DECEITFUL MEDIA

Artificial Intelligence and
Social Life after the Turing Test

SIMONE NATALE

Deceitful Media

DECEITFUL MEDIA

Artificial Intelligence and Social Life after the Turing Test

Simone Natale

OXFORD
UNIVERSITY PRESS

CONTENTS

ACKNOWLEDGMENTS

When I started working on this book, I had an idea about a science fiction story. I might never write it, so I reckon it is just fine to give up its plot here. A woman, Ellen, is awakened by a phone call. It's her husband. There is something strange in his voice; he sounds worried and somehow out of tune. In the close future in which this story is set, artificial intelligence (AI) has become so efficient that a virtual assistant can make calls on your behalf by reproducing your own voice, and the simulation will be so accurate as to trick even your close family and friends. Ellen and her husband, however, have agreed that they would never use AI to communicate between them. Yet in the husband's voice that morning there is something that doesn't sound like him. Later, Ellen discovers that her husband has died that very night, a few hours before the time of their call. The call should have been made by an AI assistant. Dismayed by her loss, she listens to the conversation again and again until she finally picks up some hints to solve the mystery. In fact, this science fiction story I haven't written is also a crime story. To learn the truth about her husband's death, Ellen will need to interpret the content of the conversation. In the process, she will also have to establish whether the words came from her husband's, from the machine that imitated him, or from some combination of the two.

This book is not science fiction, yet like much science fiction, it is also an attempt to make sense of technologies whose implications and meaning we are just starting to understand. I use the history of AI—a surprisingly long one for technologies that are often presented as absolute novelties—as a compass to orient my exploration. I started working on this book in 2016. My initial idea was to write a cultural history of the Turing test, but my explorations brought exciting and unexpected discoveries that made the final project expand much beyond that.

A number of persons read and commented on early drafts of this work. My editor, Sarah Humphreville, not only believed in this project

I remember a visit I made to one of Queen Victoria's residences, Osborne on the Isle of Wight. . . . Prominent among the works displayed there was a life-size marble sculpture of a large furry dog, a portrait of the Queen's beloved pet "Noble." The portrait must have been as faithful as the dog undoubtedly was—but for the lack of color it might have been stuffed. I do not know what impelled me to ask our guide, "May I stroke him?" She answered, "Funny you want to do that; all the visitors who pass stroke him—we have to wash him every week." Now, I do not think the visitors to Osborne, myself included, are particularly prone to magic beliefs. We did not think the image was real. But if we had not thought it somewhere we would hardly have reacted as we did—that stroking gesture may well have been compounded of irony, playfulness, and a secret wish to reassure ourselves that after all the dog was only of marble.

<div style="text-align:right">—Ernst Gombrich, Art and Illusion</div>

Introduction

In May 2018, Google gave a public demonstration of its ongoing project Duplex, an extension of Google Assistant programmed to carry out phone conversations. Google's CEO, Sundar Pichai, presented the recording of a conversation in which the program mimicked a human voice to book an appointment with a hair salon. Duplex's synthetic voice featured pauses and hesitation in an effort to sound more credible. The strategy appeared to work: the salon representative believed she was speaking with a real person and accepted the reservation.[1]

In the following weeks, Duplex's apparent achievements attracted praise, but also criticism. Commentaries following the demonstration highlighted two problems about the demo. On one side, some contended that Duplex operated "straight up, deliberate deception,"[2] opening new ethical questions regarding the capacity of an artificial intelligence (AI) to trick users into believing it is human. On the other side, some expressed doubts about the authenticity of the demo. They pointed to a series of oddities in the recorded conversations: the businesses, for instance, never identified themselves, no background noise could be heard, and the reservation-takers never asked Duplex for a contact number. This suggested that Google might have doctored the demo, faking Duplex's capacity to pass as human.[3]

The controversy surrounding Duplex reflects a well-established dynamic in the public debate about AI. Since its inception in the 1950s, the achievements of AI have often been discussed in binary terms: either exceptional powers are attributed to it, or it is dismissed as a delusion and a fraud.[4] Time after time, the gulf between these contradictory assessments has jeopardized our capacity to recognize that the true impact of AI is more nuanced and oblique than usually acknowledged. The same risk is present today, as commentators appear to believe that the question should be whether or not Duplex is able to pass as a human. However, even if Google's gadget proved unable to pass as human, we should not believe the illusion to be dispelled. Even in the absence of deliberate misrepresentation, AI technologies entail forms of deception that are perhaps less evident and straightforward but deeply impact societies. We should regard deception not just as a possible way to employ AI but as a constitutive element of these technologies. Deception is as central to AI's functioning as the circuits, software, and data that make it run.

