# Foreword to *Behind Deep Blue*
# *(*Feng-Hsiung Hsu, Princeton University Press, 25th anniversary edition, 2022*.)*[1]

**Jon Kleinberg**

There are many metaphors for artificial intelligence, beginning with the term "artificial intelligence" itself. And among these metaphors, one of the most evocative has always been the analogy with alien intelligence -- that to truly interact with powerful artificial intelligence is to make contact with something that feels otherworldly.

Science fiction has provided us with mental images for what to expect in such an encounter. The alien spacecraft sits in front of us, opaque, impassive, waiting. Lacking any shared experience with it, our communication has to be structured and stylized; we invent rules for conversation. We flash a series of lights at it: 2, 3, 5. Lights blink back in return: 7, 11, 13. The human scientists offstage in the control room jump back from their terminals, exclaiming. It knows about the prime numbers.

Growing up with these stories, I never appreciated how closely they'd match the experience of watching the world chess champion Garry Kasparov play against Deep Blue in 1996 and 1997. Two powerful intelligences, profoundly incomprehensible to each other, sit down and communicate through the cryptic, stylized language of chess moves. Kasparov puts a move on the board, and after a few minutes and a few billion computations, a move comes back in reply. The two entities are conversing with each other, each seeking to understand the other. We do not fathom, and in some very real sense cannot fathom, exactly how Deep Blue has arrived at its choice of moves. But we do know that however it is managing to play chess, it is utterly different from how its human opponent across the table is playing chess. And in this way, it is hard to escape the feeling that an alien presence has stealthily entered the playing venue -- a feeling reinforced by regular exclamations from commentators Maurice Ashley and Yasser Seirawan offstage in the control room: it knows about positional pawn sacrifices; it knows how to open an attack on a second flank; it knows the importance of blocking its opponent's counterplay. These are all human concepts that we thought we had created to talk about chess strategy, and we were seeing them emerge from the computations that went into Deep Blue's play.

Part of the subtlety to the story of Deep Blue is the way in which this metaphor of first contact with an alien intelligence is partly apt and partly an illusion. It is partly apt because the process by which Deep Blue arrived at its moves was so different from the human experience of thinking about moves in a chess game. But it is also partly an illusion, because it omits the role of the human team that built Deep Blue, and the human fields of study that developed the algorithms on which computer chess is based. We cannot give a full explanation for how the machine chose any specific move, but we know a lot about the mechanisms by which it searches, and the ideas and heuristics that drive its decisions; we created these things.

[1] This essay appears as the foreword to the re-released 25th anniversary edition of the book *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*, by Feng-Hsiung Hsu (Princeton University Press, 2022).

This tension between the human and alien dimensions of powerful AI plays a key role in Feng-Hsiung Hsu's narrative of the multi-year journey that he and his colleagues followed to create Deep Blue. And in the 25 years since Kasparov sat across the chessboard from the machine, we have seen this tension manifest itself across many of the domains where AI is adopted -- in online applications like search, recommendation, and image recognition; and in high-stakes societal contexts like medicine, education, hiring, lending, the political process, and the legal system. While these manifestations are diverse, it is useful to highlight three recurring themes in particular that we saw foreshadowed by the development of Deep Blue: the qualitative changes produced by quantitative increases in scale; the challenge of explaining the decisions made by AI applications; and the continuing role of human participants in the larger AI ecosystem.

The first of these themes is well-summarized in the mantra uttered by Kasparov in the first Deep Blue match, that "quantity becomes quality". Of course, it has never been a matter of pure quantity, since the qualitative insights that humans put into the algorithms and the architectures of AI are crucial for their success. But there remains something conceptually elusive about the way we can start with an algorithm that we understand at a step-by-step level, and when we string together billions of these steps, we see emergent behaviors that are very hard to explain in terms of the steps themselves. As Hsu writes in telling the story, "Sometimes, searching two-hundred million positions per second produces unusual new solutions." This effect has become even more potent in recent years, both in chess, where current chess engines like DeepMind's AlphaZero have achieved enormous levels of strength by training against themselves starting directly from the rules of the game, and in many other areas where AI is deployed.

The fact that quantity can become quality is a principle we know from the physical world as well. Think of a propeller. When it's turned off, you can move the blunt metal blade by hand and see its rotations. But from the hand-simulation, it's hard to understand how dramatically increasing this same propeller's speed can allow it to do things that seem magical, lifting heavy aircraft off the ground. Or how dangerous it becomes when it's doing that. Or the fact that at these speeds it will be essentially invisible to the human eye.

We suffer similar failures of imagination when we try to hand-simulate the algorithms that power modern AI applications, and to reason about what they might do. As they've grown in strength, they have come to understand our speech, paraphrase our writing, explain what they see in our photos, autonomously guide our vehicles, diagnose our medical conditions, and discover new mathematical facts and scientific phenomena. They also make damaging errors in situations that we would have imagined as relatively free of risk. We struggle to anticipate both their successes and their failures, and how they vary as we make small changes to the methods they use or the data they're trained on. And all this is happening with systems that are often operating largely in the background, behind the unassuming white screen of a search results page or the sleek interface of a phone's speech recognition system. In this way, like the propeller, AI has also become magical, dangerous, and invisible.

