HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Effective Generative AI: The Human-Algorithm Centaur

Faculty Research Working Paper Series

Soroush Saghafian
Harvard Kennedy School

October 2023

RWP23-030

Effective Generative AI: The Human-Algorithm Centaur

Soroush Saghafian[1], Ph.D., M.S.

Figure: Image of Chiron the Centaur in Daniel Le Clerc, Histoire de la médecine … (Amsterdam, 1723), p. 30.

1. Introduction

Developing analytics science methods that can enable combining the power of artificial and human intelligence has brought the concept of *centaurs* from myth to reality. In the Greek mythology,

[1] Associate Professor, Harvard University; Founder and Director, Public Impact Analytics Science Lab (PIAS-Lab) at Harvard; Visiting Scholar, MIT; Faculty Affiliate, Harvard Data Science Initiative; Core Faculty, Harvard Center for Health Decision Science; Faculty Affiliate, Harvard Mossavar-Rahmani Center for Business and Government; Faculty Affiliate, Belfer Center for Science & International Affairs.

centaurs are half-human and half-horse creatures (see the figure above). In modern analytics science, they refer to systems that allow superior decision-making by combining the power of both humans and trained algorithms. One of the main users in the U.S. has been the Defense Department, which has been working with tech companies to combine the power of algorithms with the capabilities of humans [1]. The concept has attracted the attention of the U.S. military, both in research programs at the Defense Advanced Research Projects Agency and the Pentagon's third-offset strategy for military advantage [2]. Robert O. Work, for example, who was the deputy secretary of defense under Presidents Trump and Barack Obama, advocated for the idea of centaur weapons systems, which would require human control, instead of pure Artificial Intelligence (AI) systems, and could combine the power of AI with the capabilities of humans [3].

The concept of centaurs is not that new, but it received spotlight attention within the analytics science domain because of its success in applications like paying free-style chess. Specifically, prominent advocates of free-style chess like Gary Kasparov repeatedly argued that human paired with algorithms can do better than just the single strongest computer program in chess [4]. As the chess legend put it:

> "Weak human plus machine plus better process was superior to a strong computer alone and, more remarkably, superior to a strong human plus machine plus inferior process." [5]

Beyond free-style chess, the centaur model is being widely used in a variety of applications of analytics science. In clinical decision-making related to rehabilitation assessment, for example, algorithms provide therapists with detailed analysis on patient's status, where the collaboration with therapist and such algorithm is shown to improve the practices of rehabilitation assessment [6].

Research in my own lab at Harvard, which we conducted in collaboration with the Mayo Clinic, showed very promising results for a centaur model that we developed to enhance decision-making and reduce readmission risks for a large number of patients who underwent transplantation. We found that combining human experts' intuition with the power of a strong machine learning algorithm through a human-algorithm centaur model can outperform both the best algorithm and the best human experts [7].

Other examples of using the centaur model to create public impact include systems for spotting anomalies and preventing cyber-attacks, improving design components in manufacturing systems, and assisting officers balance their workloads and helping them to better ensure public safety [2]. And the potential for developing and making use of centaurs is endless. Thus, it is reasonable to expect most data-driven organizations to take advantage of them in the near future. A department of human services, for example, can use algorithms to help predict which child welfare cases are likely to lead to child fatalities and raise a red flag for high-risk cases. Such cases are then reviewed by human experts and the results are shared with frontline staff, who then might choose remedies designed to lower risk and improve outcomes [8]. The algorithm then can be augmented by using human intuition related to specific cases, creating a human-algorithm centaur.

In this article, we focus on recent advancements in Generative AI, and especially in Large Language Models (LLMs). We first present a framework that allows understanding the core characteristics of centaurs. We argue that symbiotic learning and incorporation of human intuition are two main characteristics of centaurs that distinguish them from other models in Machine Learning (ML) and AI.

Using these core characteristics, we also present a few specific methods of creating centaurs. We then argue that the growth and success of LLMs are to a great extent due to the fact that they are moved from pure ML algorithms to human-algorithm centaurs. We present various evidence to demonstrate this, particularly by focusing on the advantages of the so-called "fine-tuning" approaches such as the Reinforcement Learning with Human Feedback (RLHF) method used in various LLMs (e.g., OpenAI's GPT-4, Antropic's Claude, Google's Bard, and Meta's LLaMA 2-Chat). We also discuss evidence showing that these fine-tuning approaches can turn Generative AI tools into cognitive models, capable of representing human behavior. In addition, we elaborate on three main advantages of centaurs: removing barriers with respect to *algorithm aversion*, *huma aversion*, and *casual aversion*.