This book argues that, since the beginning of the computer age, researchers and developers have explored the ways users are led to believe that computers are intelligent. Examining the historical trajectory of AI from its origins to the present day, I show that AI scientists have incorporated knowledge about users into their efforts to build meaningful and effective interactions between humans and machines. I call, therefore, for a recalibration of the relationship between deception and AI that critically questions the ways computing technologies draw on specific aspects of users' perception and psychology in order to create the illusion of AI.

One of the foundational texts for AI research, Alan Turing's *Computing Machinery and Intelligence* (1950), set up deception as a likely outcome of interactions between humans and intelligent computers. In his proposal for what is now commonly known as the Turing test, he suggested evaluating computers on the basis of their capacities to deceive human judges into believing they were human. Although tricking humans was never the main objective of AI, computer scientists adopted Turing's intuition that whenever communication with humans is involved, the behavior of the human users informs the meaning and impact of AI just as much as the behavior of the machine itself. As new interactive systems that enhanced communications between humans and computers were introduced, AI scientists began more seriously engaging with questions of how humans react to seemingly intelligent machines. The way this

dynamic is now embedded in the development of contemporary AI voice assistants such as Google Assistant, Amazon's Alexa, and Apple's Siri signals the emergence of a new kind of interface, which mobilizes deception in order to manage the interactions between users, computing systems, and Internet-based services.

Since Turing's field-defining proposal, AI has coalesced into a disciplinary field within cognitive science and computer science, producing an impressive range of technologies that are now in public use, from machine translation to the processing of natural language, and from computer vision to the interpretation of medical images. Researchers in this field nurtured the dream—cherished by some scientists while dismissed as unrealistic by others—of reaching "strong" AI, that is, a form of machine intelligence that would be practically indistinguishable from human intelligence. Yet, while debates have largely focused on the possibility that the pursuit of strong AI would lead to forms of consciousness similar or alternative to that of humans, where we have landed might more accurately be described as the creation of a range of technologies that provide an *illusion* of intelligence—in other words, the creation not of intelligent beings but of technologies that humans perceive as intelligent.

Reflecting broader evolutionary patterns of narratives about technological change, the history of AI and computing has until now been mainly discussed in terms of technological capability.[5] Even today, the proliferation of new communicative AI systems is mostly explained as a technical innovation sparked by the rise of neural networks and deep learning.[6] While approaches to the emergence of AI usually emphasize evolution in programming and computing technologies, this study focuses on how the development of AI has also built on knowledge about users.[7] Taking up this point of view helps one to realize the extent to which tendencies to project agency and humanity onto things makes AI potentially disruptive for social relations and everyday life in contemporary societies. This book, therefore, reformulates the debate on AI on the basis of a new assumption: that what machines are changing is primarily us, humans. "Intelligent" machines might one day revolutionize life; they are already transforming how we understand and carry out social interactions.

Since AI's emergence as a new field of research, many of its leading researchers have professed to believe that humans are fundamentally similar to machines and, consequently, that it is possible to create a computer that equals or surpasses human intelligence in all aspects and areas. Yet

entertaining a similar tenet does not forcefully contrast with and is often complementary to the idea that existing AI systems provide only the illusion of human intelligence. Throughout the history of AI, many have acknowledged the limitations of present systems and focused their efforts on designing programs that would provide at least the appearance of intelligence; in their view, "real" or "strong" AI would come through further progress, with their own simulation systems representing just a step in that direction.[8] Understanding how humans engage in social exchanges, and how they can be led to treat things as social agents, became instrumental to overcoming the limitations of AI technologies. Researchers in AI thus established a direction of research that was based on the designing of technologies that cleverly exploited human perception and expectations to give users the impression of employing or interacting with intelligent systems. This book demonstrates that looking at the development across time of this tradition—which has not yet been studied as such—is essential to understanding contemporary AI systems programmed to engage socially with humans. In order to pursue this agenda, however, the problem of deception and AI needs to be formulated under new terms.

ON HUMANS, MACHINES, AND "BANAL DECEPTION"

When the great art historian Ernst Gombrich started his inquiry into the role of illusion in the history of art, he realized that figurative arts emerge within an interplay between the limits of tradition and the limits of perception. Artists have always incorporated deception into their work, drawing on their knowledge both of convention and of mechanisms of perception to achieve certain effects on the viewer.[9] But who would blame a gifted painter for employing deceit by playing with perspective or depth to make a tableau look more convincing and "real" in the eyes of the observer?

