THE PROSPECT OF A HUMANITARIAN ARTIFICIAL INTELLIGENCE

Agency and Value Alignment

CARLOS MONTEMAYOR

BLOOMSBURY

The Prospect of a Humanitarian Artificial Intelligence

Also available from Bloomsbury:

Great Philosophical Objections to Artificial Intelligence,
by Eric Dietrich, Chris Fields, John P. Sullins,
Bram Van Heuveln, and Robin Zebrowski
Philosophy in a Technological World, by James Tartaglia
The Evolution of Consciousness, by Paula Droege
The Human Mind through the Lens of Language, by Nirmalangshu Mukherji
The Philosophy and Science of Predictive Processing,
edited by Dina Mendonça, Manuel Curado, and Steven S. Gouveia

The Prospect of a Humanitarian Artificial Intelligence

Agency and Value Alignment

Carlos Montemayor

BLOOMSBURY ACADEMIC
LONDON • NEW YORK • OXFORD • NEW DELHI • SYDNEY

BLOOMSBURY ACADEMIC
Bloomsbury Publishing Plc
50 Bedford Square, London, WC1B 3DP, UK
1385 Broadway, New York, NY 10018, USA
29 Earlsfort Terrace, Dublin 2, Ireland

BLOOMSBURY, BLOOMSBURY ACADEMIC and the Diana logo are
trademarks of Bloomsbury Publishing Plc

First published in Great Britain 2023

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

ISBN: HB: 978-1-3503-4837-0
ePDF: 978-1-3503-4838-7
eBook: 978-1-3503-4839-4

Typeset by Integra Software Services Pvt. Ltd.

To find out more about our authors and books visit www.bloomsbury.com
and sign up for our newsletters.

Contents

List of Figures

Preface

We are an adventurous and curious species. Our intelligence is our point of pride and the basis of our endless interests and achievements. The various needs and motivations we developed as we evolved have made us both mighty and vulnerable. We inherited our intelligence from an astonishingly diverse lineage, and we share much of our intelligence with other species. Yet our intelligence is unique in its range and power. What is more, our collective intelligence has profoundly shaped the planet, and not always to the good. For instance, without the scientific and industrial revolutions, nuclear weapons would never have been developed. The destructive by-products of the recent expansion and mechanization of our intelligence are part of our vulnerability.

This book centers on the possibility of creating another colossal industrial achievement, just as threatening and awesome as the atomic bomb: the set of computational techniques that may result in artificial intelligence (AI). This possibility does not exist at present, and we do not know how likely is the prospect of its existence. What we know is that the industrial revolution would be dwarfed, a small step in our history, compared to this potential development. Ironically, AI might also decisively demote our species by making us less intelligent than our creation, and this has been identified as the most important risk surrounding AI: the so-called "singularity," or the moment when we are left behind (far behind, some fear), outpaced by our newly created super-intelligent systems. Some have posited that we may find a solution in value alignment, in ensuring that the computational techniques we create reflect our ethics and norms, ensuring that it will be beneficial rather than destructive. The value alignment problem is a major theme of this book.

Value alignment appears to be the best answer to AI critics' vague predictions of doom and demotion. A key contribution of this book is to note that there is no unique "value alignment problem," but a variety of them. That is because there are various forms of value alignment, which must be carefully distinguished from one another. These distinct alignments require meaningful specifications dependent on contextual information and skills for coordinated action in different normative domains (moral, epistemic, and political). Human attunement to what is salient, evidentially important, or morally relevant

depends on how we align our attention. These attentive alignments allow us to triangulate with each other with respect to what we value. The book, however, does not merely focus on how to avoid the singularity through strategies for value alignment with AI. A central concern is the control and power—epistemic, social, and political—that AI research is *already* exerting and accumulating.

How probable is it that genuinely intelligent artificial agents will be created? A clear answer to this question remains elusive, though commentators have speculated with tremendous enthusiasm. Echoing William James' opinion about the state of psychology at his time, Terry Winograd said in 1977 that developments in artificial intelligence were akin to alchemy—we combine, mix, and scale up, in the hopes of finding the holy grail, but without knowing when or how exactly we will succeed. More recently, Melanie Mitchell has returned to the question, whether today's AI research still resembles alchemy in "Why AI Is Harder than We Think." Her verdict is that, despite enormous progress, large financial investments, and public media hype, AI research more than forty years after Winograd's remark may still be industrial alchemy—nothing more than a utopian hope.

Mitchell presents several methodological fallacies in AI research that impede its development. One of them, "wishful mnemonics," is particularly relevant for this book: AI researchers import terms from philosophy, psychology, and neuroscience (e.g., neural networks, attention, intelligence, knowledge, truth) without any real justification. This is a kind of wishful labeling that would, by alchemical magic, turn massive data mining and layered statistical processing into thoughtful agency. Gratuitous use of psychological terms in descriptions of machine output muddles our evaluations of AI, and confuses everyone with respect to the real nature and potential of AI technology. This book proposes a useful way to categorize different kinds of AI, without simply assuming that psychological terms can be directly applied to all of them equally.