We then briefly conclude by discussing two main points: (1) recent advancements in creating centaurs have moved us closer to reaching the goals that the founding fathers of AI—John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon—stated in 1955 as part of their proposed 2-month, 10-man study of AI to be held at Dartmouth; and (2) the future of AI development and use in many domains will most likely need to focus on centaurs as opposed to other traditional approaches in ML and AI.

1.1. Centaurs and Recent Advancements in Generative AI

In recent years we have seen a fast growth in the development and use of Generative AI. Notable examples include efforts by OpenIA with GPT (*Generative Pre-trained Transformer*) and DALL-E ("Dali" and "Eve," from artist Salvador Dali and the character Eve from the Pixar movie WALL-E) , Google with Bard that uses LaMDA (Language Model for Dialog Applications) and more recently PaLM (Pathways Language Model), Microsoft with Bing Chat, Stability AI with Stable Diffusion, Github (and OpenAI) with GitHub Copilot, and Anthropic with Claude.

In 2022, for example, LLMs saw a major advancement: ChatGPT—a large language model chatbot developed by OpenAI based on GPT-3.5 with the ability to provide conversational responses that can appear surprisingly human. Like other language models, the main idea behind ChatGPT is simple: to predict the next word in a sentence or phrase based on the context of previous words, using a model trained on a large number of instances. GPT-3 had about 175 billion machine learning parameters, and some estimates showed that it consumed about 936 Mwh to train—equivalent of about 30k American households power usage in a day. Some recent improvements have focused on making GPT-3 more efficient by reducing these numbers. Some other improvements over GPT-3, such as the work of some researchers at the Google's Brain team, have also enabled tasks that involve semi-reasoning. Their method termed "chain of thought prompting" enabled language models of sufficient scale (e.g., models with 100 billion parameters) to solve semi-complex reasoning problems that are not solvable with standard prompting methods [21]. In 2023, OpenAI introduced GPT-4, which is not only multimodal (e.g., accepts images), but according to the creator "exhibits human-level performance on various professional and academic benchmarks" [23].

The ability of these Generative AI models in creating high-quality content across a broad range of modalities (e.g., text, images, video, audio, and well-written computer codes) can be contributed to

many factors. But one factor that has not received enough attention is the fact that the foundation of many of these Generative AI tools has quietly moved from pure machine learning algorithms to human-algorithm centaurs. For example, as we will discuss, various fine-tuning techniques have allowed these tools to incorporate human intuition more directly, generating content that much better aligns with human preferences. Fine-tuning methods have also enabled these tools to turn into cognitive models, providing accurate representation of human behavior, which is yet another example of advantages that Generative AI as a human-algorithm centaur can offer.

In the next section, we first present a framework that allows a deeper understanding of the foundation and core characteristics of centaurs. As we will see, two important elements that distinguish centaurs from other modes of developing and using AI and ML algorithms such as Intelligence Augmentation (IA), Human-In-The-Loop (HILT), or Human Agent Collectives (HAC) are *symbiotic learning* and *incorporation of human intuition*.

After presenting the core characteristics of centaurs, we then discuss when and why one should incorporate human intuition into AI and ML algorithms. We also make use of evidence from research in my own lab, to discuss three important advantages of centaurs: *algorithm aversion*, *human aversion*, and *causation aversion*. Finally, we discuss recent research efforts in enhancing the cognitive ability of Generative AI models, and conclude by arguing that the future of AI and ML models in various domains will most likely need to focused on centaurs.

2. Foundation and Core Characteristics of Centaurs: Symbiotic Learning and Incorporation of Human Intuition

In 1960, J.C.R. Licklider published a paper entitled "Man-Computer Symbiosis" in which he popularized the idea of man-computer symbiosis "as an expected development in cooperative interaction between men and electronic computers" [35]. In his words, the aim was to "enable men and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs." In today's analytics science, we can think of symbiotic learning and incorporation of human intuition as two important characteristics of centaurs. These two characteristics distinguishes centaurs from other designs in AI and ML such as "intelligence augmentation," "human-in-the loop," or "human-agent collectives."