While this is easily accepted from an artist, the idea that a software developer employs knowledge about how users are deceived in order to improve human-computer interaction is likely to encounter concern and criticism. In fact, because the term *deception* is usually associated with malicious endeavors, the AI and computer science communities have proven resistant to discussing their work in terms of deception, or have discussed deception as an unwanted outcome.[10] This book, however, contends that deception is a constitutive element of human-computer interactions rooted in AI technologies. We are, so to say, programmed to be deceived, and modern media have emerged within the spaces opened by the limits and affordances of our capacity to fall into illusion. Despite their resistance

to consider deception as such, computer scientists have worked since the early history of their field to exploit the limits and affordances of our perception and intellect.[11]

Deception, in its broad sense, involves the use of signs or representations to convey a false or misleading impression. A wealth of research in areas such as social psychology, philosophy, and sociology has shown that deception is an inescapable fact of social life with a functional role in social interaction and communication.[12] Although situations in which deception is intentional and manifest, such as frauds, scams, and blatant lies, shape popular understandings of deception, scholars have underlined the more disguised, ordinary presence of deception in everyday experience.[13] Many forms of deception are not so clear-cut, and in many cases deception is not even understood as such.[14]

Moving from a phenomenological perspective, philosopher Mark A. Wrathall influentially argued that our capacity to be deceived is an inherent quality of our experience. While deception is commonly understood in binary terms, positing that one might either be or not be deceived, Wrathall contends that such a dichotomy does not account for how people perceive and understand external reality: "it rarely makes sense to say that I perceived either truly or falsely" since the possibility of deception is ingrained in the mechanisms of our perception. If, for instance, I am walking in the woods and believe I see a deer to my side where in fact there is just a bush, I am deceived; yet the same mechanism that made me see a deer where it wasn't—that is, our tendency and ability to identify patterns in visual information—would have helped me, on another occasion, to identify a potential danger. The fact that our senses have shortcomings, Wrathall points out, represents a resource as much as a limit for human perception and is functional to our ability to navigate the external world.[15] From a similar point of view, cognitive psychologist Donald D. Hoffman recently proposed that evolution has shaped our perceptions into useful illusions that help us navigate the physical world but can also be manipulated through technology, advertising, and design.[16]

Indeed, the institutionalization of psychology in the late nineteenth and early twentieth centuries already signaled the discovery that deception and illusion were integral, physiological aspects of the psychology of perception.[17] Understanding deception was important not much or not only in order to study how people misunderstood the world but also to study how they perceived and navigated it.[18] During the nineteenth and twentieth centuries, the accumulation of knowledge about how people were deceived informed the development of a wide range of media technologies and practices, whose effectiveness exploited the affordances and limitations

of our senses of seeing, hearing, and touching.[19] As I demonstrate in this book, AI developers, in order to produce their outcomes, have continued this tradition of technologies that mobilize our liability to deception. Artificial intelligence scientists have collected information and knowledge about how users react to machines that exhibit the appearance of intelligent behaviors, incorporating this knowledge into the design of software and machines.

One potential objection to this approach is that it dissolves the very concept of deception by equating it with "normal" perception. I contend, however, that rejecting a binary understanding of deception helps one realize that deception involves a wide spectrum of situations that have very different outcomes but also common characteristics. If on one end of the spectrum there are explicit attempts to mislead, commit fraud, and tell lies, on the other end there are forms of deception that are not so clear-cut and that, in many cases, are not understood as such.[20] Only by identifying and studying less evident dynamics of deception can we develop a full understanding of more evident and straight-out instances of deception. In pointing to the centrality of deception, therefore, I do not intend to suggest that all forms of AI have hypnotic or manipulative goals. My main goal is not to establish whether AI is "good" or "bad" but to explore a crucial dimension of AI and interrogate how we should proceed in response to this.