The difficulty in understanding how different behaviors emerge from a complex AI system underpins a second central theme, the challenge of interpretability: to provide explanations of a system's behavior that are comprehensible to humans. Often this is a functionality urgently needed by the engineers who build these applications, faced with the challenge of finding and correcting bugs in a system as complex as Deep Blue. But sometimes the problem of interpretability reaches the end user of the system as well -- for example, when we want to know why an algorithmic system gave us a particular medical diagnosis, or a particular decision on a loan application.

This breakdown of interpretability plays a key role in one of the central episodes of the matches between Kasparov and Deep Blue, a hurricane of controversy and debate around a single move made by the computer: its decision in the second game of the second match to play its bishop to the square e4. Computers had traditionally favored moves that -- all else being equal -- won their opponents' pieces and pawns, but on this move Deep Blue spurned such temptations and instead smothered Kasparov's hopes of a potential counterattack with a quiet but deadly defensive move. The move violated the ubiquitous mental model for how chess programs at the time were supposed to work, and it threw the planning of Kasparov's team off balance. So how did Deep Blue arrive at the decision to make this move? At some fundamental level, as Hsu concludes in his story of the match, there is no easy answer to this question. In the few minutes while it was considering this move, Deep Blue performed billions of computations. Even with a complete transcript of its reasoning, it would take more than a human lifetime to read through it; and at the end, there would likely still be no short explanation for why the computation reached the conclusion it did. In some situations, even with all the knowledge and motivation that experts can bring to bear, a computer's decision may be inherently incomprehensible to human beings.

So a lack of interpretability can sometimes be inevitable, but it is not all-or-nothing; we can work to create structures that increase the extent of explanation, that build dependability and trust into the system. The story of Deep Blue is the story of this process as well, and it is still ongoing throughout the places where AI is used. Because without some basis for trusting the system, we lose faith in the decisions we see; we form doubts about our own courses of action. The conflict turns inward, and we start to defeat ourselves.

And this brings us to the third theme -- the human dimension of AI. The need to design AI in the context of its intended applications, to achieve dependability and trust in each domain, suggests that at least for now there will continue to be a strongly human component to its use. It makes us appreciate the myriad ways in which humans still chaperone our interactions with AI -- when they are built, when they are used, and when their results are interpreted -- and the importance these roles can have.

We can likely expect an increasing variety of these triangular interactions between an expert facilitator, an end user, and an underlying AI system --- for example between a health-care worker, a patient, and a medical AI application; or a teacher, a student, and a piece of educational technology. One of my earliest experiences with this chaperoned three-party dynamic of human-AI interaction involved Deep Blue as well. As a graduate student in summer 1995, feeling slightly star-struck, I sat down to play the Deep Blue prototype during an afternoon of demos for summer interns at IBM Watson. Beyond the game itself, I remember the conversation with Murray Campbell on the other side of the board as he moved the pieces for the computer, discussing at various points what Deep Blue was seeing and expecting, and at the inevitable end, explaining in the gentlest possible way that Deep Blue's king was going to make it all the way across the board from the square g8 to a hiding place on a7, and I wasn't going to get the draw by perpetual check that I was looking for.

Increasing the richness of the human-AI ecosystem in a way that makes AI safer, more trustworthy, and less disruptive to use is a goal that will continue to require all our innovation. Adding appropriate layers of human context to the interaction is one important direction toward this goal. Another potential direction is the creation of AI systems that are more human-like in their behavior, so that the experience of working with them directly may begin to feel less alien. In the context of chess, several colleagues (Ashton Anderson, Reid McIlroy-Young, and Siddhartha Sen) and I have explored this idea by creating a new chess-playing program, Maia, that can be tuned to match the moves of human players of different skill levels much more accurately than standard chess engines can achieve. It is an example of a different kind of optimization -- not aimed at coming closer to perfection in solving the task, but at more closely resembling the human approach to the task.

In this, as in so many related questions, chess is a potent model system precisely because it is so structured, so well-defined -- a self-contained world that is complex enough to force us to confront and solve hard challenges, but clear-cut enough to provide a relatively concrete sense for when we've succeeded. Deep Blue is arguably the most visible chapter in the history of this research paradigm, and it foreshadowed a set of challenges that the broader field of AI is still actively in the middle of resolving: how to anticipate the consequences of these systems as we introduce them into a series of increasingly high-stakes domains; how to make sure we can explain their decisions to the individuals whose lives they affect; and how to ensure productive coexistence between the behavior of AI and the behavior of human beings.

These questions will continue to challenge us for the foreseeable future. For now, though, let's think back to a moment in the history of computing when two intelligent beings -- a human with his support team; a computer with its designers -- briefly engaged across a chessboard and struggled to understand each other before going their separate ways.