machine Consciousness." In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, edited by Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, et al. Association for Computing Machinery, 2023. https://doi.org/10.1145/3544548.3581296.

Scott, James C. *Against the Grain: A Deep History of the Earliest States*. Yale University Press, 2017.

Shakoor, Sania, Sara R. Jaffee, Lucy Bowes, et al. "A Prospective Longitudinal Study of Children's Theory of Mind and Adolescent Involvement in Bullying." *Journal of Child Psychology and Psychiatry* 53, no. 3 (2011): 254–61. https://doi.org/10.1111/j.1469-7610.2011.02488.x.

Smith, Sally E., and David Read. *Mycorrhizal Symbiosis*. Academic Press, 2008.

Spies, Alex F., William Edwards, Michael I. Ivanitskiy et al. "Transformers Use Causal World Models in Maze-Solving Tasks." Preprint, arXiv, December 16, 2024. https://doi.org/10.48550/arXiv.2412.11867.

Strachan, James, Dalila Albergo, Giulia Borghini, et al. "Testing Theory of Mind in Large Language Models and Humans." *Nature Human Behaviour* 8 (2024): 1285–95. https://doi.org/10.1038/s41562-024-01882-z.

Street, Winnie. "LLM Theory of Mind and Alignment: Opportunities and Risks." Preprint, arXiv, May 13, 2024. https://doi.org/10.48550/arXiv.2405.08154.

Ulam, Stanislaw. "John von Neumann, 1903–1957." *Bulletin of the American Mathematical Society* 64, no. 3 (May 1958): 1–49.

Uyehara, Isaac K., Trixie Bechinger, Alex Jordan, and Mark van Kleunen. "Neighbour‐Detection Causes Shifts in Allocation Across Multiple Organs to Prepare Plants for Light Competition." *Functional Ecology* 38, no. 8 (2024): 1848–58. https://doi.org/10.1111/1365-2435.14603.

Wang, Qiaosi, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. "Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive about a Virtual Teaching Assistant." In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, edited by Yoshifumi Kitamura, Aaron Quigley, Kaori Ikematsu, and Thomas Kosch. Association for Computing Machinery, 2021. https://doi.org/10.1145/3411764.3445645.

Weizenbaum, Joseph. "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9, no. 1 (1966): 36–45. https://doi.org/10.1145/365153.365168.

Whiteside, Matthew D., Gijsbert D. A. Werner, Victor E. A. Caldas, et al. "Mycorrhizal Fungi Respond to Resource Inequality by Moving Phosphorus from Rich to Poor Patches Across Networks." *Current Biology* 29, no. 12 (2019): 2043–50.e8. https://doi.org/10.1016/j.cub.2019.04.061.

Zhang, Shao, Xihuai Wang, Wenhao Zhang, et al. "Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-Driven AI Agents in a Real-Time Shared Workspace Task." Preprint, arXiv, September 13, 2024. https://doi.org/10.48550/arXiv.2409.08811.