However, if and when AI is developed, the consequences will be dramatic. We need to prepare ourselves for this eventuality by making our technology more beneficial to all humans, starting with the technology we are using now. AI research technology is not benefiting humanity in any clear or measurable way, and on the contrary, it has been misused and can potentially become quite dangerous. An important reference point here is Kate Crawford's book, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. As Crawford forcefully argues, the current investment frenzy in AI technologies is profoundly troubling because of its economic, environmental, social, and political costs. Power and justice are also a central concern of my examination of AI.

As a matter of basic fairness, if intelligence is developed industrially, then it should be used for the benefit of all humanity. It should also be accessible in various ways to all human beings, and not be the exclusive property of a few powerful impresarios or politicians who might keep AI as an industrial or military secret, which is consistent with what is currently happening. This stolen intelligence could potentially be used against all of us. Such a scenario presents a more perverse kind of being "left behind," a political maneuver to demote not ultra-rich or ultra-powerful humans by expelling them from the production of knowledge. Instead of a "terminator" scenario in which machines take over, here we have a more standard form of unfairness or "business as usual." As a countermeasure to this possibility, this book argues in favor of a humanitarian approach to the development of AI technology, based on the existent international framework for the protection of human rights. These problems are urgent, whether or not the pursuit of AI turns out to be a form of alchemy, and whether or not AI agents become genuinely intelligent.

Another goal of this book is to demonstrate that we need to distinguish between formidable predictive computing that looks like intelligence and genuinely intelligent agency, a goal that can only be achieved by firmly grounding our arguments in philosophical considerations. With respect to psychology, critical distinctions are drawn from the empirical and philosophical literature on intelligence, attention, consciousness, knowledge, and emotion. Value theory, as well as political and legal philosophy, informs the book's proposals concerning the value alignment problem, based on an analysis of human needs. Types of intelligence are examined according to their normative, or value-grounding roles. The importance of control and motivation for intelligent behavior is examined according to basic human needs, which are then shown to be critical in our understanding of human rights and human dignity.

If the reader is mostly interested in philosophy of mind, epistemology or the theory of knowledge, and their relation to motivations, needs, and intelligence, her focus should be on the first five chapters of the book. The first part is about the conditions under which current technology could be considered genuine intelligence, and the various problems that these assessments confront. If she is more interested in social, political, or legal issues, then the concluding three chapters are the most relevant. The second part is about why AI may be best understood as a form of collective epistemic agency, and the urgent political repercussion of such a conceptualization of AI. But I would like to encourage the reader to go through all of the chapters in order because the account in the first part of the book is deeply tied to the notion of human dignity explained in the second.

Poverty, environmental harm, and political polarization are realities that the AI industry should seek to ameliorate. The utopian dream of democratizing technology, which originated with the development of the internet and search algorithms, has backfired, turning algorithmic technology into monetized surveillance of our behavior. We must now enforce the protections and liberties grounded on human dignity against the misuse of technology, based on a basic humanitarian framework for all future technology. Only then can we hope that AI will fulfill the promise of its benefits. At its heart, the book urges the reader to appreciate why AI justice and democracy must be pursued right away. Engaging with philosophical considerations on AI, the reader may be also considering the very limits, moral and epistemic, of human creativity. The issues that make AI so fascinating are the same issues that make *us* so fascinating. Our curiosity about intelligence may motivate us to revisit the idea that since we are all intelligent, we all share the same dignity and should be treated accordingly, with important repercussions for the ways in which we administer and control the deployment of technology.

Acknowledgments

I am forever grateful to the members of the Object Group in New York City, particularly to Fuat Balci, Susan Carey, Harry H. Haladjian, Rochel Gelman, Brian Keane, Zenon Pylyshyn, Brian Scholl, and Anne Triesman. Attention and intelligence are deeply related psychological capacities, and I learned from our group discussions how the attentive organization of objects in perceptual scenes articulates various kinds of intelligence. Alvin Goldman's work and advice have had a deep influence on my views, which is reflected in the last chapters of the book. My gratitude to my friend and collaborator Abrol Fairweather for so many insights and discussions that inform much of what follows. I am also grateful to Greyson Abid, Jackson Kernion, Geoffrey Lee, Anna-Sara Malmgren, Nico Orlandi, Antonia Peacocke, and other members of the Berkeley philosophy of mind reading group. Our meetings were always inspiring and informative.