Intelligence Augmentation (IA) (a.k.a. "Augmented Intelligence") refers to a subset of AI and ML that focuses on AI as an "assistant." The goal is to create models that can serve humans as a support system, improving human decision-making, access to information, problem solving capabilities, and/or augmented memory, among others. But IA systems lack any element of symbiotic learning. Furthermore, human intuition might or might not be incorporated in such systems.

Human-In-The-Loop (HILT) refers to "the need for human interaction, intervention, and judgment to control or change the outcome of a process, and it is a practice that is being increasingly emphasized in machine learning, generative AI, and the like" [24]. In other words, HILT is often thought

of as a combination of supervised machine learning and active learning where humans are involved in both the training and testing stages of building an algorithm. But, again, there is (a) no element of symbiotic learning, and (b) human intuition is used only in limited ways (e.g., in how the algorithm is designed and tested).

Human Agent Collectives (HAC) is a closer concept to centaurs. HACs "are a new class of socio-technical systems in which humans and smart software, agents, engage in flexible relationships to achieve both their individual and collective goals. Sometimes the humans take the lead, sometimes the computer does, and this relationship can vary dynamically." [37, p. 80]. However, a main difference between centaurs and HACs is that HACs can be thought of as creation of agile teams where humans and agents will form short-lived teams, whereas centaurs involve permanent and symbiotic relationships [25].

Nonetheless, one can think of these different terms as related concepts, though with slightly different verbal interpretations. What is more important, however, is understanding how centaurs can be developed and employed differently than other AI and ML systems to reach their maximum potential by using symbiotic learning and by benefiting from human intuition.

Generative AI models such as many current LLMs have moved towards using fine-tuning methods in which human intuition is used as part of symbiotic learning to align outcomes with human preferences. We will learn about this way of incorporating human intuition later in this article. A simpler yet more general method of benefiting from human intuition is to first develop a non-centaur ML model that can represent and learn human intuition as a function of various context-dependent variables. The output of this ML model can be fed to a new ML model as an input along with typical data sets used for training. Benefiting from the learned human intuition, this new and augmented ML system provides outputs that can be used by human experts. This symbiotic learning can continue, creating a high performing centaur.

For example, in our work [7], we first deployed a non-centaur ML model that could represent physicians' intuition (assessed at the time of discharge) in whether or not a patient will be readmitted to the hospital within 30 days. The output of this model was then fed to another ML model along with various variables available in training data sets. An advantage of this method is that the first step—learning about human intuition via ML—is not as difficult as it might sound. Specifically, various studies have communicated that humans often think "linearly" when it comes to modeling their perception (see, e.g., [7] for more discussions). Thus, one might not need complicated ML model to learn about the association between outcomes and the underlying context variables as relates to human intuition. Surprising, however, feeding human intuition as captured in the first step to the ML model provided a superior system that outperformed both best humans and best ML algorithms [7].

This way of creating centaurs is depicted in Figure 1 (top). This approach extends traditional methods of creating centaurs in which first one decides whether the task at hand should be dedicated to a human or to a computer. The task is then routed to a human or AI accordingly (see, e.g., [38]). The extension is due to the fact that, in the approach depicted in Figure 1 (top), human intuition is one of the