Home robots such as Jibo and companion chatbots such as Replika, for example, are designed to appear cute and to awaken sentiments of empathy in their owners. This design choice looks in itself harmless and benevolent: these technologies simply work better if their appearance and behavior stimulate positive feelings in their users.[21] The same characteristics, however, will appear less innocent if the companies producing these systems start profiting from these feelings in order to influence users' political opinions. Home robots and companion chatbots, together with a wide range of AI technologies programmed to enter into communication with humans, structurally incorporate forms of deception: elements such as appearance, a humanlike voice, and the use of specific language expressions are designed to produce specific effects in the user. What makes this less or more acceptable is not the question whether there is or is not deception but what the outcomes and the implications are of the deceptive effects produced by any given AI technology. Broadening the definition of deception, in this sense, can lead to improving our comprehension of the potential risks of AI and related technologies, counteracting the power of the companies that gain from the user's interactions with these technologies and stimulating broader investigations of whether such interactions pose any potential harm to the user.

To distinguish from straight-out and deliberate deception, I propose the concept of *banal deception* to describe deceptive mechanisms and practices that are embedded in media technologies and contribute to their integration into everyday life. Banal deception entails mundane, everyday situations in which technologies and devices mobilize specific elements of the user's perception and psychology—for instance, in the case of AI, the all-too-human tendency to attribute agency to things or personality to voices. The word "banal" describes things that are dismissed as ordinary and unimportant; my use of this word aims to underline that these mechanisms are often taken for granted, despite their significant impact on the uses and appropriations of media technologies, and are deeply embedded in everyday, "ordinary" life.[22]

Different from approaches to deliberate or straight-out deception, banal deception does not understand users and audiences as passive or naïve. On the contrary, audiences actively exploit their own capacity to fall into deception in sophisticated ways—for example, through the entertainment they enjoy when they fall into the illusions offered by cinema or television. The same mechanism resonates with the case of AI. Studies in human-computer interaction consistently show that users interacting with computers apply norms and behaviors that they would adopt with humans, even if these users perfectly understand the difference between computers and humans.[23] At first glance, this seems incongruous, as if users resist and embrace deception simultaneously. The concept of banal deception provides a resolution of this apparent contradiction. I argue that the subtle dynamics of banal deception allow users to embrace deception so that they can better incorporate AI into their everyday lives, making AI more meaningful and useful to them. This does not mean that banal deception is harmless or innocuous. Structures of power often reside in mundane, ordinary things, and banal deception may finally bear deeper consequences for societies than the most manifest and evident attempts to deceive.

Throughout this book, I identify and highlight five key characteristics that distinguish banal deception. The first is its everyday and ordinary character. When researching people's perceptions of AI voice assistants, Andrea Guzman was surprised by what she sensed was a discontinuity between the usual representations of AI and the responses of her interviewees.[24] Artificial intelligence is usually conceived and discussed as extraordinary: a dream or a nightmare that awakens metaphysical questions and challenges the very definition of what means to be human.[25] Yet when Guzman approached users of systems such as Siri, the AI voice assistant embedded in iPhones and other Apple devices, she did not find that the users were questioning the boundaries between humans and machines.

Instead, participants were reflecting on themes similar to those that also characterize other media technologies. They were asking whether using the AI assistant made them lazy, or whether it was rude to talk on the phone in the presence of others. As Guzman observes, "neither the technology nor its impact on the self from the perspective of users seemed extraordinary; rather, the self in relation to talking AI seemed, well, ordinary—just like any other technology."[26] This ordinary character of AI is what makes banal deception so imperceptible but at the same time so consequential. It is what prepares for the integration of AI technologies into the fabrics of everyday experience and, as such, into the very core of our identities and selves.[27]

The second characteristic of banal deception is functionality. Banal deception always has some potential value to the user. Human-computer interaction has regularly employed representations and metaphors to build reassuring and easily comprehensible systems, hiding the complexity of the computing system behind the interface.[28] As noted by Michael Black, "manipulating user perception of software systems by strategically misrepresenting their internal operations is often key to producing compelling cultural experiences through software."[29] Using the same logic, communicative AI systems mobilize deception to achieve meaningful effects. The fact that users behave socially when engaging with AI voice assistants, for instance, has an array of pragmatic benefits: it makes it easier for users to integrate these tools into domestic environments and everyday lives, and presents possibilities for playful interaction and emotional reward.[30] Being deceived, in this context, is to be seen not as a misinterpretation by the user but as a response to specific affordances coded into the technology itself.