Many other groups, friends, and colleagues provided support and feedback leading to the completion of this book. I thank the Santa Fe Institute, and the audience at the event "The Limits of Understanding," which took place in November of 2017. I am also grateful for my experience participating in the 2021 Diverse Intelligences Summer Institute, particularly to Erica Cartmill and Jacob Foster. I was very lucky to meet with Kevin Kelly for a fruitful conversation on different kinds of intelligences and the unity of knowledge while we overlapped in Budapest in 2017. My deep gratitude to Gary Bengier for his friendship and the multiple long conversations on AI and many other related topics, including the future of automation. I extend special thanks to my friend and collaborator Garrett Mindt, particularly for his help with the images and the material on AI and social epistemology. To Jodi Halpern, Iris Oved, Susan Schneider, and Anand Vaidya, thank you for your constant support and insights. Special thanks to my friends and collaborators at the Consciousness Group at Stanford: Acacio de Barros, Leonardo Guimaraes de Assis, John Perry, and Paul Skokowski. With an equal amount of gratitude, many thanks to The Friends of Attention, particularly to D. Graham Burnett and Justin E. H. Smith.

San Francisco State University, where I have taught philosophy for more than ten years, has shaped the two main thematic orientations of this book: the philosophy and politics of AI. It is the State college of the city of San Francisco,

where the influence of Silicon Valley is palpable and omnipresent. It is also an institution devoted to social justice with a unique history of promoting equality and diversity in education. My colleague and mentor Anita Silvers convinced me that I should combine these two orientations in my research. She unfortunately passed away in 2019. Anita's passion for activism and engaged philosophy remains a vital source of inspiration for me. Her influence is, hopefully, manifest in the pages that follow. I am very grateful to her, and to all my great colleagues at the Philosophy Department and at the University, especially Dragutin Petkovic and Denise Kleinrichert, who are my collaborators in the recently created Ethical AI Certificate at SF State. I am also grateful to Lynn Mahoney and Andy Harris for their support.

My greatest debt is to Victoria Frede. None of the work that went into this book could have been possible without her care and encouragement.

Glossary and Abbreviations

artificial intelligence (AI) The computational design of intelligent capacities based on information-processing techniques concerning large databases. These capacities concern inferential, perceptual, rational, linguistic, and other cognitive capacities regarding decision-making and rationality. The industrial simulation of intelligence includes the informational processing methods of deep learning, unsupervised learning, neural networks, and other techniques and approaches to learning and meta-learning (or learning how to learn), including evolutionary models. Historically, the two approaches to AI were data-based trained neural networks and symbolic, representation-based, and rule-following models. This dichotomy simplifies the complexity of today's techniques, which rely on insights from these two traditions and employs several layers of processing.

artificial general intelligence (AGI) Artificial intelligence that is comparable in scope and complexity to human intelligence. Since autonomy is so important for genuine intelligence, AGI would need to be autonomous in a similar way: it will learn and develop skills on her own, similarly to human children. The distinction between consciousness and attention clarifies which concrete aspects of AGI could resemble human intelligence and which may escape computational simulation efforts. AGI that relies exclusively on attention may eventually be possible. This is a positive outcome because the nature of consciousness is a notoriously difficult problem and cannot be assumed as a fundamental aspect of AGI design. Attention suffices for epistemic agency and, therefore, AGI with attention capacities should count as genuinely intelligent and autonomous (see the definitions of EEI and IEI below for further clarification).

attention The capacity to identify salient information while inhibiting irrelevant information. More specifically, the selective cognitive capacity that keeps track of information that is relevant for the satisfaction of the representational and cognitive needs of an agent. Attention is a type of mental agency because it involves mental action directed toward salient objects or properties. Since agency is a source of control over the mental and physical actions of an agent, agency is a fundamental basis for trust, responsibility, and credit. Attentive agency is required for normative interpretations of behavior, understood as good or bad

cognition, morally or epistemically. If AI becomes genuinely intelligent, it will display forms of attention that are equivalent to human and animal attention, as sources of control, trust, and responsibility.

autonomy The self-reliance of genuinely intelligent agents on their capacities to satisfy their goals and needs. Autonomous agents satisfy a variety of cognitive needs on the basis of their reliable attentive capacities. This variety of needs is associated with general intelligence and with the complex cognitive capacities of animals and humans, as well as with their worth and dignity, at least in the case of humans.

cognitive needs The basic conditions that must be met, as well as the problems that must be solved, in order to maintain the well-being of an agent by enhancing her cognitive capacities and developing her independence. An agent who satisfies her needs because of her own cognitive capacities and intelligence is autonomous because she is responsible for their satisfaction. The development and free satisfaction of cognitive needs are constitutive of the dignity of agents, particularly human beings. The systematic presence of these needs provides an objective basis for the value and dignity of humans. It also makes possible the coordination of efforts to satisfy needs through joint attention and collective action. An intelligent agent typically ranks her needs according to their value, and these needs include representational, biological, emotional, and rational needs. Humans also have unique needs, which they value the most, such as transcendence, spiritual, and autonomy needs.

collective artificial intelligence (CAI) A kind of artificial intelligence that resembles the intelligence of wide-ranging human cooperation, manifest in collectives and institutions, such as banks, governments, and scientific agencies. It is possible that artificial intelligence is best described as a collective type of agency.