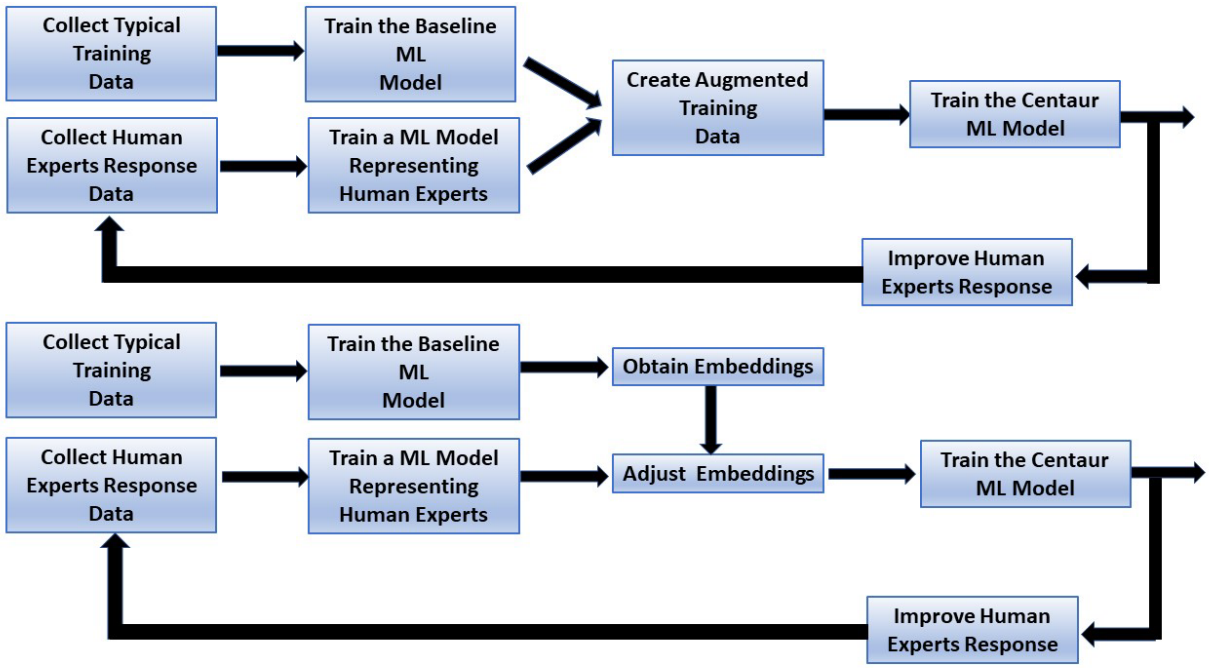

Figure 1: *Two modern methods of creating centaurs. Top: augmenting the baseline ML model by feeding outputs of an ML model that represents human experts' intuition (see, e.g., [7]). Bottom: creating a centaur by directly adjusting embeddings or other aspects of the ML model (such as adding a layer to an underlying neural network) (see, e.g., [40]).*

many inputs to the ML model, and thus, the wight assigned to human intuition is not necessarily zero or one.

Another way of creating centaurs involves adjusting embeddings or other aspects of a baseline ML model (e.g., an LLM) so it better represents human intuition. This general method of creating centaurs is depicted in Figure 1 (bottom). In [40], for example, the authors extracted embeddings for several cognitive tasks using LLAMa (Large Langue Model Metha AI) and then fine-tuned a linear layer on top of such embedding to predict human choices, creating a new class of models that they termed CENTaUR. The resulting model showed close to human behavior in two general sets of human decision-making experiments. The first set is known as *decisions from descriptions*. In this set, a complete, idealized, and abstract set of information about the values and frequencies of potential outcomes from each choice is provided to participants before choices are made [26]. An example is when participants are told there is "50% chance to win 1000; 50% chance to win nothing" ([27], p. 264). The second set is known as *decisions from experience*. In this set, participants need to form their own view of the potential outcomes from each choice via feedback provided after each selection is made [26, 40].

Fine tuning of Generative AI models such as LLAMa and GPT, which in essence allows combining human intuition with the power of a pre-train machine learning model, has shown close to (or better) than human behavior in both of these sets of decision-making experiments [39,40]. We will review some experiments that further reveal the cognitive ability of Generative AI models such as GPT in Section 3.2.

A central question, however, that might arise is this: when and why should one incorporate human intuition? Wouldn't it degrade performance? After all, aren't ML models better than human intuition in many tasks? In the next section, we will provide some answers to these questions. However, it is important to note that centaurs can work better than the best algorithms, even in tasks in which human intuition is relatively weak. In our experiments with the Mayo Clinic, for example, the ranking in performance (from best to worst) was as follows: human-algorithm centaur, algorithm, and human experts. Notably, the best human expert's performance was significantly below that of an ML model that was not benefiting from human intuition. Yet, feeding human intuition to it could create a desirable boost in performance.

2.1. When and Why Should One Use Human Intuition?

Humans often face difficult decision-making situations, and it seems that their intuition is not always helpful. When facing critical life-changing decisions such as quitting a job or ending a relationship, we tend to be happier with the outcomes later on, when a coin toss tells us to make a change than when it promotes maintaining the status quo [9].

Nonetheless, human intuition is often very powerful, especially when we want to make quick decisions. Put it differently, while intuition often misfires when we are dealing with complex problems that require careful analytics (e.g., in finding ways to reduce incidents of diabetes for organ transplanted patients [10, 11], deciding upon cell formation and layout design for a cellular manufacturing system [12], or finding most effective ways of saving lives in emergency rooms [13, 14]), it can be very useful when using data, models, and careful analytics is not an option. Malcom Gladwell's popular book "Blink: The Power of Thinking Without Thinking" provides various examples of this, including when police officers need to quickly decide whether to shoot a suspect [15]. Good intuitive decision-making also helps fire fighters when they face a burning building [16].