The third characteristic of banal deception is obliviousness: the fact that the deception is not understood as such but taken for granted and unquestioned. The concept of "mindless behavior" has been already used to explain the apparent contradiction, mentioned earlier, of AI users understanding that machines are not human but still to some extent treating them as such.[31] Researchers have drawn from cognitive psychology to describe mindlessness as "an overreliance on categories and distinctions drawn in the past and in which the individual is context-dependent and, as such, is oblivious to novel (or simply alternative) aspects of the situation."[32] The problem with this approach is that it implies a rigid distinction between mindfulness and mindlessness whereby only the latter leads to deception. When users interact with AI, however, they also replicate social behaviors and habits in self-conscious and reflective ways. For instance, users carry out playful exchanges with AI voice assistants, although they

know too well the machine will not really get their jokes. They wish them goodnight before going to bed, even if aware that they will not "sleep" in the same sense as humans do.[33] This suggests that distinctions between mindful and mindless behaviors fail to capture the complexity of the interaction. In contrast, obliviousness implies that while users do not thematize deception as such, they may engage in social interactions with the machine deliberately as well as unconsciously. Obliviousness also allows the user to maintain at least the illusion of control—this being, in the age of user-friendliness, a key principle of software design.[34]

The fourth characteristic of banal deception is its *low definition*. While this term is commonly used to describe formats of video or sound reproduction with lower resolution, in media theory the term has also been employed in reference to media that demand more participation from audiences and users in the construction of sense and meaning.[35] For what concerns AI, textual and voice interfaces are low definition because they leave ample space for the user to imagine and attribute characteristics such as gender, race, class, and personality to the disembodied voice or text. For instance, voice assistants do not present at a physical or visual level the appearance of the virtual character (such as "Alexa" or "Siri"), but some cues are embedded in the sounds of their voices, in their names, and in the content of their exchanges. It is for this reason that, as shown in research about people's perceptions of AI voice assistants, different users imagine AI assistants in different, multiple ways, which also enhances the effect of technology being personalized to each individual.[36] In contrast, humanoid robots leave less space for the users' imagination and projection mechanisms and are therefore not low definition. This is one of the reasons why disembodied AI voice assistants have become much more influential today than humanoid robots: the fact that users can project their own imaginations and meanings makes interactions with these tools much more personal and reassuring, and therefore they are easier to incorporate into our everyday lives than robots.[37]

The fifth and final defining characteristic of banal deception is that it is not just imposed on users but also is programmed by designers and developers. This is why the word *deception* is preferable to *illusion*, since *deception* implies some form of agency, permitting clearer acknowledgment of the ways developers of AI technologies work toward achieving the desired effects. In order to explore and develop the mechanisms of banal deception, designers need to construct a model or image of the expected user. In actor-network theory, this corresponds to the notion of script, which refers to the work of innovators as "inscribing" visions or predictions about the world and the user in the technical content of the new object and

technology.[38] Although this is always an exercise of imagination, it draws on specific efforts to gain knowledge about users, or more generally about "humans." Recent work in human-computer interaction acknowledges that "perhaps the most difficult aspect of interacting with humans is the need to model the beliefs, desires, intentions preferences, and expectations of the human and situate the interaction in the context of that model."[39] The historical excavation undertaken in this book shows that this work of modeling users is as old as AI itself. As soon as interactive systems were developed, computer scientists and AI researchers explored how human perception and psychology functioned and attempted to use such knowledge to close the gap between computer and user.[40]

It is important to stress that for us to consider the agency of programmers and developers who design and prepare for use AI systems is perfectly compatible with the recognition that users themselves have agency. As much critical scholarship on digital media shows, in fact, users of digital technologies and systems often subvert and reframe the intentions and expectations of companies and developers.[41] This does not imply, however, that the latter do not have an expected outcome in mind. As Taina Bucher recently remarked, "the cultural beliefs and values held by programmers, designers, and creators of software matter": we should examine and question their intentions despite the many difficulties involved in reconstructing them retrospectively from the technology and its operations.[42]

Importantly, the fact that banal deception is not to be seen as negative by default does not mean that its dynamics should not be the subject of attentive critical inquiry. One of the key goals of this book is to identify and counteract potentially problematic practices and implications that emerge as a consequence of the incorporation of banal deception into AI. Unveiling the mechanisms of banal deception, in this sense, is also an invitation to interrogate what the "human" means in the discursive debates and practical work that shape the development of AI. As the trajectory described in this book demonstrates, the modeling of the "human" that has been developed throughout the history of AI has in fact been quite limited. Even as computer access has progressively been extended to wider potential publics, developers have often envisioned the expected user as a white, educated man, perpetuating biases that remain inherent in contemporary computer systems.[43] Furthermore, studies and assumptions about how users perceive and react to specific representations of gender, race, and class have been implemented in interface design, leading for instance to gendered characterizations of many contemporary AI voice assistants.[44]