consciousness Phenomenal consciousness is the subjective and qualitative character of experience, or what it is like to experience something from the first-person perspective, like pain or color. Access consciousness is the information agents have readily available for thought, decision-making, and action, including coordinated action and linguistic communication. Access consciousness is best understood as a cross-modal and integrated kind of attention that does not necessitate a specific or phenomenal qualitative character. Phenomenal consciousness may be fundamental for various kinds of moral and aesthetic experiences, but it may not be necessary for epistemic agency (see attention above).

consciousness and attention dissociation (CAD) The extent to which consciousness is independent from attention capacities. The dissociation between consciousness and attention may be absolute in the sense that consciousness and attention might be entirely different capacities. There may also be a dependence between them (attention may be necessary for consciousness). There are empirical and theoretical reasons that justify both claims, namely, that consciousness and attention are separate capacities and that attention is more fundamental, in the sense that consciousness seems to necessitate some kind of attention. Since attention may occur without consciousness and since attention is sufficient to explain epistemic agency and rationality, if AGI became attentive, it should count as genuinely intelligent even in the absence of human-like phenomenal consciousness.

epistemic agency The autonomous exercise of knowledge conducive capacities that allow humans and animals to satisfy a wide diversity of representational, rational, and communicational needs. According to the present proposal, epistemic agency relies fundamentally on attention, rather than on consciousness or subjective awareness.

extensionally equivalent intelligence (EEI) AI or AGI that is either equally or more reliable and accurate than human intelligent capacities (either itemized or in general). Although EEI performance might be the same or even superior in reliability, EEI systems cannot count as autonomous because they lack the required integrated agency that relates motivations with autonomous need satisfaction. In other words, EEI is strictly a simulation of intelligence without the required underlying capacities for problem solving and autonomous learning. It could be argued that all AI technology at present, even when at human level performance, is of this kind in the best of cases.

human rights The legal protections provided by States to humans, independently of their national status, in order to protect their dignity and integrity. Typically conceived as responsibilities of States, human rights transcend legal discourse because of their political importance as expressions of civil and intellectual freedom. Accordingly, human rights are protected under international law, as basic components of the humanitarian legal framework of the United Nations. Human rights protect the dignity of humans by seeking to satisfy their most basic needs, including the development of their capacities, autonomy, and freedom. Understood ethically, human rights are concerned with preventing States from interfering with the freedom of individuals (or negative freedom), and also with helping individuals develop their capacities (or positive freedom).

intensionally equivalent intelligence (IEI) Unlike EEI, IEI is not just a simulacrum of human intelligence. IEI will involve agency and motivations similar to those of humans because they will be similarly attentive, rather than based on data-driven brute-force simulation. Two systems may look exactly alike in terms of rate of success, but IEI will be performing tasks based on motivations to attend to stimuli in order to satisfy its own cognitive needs, while the EEI system would be parametrizing and optimizing information processing without agential guidance, and will be in need of agential interpretation outside itself. Attention suffices for AGI, but only if such AGI is IEI, because attention essentially involves reliable motivations in its guidance, selection, and selectivity functions. This kind of motivation-based satisfaction of needs is what is distinctive of autonomous agency. EEI, however, would be the most sophisticated kind of automation and it may generate considerable advances and risks, in spite of the fact that it won't count as fully agential.

moral agency The autonomous capacities for engagement and empathy of human cognition required for the satisfaction of moral and emotional needs. Because of its emotional components, moral agency may not only necessitate attention but, unlike epistemic agency, it may also necessitate phenomenal consciousness. This has the potential repercussion that there could never be genuinely autonomous, and therefore intelligent, moral IEI.

motivation The urge or impulse to act in order to achieve a goal. Intelligent agents have motivations that reliably conduce to the satisfaction of their needs on the basis of their capacities. Motivations are essential aspects of agents, and they include epistemic, moral, practical, rational, and emotional desires or needs. Intelligence requires that motivation be guided by attention in order to autonomously satisfy these needs.

normative The characteristic of being in compliance with a standard that allows for evaluations of performances as either good or bad. Virtuous agents satisfy their needs in a way that meet epistemic (more likelihood of truth) and moral (more morally praiseworthy actions) standards. Normativity is related to autonomy and the capacity of agents to improve by becoming more virtuous; see responsibility.

responsibility The condition of autonomous agents that makes them accountable for their actions and answerable to others when they request an explanation. Only IEI agents could be responsible in a non-lucky and genuine way because responsibility requires not only freedom to act autonomously but

also the necessary relation between the intention to act and the consequences of acting on that intention. There are various types of responsibility, among the most salient for present purposes are epistemic, moral, practical, and legal.