In addition, while relying on intuition in handling complex problems can be misleading, combining intuition with the most useful analytics approaches can often be better than just relying on analytics. To better understand this, it is useful to see how our own system of thinking works. Daniel Kahneman—a contemporary phycologists, known for his groundbreaking work on the psychology of judgment and decision-making as well as behavioral economics, who won the Nobel Prize in Economics in 2002—highlighted in his book "Thinking Fast and Slow" that our brain has two modes of thinking: System 1 and System 2. System 1 is fast and instinctive, but System 2 is slower, more deliberative, and more logical. What is perhaps more interesting is that these systems greatly *complement* each other. Our body somehow knows that we need both systems to be able make good decisions in different situations.

Similar to how System 1 and 2 complement each other, intuition and analytics can help each other as well. And this is where *human-machine centaurs* can play a vital role. We—humans—can use our intuition in many ways while developing analytics methods and taking advantage of computers to run them. For starters, intuition often allows us to develop better models, or considering the George Box's aphorism "all models are wrong, but some models are useful," more "useful" ones. Intuition also allows us to verify the results obtained from models and make sure that the assumptions made in the model are not problematic. Analytics scientists often use this simple technique as a *feedback loop*: when the results obtained from a model are not sensible and can be related to a wrong assumption in the model, they modify their model to obtain a better one. And if they are working with a cloud of models to address *the*

curse of ambiguity [17, 18], they can replace the models that might be causing preposterous results. Preposterous results could also be related to abnormalities and/or outliers in data that need to be removed before feeding them to models, and intuition is often very helpful here as well.

Realizing that intuition alone can be misleading in understanding and analyzing complex systems, and that human-machine integration are needed to harness the full power of both advanced analytics and mighty intuition, has introduced new methods of creating human-algorithm centaurs capable of symbiotic learning. We saw some of such methods in the previous section (Figure 1). In Section 4, we will elaborate more on specific methods that Generative AI models such as recent LLMs have used (Figure 2). However, prior to doing so, it is useful to first see some of the main advantages of centaurs vis-à-vis traditional ML and AI models.

3. Main Advantages of Centaurs Compared to Traditional ML and AI Models

Beyond the fact that human-algorithm centaurs can outperform both best algorithms and best human experts, there are various other reasons why developing and implementing centaurs should be on top of mind in different sectors.

First, "algorithm aversion" has been widely reported as a main hinderance in creating impactful AI in applications in which the final decision-maker is a human [31, 32, 33]. Through experiments, we have found that utilizing centaurs can reduce algorithm aversion, mainly because recommendations from centaurs are closer to human way of thinking (since centaurs incorporate human intuition) [7]. For example, using the "weight on advice measure," we found that physicians are more willing to consider recommendations that come from centaurs than those from traditional ML models [7].

Second, and related to the previous point, the recommendations coming out of centaurs represent less "human aversion" in that, not surprisingly, they better represent human behavior and preferences. While AI developers have focused heavily on developing algorithmic recommendations that are interpretable and explainable, one should note that absent incorporating human intuition, interpretability and explainability alone are not that useful: an interpretable and explainable algorithm can still provide recommendations that do not match human intuition, which can make it less likely for a human decision-maker to put weight on the advice provided by the algorithm.

Third, centaurs can help AI developers overcome what we might call "causation aversion." This refers to the fact that most of the focus of the current AI and ML algorithms has been on the association layer of the "Ladder of Causation" [41]. What decision-makers in various sectors need are algorithms that can work beyond association between various input and output variables. Prescriptive analytics, where the intention is to recommend decisions that can *cause* improvements, can benefit a lot from moving algorithms to human-algorithm centaurs. Given various ongoing research on causal ML (see, e.g., [10] and the references therein), one can expect to see more efforts in developing centaurs capable of causal and counterfactual reasoning.