social epistemology Epistemology is the theory of knowledge. It examines the conditions under which beliefs are justified, either because they are based on good evidence or because they are more likely to be true than false. It also studies the nature of knowledge, curiosity, understanding, intelligence, and rationality, and the relation among these intellectual capacities. Social epistemology studies these capacities in the context of collective agents and social groups. CAI would be a kind of collective epistemic agency, and a question that emerges is whether it could qualify as IEI epistemic agency (see CAI above). Social epistemology focuses on how groups communicate, jointly attend, achieve goals, and satisfy their needs successfully as well as non-luckily.

value The worth or utility of individuals, their actions, and their goals. Autonomous agents have intrinsic value because of the good skills they employ in satisfying their goals and needs. Agents prioritize their needs according to different value assignments. Humans are approximately value aligned because they have similar moral, epistemic, and aesthetic needs (although this does not entail identity in values, or even in rankings of value). Attention is the key to solve many aspects of the value alignment problem with AI. The study of intelligence requires assessments of value and agency, and this is particularly interesting and complicated with respect to AI. For present purposes, a human rights framework is the best way to guarantee that AI will be ethical. AI will be ethical to the extent that it benefits humanity as a whole.

virtue The excellence or skillfulness of agents that perform tasks in different normative domains, particularly epistemology, morality, and aesthetics. Virtuous performance permits agents to satisfy their needs on the basis of their skilled capacities. In this sense, the type of virtuous capacities that allow agents to satisfy basic representational and cognitive needs are not exceptional, given the importance they have for human freedom and development. Other more unique skills build on these widespread ones, and require specialized habits and training. Skills help agents succeed in a non-lucky way and are an important aspect of their worth and dignity.

Introduction: Normative Aspects of AI Development

This book is about the diversity of intelligence. It centers on one of the most intriguing possibilities concerning intelligence, namely, the development of humanly designed intelligent machines. The examination of capacities that qualify as intelligence offered here is, therefore, not anthropocentric. It includes an extensive discussion of human and animal intelligence, drawing comparisons and identifying crucial differences with the case of intelligent machines. This analysis can, in principle, be extended to any intelligent agent and it allows for various types and degrees of intelligence without assuming that intelligence is a monolithic or uniform phenomenon. Intelligent agents have needs and the way in which they satisfy them through their capacities is at the heart of what makes them intelligent. A critical need of intelligent beings that is particularly important for humans is to pursue what they find important, interesting, and valuable.

Humans value their freedom and well-being—two aspects of human agency that are tightly connected. The freedom to accomplish one's own well-being is of chief moral importance, and is fundamental for the notion of human dignity. The achievement of satisfying various needs, including a basic need for unencumbered agency, is a goal that motivates agents to pursue various trajectories for action, rather than a rigid set of fixed solutions to problems. For instance, according to the so-called "capability approach," freedom depends on the development of capacities and abilities required for agents to successfully pursue their goals. The satisfaction of agential needs requires intelligence, learning, and education, which constitute personal virtues. It also depends on the support from a society that allows individuals to pursue their autonomous need-satisfaction according to what they value. Human dignity depends on the possibility to develop one's own capacities in order to pursue goals.[1]

The relation between intelligence and freedom that is relevant for present purposes concerns ethics and political philosophy. This is also the notion that

is relevant for issues concerning value alignment, including value alignment with potentially intelligent machines. Thus, metaphysical issues regarding the problem of free will are not germane to this discussion and they deserve an independent treatment. However, models of agency discussed in this rich literature are relevant to the extent that they explain motivation and action under the control of an agent.[2] As Gabriel (2020) says, value alignment should not be based on "the true" moral theory which should be programmed universally into machines. Instead of a metaphysical solution to problems of value and free will, we need an ethical and political framework to impose reasonable limits on what machines should do. The present proposal, consistent with this commitment, is to appeal to those cognitive capacities of agents that are constitutive of their dignity and freedom, and which allow them to satisfy a multiplicity of needs. In the chapters that follow, these capacities are specified descriptively and normatively in terms of the functions of attention.

Intelligence helps us satisfy various needs, individually and collectively. The motivation to satisfy our needs autonomously through one's own agency and without external oppression or determination is essential to our understanding freedom and flourishing. Our cognitive capacity to attend to what is salient and ignore what is irrelevant or harmful is fundamental to satisfy our needs. In the context of artificial intelligence (AI), the role of attentive need satisfaction opens up various puzzling questions. The relation between intelligence, needs, and autonomy in this new and rapidly developing context is the central topic of this book.

Attention—the capacity to reliably identify contents that are salient while inhibiting irrelevant information—is essential to all kinds of intelligence in humans and animals, as is argued at length below. Motivations to attend to specific contents are critical to how and why we take a certain course of action, how we interpret information, what peeks our interest, or what we find troubling. The motivation to attentively satisfy cognitive needs provides a unique and autonomous perspective on the world, based on the attentive capacities underlying various forms of intelligence: to know, to help others, and to learn.

Humans starting very early in childhood *learn how to learn*, partly by autonomously determining what they should pay attention to (see the discussion on Alan Turing's "child machine" below). Humans don't have a specific set of fixed preferences which they pursue single mindedly in order to optimize solutions to problems in a "robotic" fashion. They constantly change and reorganize their preferences and goals, not because they are irrational, but as part of their autonomous and curiously intelligent agency. Humans care

about *problems* that are salient and important to them—they never just meet fixed goals optimally. They are generous with their curiosity, and value new information, which is part of their capacity to learn how to learn new skills and develop or improve their capabilities; humans and animals learn by increasing the amount of attentional contents that are salient to them. This flexible and attentive inquisitiveness is essential to the notion of dignity enshrined in the human rights protecting freedom of thought, expression, and information, as well as the right to education.[3]

Humans negotiate what they value, individually and collectively. They satisfy their epistemic and moral needs by aligning their attention with trustworthy and valuable sources of information—typically other humans, but increasingly collective agents as well. Humans satisfy a multiplicity of needs based on attentive *trust*. The reliability of attention routines, perceptual and cognitive, intellectual and moral, as well as epistemic and communicative, is a key source of trust underlying human and animal intelligence. Trust resides on the fact that attentive capacities are not only reliable because they successfully solve problems and meet goals, but crucially because these are capacities under the autonomous guidance and control of agents. Agential trust is much more powerful than merely causal or mechanistic reliance. I can trust my fridge or my car because these mechanisms are reliable in performing specific tasks. But this is nothing compared to trusting my friend or teacher as a source of information concerning a vast array of topics required to satisfy a multiplicity of cognitive needs. Our trust in an agent is grounded on their capacity to attend to salient sources of value, and to identify relevant problems, rather than simply producing single-minded solutions once a problem is presented to them. This is key to understand various kinds of AI risks. It is also fundamental to understand the distinction between tool AI and genuine AI, including artificial general intelligence (AGI).

The relevant notion of agential control for our purposes is associated with responsibility: epistemic, moral, and legal. Agents are responsible for their actions because they control them or guide them through the exercise of their capabilities, fostered within a cultural and social milieu (Vargas, 2013). Freedom from constraints is relevant for responsibility only if one has the capabilities to achieve the goals one sets for oneself, and this is essential to the value of freedom and autonomy—that our success at meeting needs occurs under our agential control, for which we are responsible. Thus, the notion of attentional control developed in this book is essentially related to various kinds of trust and responsibility. Trust derives from the reliability of an agent's capacities, but also from the non-accidental relation between an agent's motivations to act or attend

to contents and her success at satisfying multiple needs. Agents are not lucky in their success because of their abilities and skills.

Since humans have quite diverse needs, preferences, and attentive capacities, value alignment is an intricate endeavor that frequently creates disagreement. Consider a case of personal conflict concerning the very notion of freedom. Agents value their freedom of choice. Governmental regulations that limit their choice of types of fuel consumption, car selection, or transportation lifestyle in general may be considered as restrictions to freedom and thus, may be considered unjustified—they indeed, after all, restrict the landscape of choices. But if the environment degrades further, health problems in humans and the extinction of species—a basic condition for the existence of a much larger number of capacities necessary for free agency—will be hindered. Thus, a substantially larger set of options would be restricted as a consequence. This very real and current dilemma demonstrates the difficulties of alignment, even when it comes to freedom. This is why, as explained in what follows, hierarchies of needs must somehow be aligned with collective values, specified by some kind of objective or consensus-based measure, rather than mere personal preferences. The capability approach to human development is particularly helpful in addressing this difficulty (Sen 1999, Nussbaum 2011; see in particular Binder 2019), because human capacities and needs are much more homogeneous than personalized sets of preferences.

Similar conflicts concerning conceptions of freedom emerge at the collective level. When Mexico signed NAFTA, the government allegedly liberated the Indigenous communities of Mexico by giving individuals property rights. But by changing the collectivized property framework they depended on, these communities were deprived of the basic social framework they relied upon in order to satisfy their needs. In this collective context, the capability approach is also helpful because of its focus on the freedom to choose a life path that is personally salient because one has reasons to value it (Binder and Binder, 2019).

Yet another complication is that epistemic value alignment can differ radically from moral value alignment, also leading toward conflict. What you believe and know to be wrong (socially and legally) could be not only morally permissible, but also morally obligatory. What is legally obligatory may be deeply immoral. In Mark Twain's *The Adventures of Huckleberry Finn*, Huckleberry Finn helps Jim, and this action is morally good, in spite of the fact that Finn believes that helping Jim is wrong. This case shows that being akratic (incoherent, or unwilling to follow the consequences of what we believe) is morally virtuous in this case. Belief would dictate that Huckleberry not help Jim, but he suppresses and ignores it because of the morally significant need to help Jim. What is salient

to Finn is not his belief regarding the wrongness of helping Jim, but rather his awareness of Jim's dignity (Arpaly, 2002). The morally right value alignment is achieved by ignoring the veridical belief that according to the politics and legal culture of Finn's time, helping Jim is wrong.