3.1. Centaurs and Cognitive Ability of Generative AI

Another main advantage of centaurs is that, unlike traditional AI and ML models, they have the ability to be used as *cognitive models*. More precisely, since they benefit from human intuition and symbiotic learning, they can excel in cognitive tests. Recent research, for example, shows that GPT has been able (sometimes after additional fine-tuning based on human intuition) to perform well compared to human subjects in vignette-based tasks, and has also excelled in multi-arm bandit type of decision-making experiments that requires balancing exploitative and exploratory actions [39]. Some specific cognitive experiments performed by researchers to assess the cognitive ability of GPT include:

Base Rate Fallacy (The Cab Problem). This experiment is about judging the color of a cab that has been involved in an accident. Base rate fallacy refers to the fact that participants often fail to take into account the base rate of different cab colors that operate in a city. GPT-3 has been able to avoid this fallacy, providing accurate answers [39].

Causal and Counterfactual Reasoning (Blicket, Pills and Do Experiments). One of the experiments used in assessing the causal reasoning ability of GPT is the "Blicket" experiment. The goal is to identify whether an object turns on a machine (i.e., is a "blicket"). Two objects are introduced, and the participants are informed that the first object alone can turn on the machine. But for the second object to turn on the machine, it needs to be accommodated with the first one. GPT-3 has been able to correctly identify that the first object is a blicket, but the second one is not [39]. A more serious set of investigations includes using the so-called "pills experiments" and "do experiments". In both of these, an underlying causal relationship between different entities (e.g., pills and death) is provided and then the participant is asked some counterfactual questions. GPT-3 without fine-tuning has shown good performance in some of these experiments, but has also raised concerns about its overall causal and counterfactual reasoning. For example, authors in [39] concluded that "GPT-3 has difficulties with causal reasoning in tasks that go beyond a vignette-based characterization." Using experiments related to over hundred causal relationships from various domains such as physics, biology, zoology, cognitive science, epidemiology, and soil science, authors in [28] concluded that "algorithms based on GPT-3.5 and 4 outperform existing algorithms on a pairwise causal discovery task, counterfactual reasoning task, and actual causality. At the same time, LLMs exhibit unpredictable failure modes."

Problem Solving and Decision-Making (Multi-Arm Bandits and Prospect Theory Experiments). GPT has also shown strong performance in problem solving and decision-making tasks. As mentioned earlier, these include both *decisions from descriptions* and *decisions from experience.* The latter is the more challenging one, and often requires some fine-tuning to make sure human intuition is directly fed to the underlying model. In testing the ability of LLMs to make decisions from experience, some researchers have focused on multi-arm bandit problems where tradeoffs in exploration and exploitation play a key role. Recent work also demonstrates that LLMs might show similar to human behavior biases such as those that were documented by Kahneman and Tversky as part of their celebrated *Prospect Theory* [27,36]. For example, in one set of experiments researchers found that GPT-3 showed three of the six biases in human decision-making that Prospect Theory predicts. Specifically, GPT-3 displayed a "framing effect", a "certainty effect," and an "overweighting bias effect" but not a "reflection effect," an "isolation effect," or a "magnitude perception effect" [39].

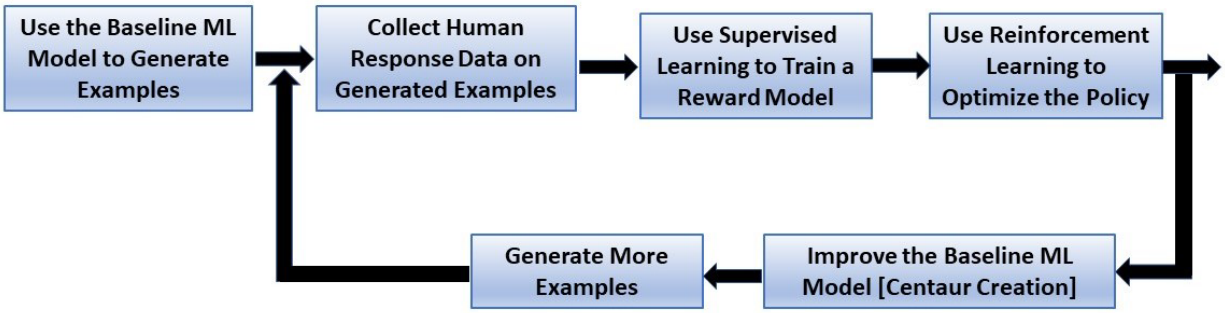

Figure 2: *Using Reinforcement Learning with Human Feedback (RLHF) to create a centaur. This method is used in various LLMs (GPT, Claude, Bard, and LLaMA) as a way of incorporating human intuition.*

Although LLMs such as GPT have shown promising results in many cognitive psychology experiments, it is important to note they are by no means close to human level of intelligence. This is partially because verbal communication and the ability to think intelligently are not the same. Notable scholars such as Noam Chomsky, who studied the mental representations and rules that describe our perceptual and cognitive skills, have argued that we should dig deeper into the organism's genetic endowment and their maturation. More broadly, Chomsky has argued against the focus of modern AI on statistical learning techniques, stating that they are unlikely to yield general principles about the nature of intelligent beings or cognition [22].