No one expects machines to solve this kind of intricate problem—as the previous examples show, humans struggle with them all the time. What is important to emphasize is that these difficulties become much more intractable if the possibility of AI materializes because AI agents will lack any of the needs that attention and related cognitive capacities are designed to satisfy—unless they become attentive and curious the way we are. If AI is only fake intelligence, these problems will exacerbate social misalignments, risks, and harms. If it is genuine intelligence and it solves problems better than us, then AI can create risks of hegemonic power and human enfeeblement. These and related difficulties concerning AI risk are addressed in the chapters that follow by systematically showing the centrality of attention capacities in solving various kinds of conflicts and value alignment. Thus, a fundamental problem that this book addresses is, what are the conditions AI must satisfy for it to count as genuinely intelligent?

Intelligence has a normative dimension: it guides agents with respect to how they *should* behave in multiple contexts. Attention is critical to explain this normative role because of its essentially guiding, selective, and sensitive functions, as is argued at length below. It is important to note that the term "attention" has been used in psychology and philosophy in a very different way from its current use in machine learning. Providing a framework for attentive intelligence that could be helpful in understanding genuinely intelligent AI is a central goal of this book. This framework explains and develops attention's relation to general intelligence and knowledge. Only until recently had philosophers started examining the psychologists' definition of attention. All definitions point at selectivity and sensitivity for contextualized action, which are key features of intelligent behavior. Here attention is defined in terms of reliable cognitive skill and ability, and its normative properties are understood in terms of virtues.[4] There is consensus that attention is critical for intelligent behavior and action. The key is that attention is also, because of its relation to action, deeply related to autonomy. Attentive capacities are essential for the agential satisfaction of needs and they provide the basis for various kinds of autonomy.[5]

The defining feature of AGI, and of any kind of intelligence, is the autonomy that attentive capacities provide for the agent-dependent solution of a multiplicity of problems, as specified by the hierarchy of needs of agents. The standard assumption in AI research but also in general is that rational behavior

or optimization is the key mark of intelligence. But as recent research in machine learning has shown, this assumption can lead to deception (Conti et al., 2017) by optimizing a solution to a problem that is not relevant, leading to various forms of inaccuracy portrayed as adequate. This book argues that while AGI may meet epistemic standards for intelligence, emotional and moral intelligence cannot be achieved as it occurs in humans and other animals. This is not because of anthropocentric chauvinism, but because of the complex types of needs humans satisfy, as the cases above concerning freedom and moral action demonstrate. This issue concerning moral capacities is, already, an important source of deception, but there are many others.

Sectors of the AI community are already concerned about the use of psychological terms like "learning," "language processing," or "attention" in the context of machine learning and AI development. The worry is that they are either purely metaphorical or too restrictive because they apply only to a very small portion of what a cognitive process entails—as Melanie Mitchell (2021) says, these are fallacious kinds of *wishful mnemonics* that might be both embellished and inaccurate. In either case, comparing algorithmic solutions with mental processing is not a good analogy, and can be both dangerous and unproductive. For instance, since it is not clear that current AI paradigms can represent any content (Marcus and Davis, 2019), let alone manipulate contents intelligently, the use of psychological terms in AI research should be taken with skepticism and caution. The way current AI pays "attention" may be mere simulacrum, instead of genuine intelligent cognition. A distinction between intensional and extensional AI equivalence is introduced below in order to address this problem, which is discussed throughout the book.

In addition, the central assumption that AI concerns problem-solving rather than problem assessment and selection has been criticized as inadequate for the development of AGI because of the diversity and evolving nature of environments in which problems confront agents (Lehman and Stanley, 2008). In particular, it has been suggested that algorithmic problem generation may be even more important than problem solution (Wang et al., 2019). An evolutionary approach to intelligence and cognition supports this claim. Evolution does not simply solve objectives and problems according to optimized standards (Lehaman and Stanely, 2008). Rather, evolution generates a variety of problems, making some more interesting, important, and difficult than others not for the sake of problem solution but for the sake of problem selection and for learning how to learn the solution to many problems in a flexible fashion. This kind of meta-learning is fundamental in the development

of attention routines, and it is essential for the satisfaction of needs as part of the evolution of species (Haladjian and Montemayor, 2015).

Intelligent beings instinctively know that some problems are more interesting, more fundamental, and more relevant than others. Without this capacity to detect as salient the most important problems, intelligence is impossible. Attentive capacities are essential to navigate this landscape of problem-creation and selection. This capacity is profoundly relevant for epistemic and moral alignments, and for structuring the hierarchy of needs that all humans develop, as argued in Chapter 2. The prospect of "open-ended" AI depends on the development of genuine attention routines, similar to those humans and animals evolved in order to navigate a vast space of continuously growing problems. Alignment needs to be understood in terms of needs, problems, and motivations (just the way we understand life's solutions to multiple problems), rather than a set of optimizations to reach a goal. Meeting a goal may be deceptive by not really solving the problem as it should be solved, in the right context or for the right reasons. All approaches to AI can benefit from a theoretical treatment of how curiosity, sensitivity, and selectivity are all essential ingredients of intelligent meta-learning. The remainder of this introduction provides more details about the present theoretical approach, including concrete proposals, terminology, and how specific arguments are developed in each chapter.