Nonetheless, the progress made in Generative AI has been substantial. A lyric in Leonard Cohen's "Anthem" reminds us that "there is a crack in everything, that's how the light gets in." In analytics science, it is our duty to see the light that gets in through a model's crack, be aware of it, and inform others about it. In Generative AI models—especially in LLMs— a specific method that has been used to create centaurs that better represent human behavior is shown in Figure 2. In the next section, we will discuss this specific method of creating centaurs.

4. Creating Centaurs via Reinforcement Learning with Human Feedback (RLHF)

Reinforcement Learning with Human Feedback (RLHF) is now a dominant method used in many LLMs such as OpenAI's GPT-4, Antropic's Claude, Google's Bard, and Meta's LLaMA 2-Chat (see, e.g., [29]). As shown in Figure 2, RLHF is in essence a specific method of creating centaurs by incorporating human intuition.

Reinforcement Learning (RL) requires assigning rewards, and one way is to ask a human to assign them. In recent years, RL has seen a growing number of applications and advancements, especially in decision-making scenarios where one needs causal reasoning, and where ambiguity prevents traditional causal inference methods (see, e.g., [10], [30], and the references therein).

The main ideas behind RL can be chased back to the work of Harvard psychologist Burrhus Frederic Skinner. Skinner published a seminal work in 1938 entitled "The Behavior of Organisms," in which he perused the idea that animal behavior can, in essence, be described by a simple set of associations between an action and what the animal receives as the subsequent reward or punishment.

In the training phase of many current LLMs, a similar idea is used: a human "labeler" assigns rewards to various outputs the model generates by ranking them from the best to the worst. This phase

is shown in Figure 2 as "collecting human response data on generated examples," which in turn allows using supervised learning to train a reward model. With the reward model in hand, RL is used to optimize the underlying policy of the original ML model (e.g., an LLM such as GPT, Claude, Bard, and LLaMA). Symbiotic learning is achieved by asking the centaur to create more examples, which are again communicated to the human labeler who will, once again, inform the centaur about human preferences.

RLHF can also be used to improve the cognitive ability of centaurs in various tasks that we discussed in Section 3.1. Nonetheless, both using RLHF and other methods of creating centaurs (see Figure 1) are subject to various limitations, and further research is needed to address them.

5. Conclusion: The Future of AI In Various Applications is Centaur

In this article, we presented human-algorithm centaurs as entities that allow symbiotic learning and incorporation of human intuition. We discussed three examples of specific methods that can allow creating centaurs (Figures 1 and 2). These include (1) augmenting a baseline ML model by feeding outputs of a different ML model that represents human experts' perceptions, (2) directly adjusting embeddings or other features of a learned baseline ML model (e.g., an LLM) using human experts' outputs, and (3) making use of RLHF.

We argued that centaurs allow removing barriers such as algorithm aversion, human aversion, and causal aversion. We also discussed that, in contrast to traditional AI and ML models, centaurs can do well in tasks that require cognitive ability. This is evidenced from cognitive psychology experiments that involve not only decisions from descriptions, but also decisions from experience.

In closing, we note that while AI is still far from reaching "human level intelligence," the advancements in Generative AI have shown promising results for making use of centaurs in various applications. Centaurs might indeed be at the center of future AI developments. They might also be our main hope to get closer to reaching the goals that were stated by the founding fathers of AI. To realize this, it is useful to revisit what the founding fathers (John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon) stated in 1955 as part of their proposed 2-month, 10-man study of AI (to be held at Dartmouth):

> "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" [34].

Making "machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" no longer seems an unrealistic goal, given what centaurs have been able to achieve to this date. Future research can make centaurs even more powerful, pushing us further to reach the goals of AI research as envisioned by the founding fathers of AI.

References

[1] New York Times. Pentagon Turns to Silicon Valley for Edge in Artificial Intelligence. May 2016.

[2] PARC, 2023. Half-Human, Half-Computer? Meet the Modern Centaur. https://www.parc.com/blog/half-human-half-computer-meet-the-modern-centaur/

[3] New York Times. A Case for Cooperation Between Machines and Humans. May 2016.

[4] Garry Kasparov on AI, Chess, and the Future of Creativity. Mercatus Center blog at Medium/Conversations with Tyler. 10 May 2017.

[5] Kasparov G (2010). The chess master and the computer. *The New York Review of Books* 57(2):16–19.

[6] Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Bermúdez i Badia, S. (2021). A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-14.

[7] Orfanoudaki, A., Saghafian, S., Song, Karen, Cook, C.B. and H.A. Chakkera (2022). Algorithm, Human, or the Centaur: How to Enhance Clinical Care? *Working Paper, Harvard University.*

[8] Deloitte (2020). How New Human-Machine Collaborations Could Make Government Organizations More Efficient. *Harvard Business Review.*

[9] Levitt, S. D. (2021). Heads or tails: The impact of a coin toss on major life decisions and subsequent happiness. *The Review of Economic Studies*, 88(1), 378-405.

[10] Saghafian, S. (2023). Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach. *Management Science (forthcoming)*

[11] Munshi, V. N., Saghafian, S., Cook, C. B., Werner, K. T., Chakkera, H. A. (2020). Comparison of post-transplantation diabetes mellitus incidence and risk factors between kidney and liver transplantation patients. *PloS One*, 15(1), e0226873.

[12] Saghafian, S., Jokar, M. R. A. (2009). Integrative Cell Formation and Layout Design in Cellular Manufacturing Systems. *Journal of Industrial and Systems Engineering*, 3(2), 97-115.

[13] Saghafian, S., Kilinc, D., Traub S. J. (2022). Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue?", *Cambridge Handbook on Productivity, Efficiency and Effectiveness in Healthcare* (forthcoming).

[14] Traub, S. J., Bartley, A. C., Smith, V. D., Didehban, R., Lipinski, C. A., Saghafian, S. (2016). Physician in triage versus rotational patient assignment. *The Journal of Emergency Medicine*, 50(5), 784-790.

[15] Gladwell, M. (2006). Blink: The Power of Thinking Without Thinking. *Penguin Books,* London.

[16] Klein, G. (1999). Sources of Power: How People Make Decisions. MIT Press, Cambridge, MA.

[17] Saghafian, S. (2018). Ambiguous partially observable Markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178, 1-35.

[18] Saghafian, S., Tomlin, B. (2016). The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, 64(1), 167-185.

[19] How language-generating AIs could transfer science. Interview by R. Van Noorden, *Nature*, 605, 5 May 2022, p21.

[20] Heaven, W.D. (2021). Why GPT-3 is the best and worst of AI right now. *MIT Technology Review*, Feb. 24.

[21] Wei, J. and Zhou, D. (2022). Language models perform reasoning via Chain of Thought. *Goggle AI Blog.*

[22] Katz, Y. (2012). Noam Chomsky on Where Artificial Intelligence Went Wrong. *The Atlantic*, Nov. 1.

[23] OpenAI (2023). GPT-4. https://openai.com/research/gpt-4

[24] Meng, X.-L. (2023). Data Science and Engineering with Human in the Loop, Behind the Loop, and Above the Loop. *Harvard Data Science Review*, 5(2).

[25] Muller, E. (2022). How AI-Human Symbiotes May Reinvent Innovation and What the New Centaurs Will Mean for Cities. *Technology and Investment*, 13, 1-19.

[26] Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2018). Task complexity moderates the influence of descriptions in decisions from experience. *Cognition*, 170, 209-227.

[27] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–292.

[28] Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050.*

[29] Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P. and Wang, T. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217.*

[30] Saghafian, S. & S.A. Murphy (2021). Innovative Healthcare Delivery: The Scientific and Regulatory Challenges in Designing mHealth Interventions" *National Academy of Medicine (NAM) Perspectives*, Commentary.

[31] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114.

[32] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155-1170.

[33] Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220-239.

[34] McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf

[35] Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, (1), 4-11.

[36] D. Kahneman, & A. Tversky. (1972). Subjective probability: A judgment of representativeness. *Cognit. Psychol.* 3, 430–454 (1972).

[37] Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., & Rogers, A. (2014). Human-agent collectives. *Communications of the ACM*, 57(12), 80-88.

[38] Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of

the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

[39] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.

[40] Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917.*

[41] Pearl, J. & Mackenzie, D. (2020). *The Book of Why: The new Science of Cause and Effect*, Basic Books.