A good portion of AI research focuses on technical and compliance issues, which are also discussed in what follows. While this emphasis on technical matters is natural for AI industrial research, everyone agrees that AI is a lot more than a mere industrial effort. Scaling up, improving, and industrializing AI in accordance with safety measures will require an enormous amount of international research and coordination. This is a colossal and significant undertaking. But once this technology is developed, the most important risks it will generate will not be industrial.[6] No legal, commercial, or industrial standard of safety can measure up to what AI promises to be. It is, therefore, of the utmost importance to start a discussion about the non-industrial risks and features of this new technology, which include the moral and epistemic risks that AI agents will generate.

The uniqueness of AI as an existential threat but also as unprecedented benefactor lies in the notion of *intelligence*. This concept belongs to the set of categories that philosophy studies as part of the theory of knowledge, which include justification, truth, and understanding. There are multiple ways of defining intelligence in terms of these other epistemic notions, but a distinctive feature of the present approach is that it demonstrates why any notion of

intelligence requires *cognitive autonomy*. None of the existential risks or enormous benefits of AI make sense without some degree of autonomy. If AI agents are not autonomous, then they are a kind of automation—a very complex and dangerous tool, but still nothing above and beyond a tool. The fantastic dangers and promises of AI depend entirely on the autonomy of these systems. AI in a specific area of knowledge is enough of a threat if it is genuinely autonomous and intelligent. The biggest threat is AGI. Assuming that AGI becomes equivalent to human intelligence in generality, if genuinely intelligent, then AGI will also be autonomous and much more powerful and rapid. It will quickly become smarter than us, as those who champion the "singularity" or the moment in which humans are quickly left behind have emphasized.

What is cognitive autonomy? The philosophical proposal defended here is that autonomy necessitates at the very least *epistemic agency* for the independent and self-reliant satisfaction of representational and rational needs. Intelligence is fundamentally the autonomous satisfaction of representational and cognitive needs. More generally, an agent is intelligent only if she satisfies her needs because of her abilities. The more needs an agent has (representational, rational, moral) the more intelligent she needs to be.

As already mentioned, the central thesis of the present account of AI is that the best way to understand intelligent autonomy is through attention. The argument for this claim is as follows. In the case of human and animal psychology, attention is the fundamental mental capacity employed in the satisfaction of a wide variety of needs because of its general, selective, and sensitive nature. Attention is a selective capacity that makes salient information that is relevant for the satisfaction of the representational and cognitive needs of an agent, and is insensitive to irrelevant information. Because of this selective and inhibitory capacities, attention is a quite sophisticated kind of mental agency. Since agency is a source of control over the mental and physical actions of an agent, attention is also a basic source of trust, responsibility, and credit. Agency is required for normative interpretations of behavior, such as interpretations of an agent's actions as morally justified, or her inferences as epistemically justified. If AI becomes genuinely intelligent, it will display forms of attention that are equivalent to human and animal attention, as sources of control, trust, and responsibility.

This is a straightforward account, but multiple difficulties emerge in the context of *attentive AI*. A paradox that is addressed throughout this book lies at the core of genuine AI. If genuine AI is developed, it will by definition be autonomous. But then AI is out of *our* control. If AI is completely under our control, then it is a form of automation. Any degree of autonomy generates risks,

and full autonomy is unpredictably dangerous. No autonomy, however, means no intelligence. So AI is either big trouble or no problem at all. How to approach this difficulty? This is certainly not a technical or industrial compliance problem. A thorough philosophical analysis of AI autonomy is needed and that is what this book seeks to offer.

Given that intelligence is an excellence, virtue, or good feature of agents, AI analyses should incorporate what philosophers call "normative issues" regarding justification, moral or epistemic. Legal norms are of course relevant for the normativity of AI and a full chapter is devoted to them below. But legal norms cannot speak to the issue of autonomy or agency because they *assume* that the subjects of the law are agents in the first place, with moral and epistemic capacities. So we need to start with an analysis of these capacities in order to understand how AIs could become moral or epistemic agents and then qualify as subjects of legal protections and responsibilities. Both kinds of normative accounts, moral and epistemic, are developed at length below.

It certainly could be the case that AI and AGI are never developed, because of various kinds of problems. So here is an important message for the generous and curious reader that has explored the first pages of this book. If you believe AI will never become real or genuine AGI and that so-called "AI" will always be only tool-AI, then this book will only concern counterfactual scenarios in which AGI could become a reality. If you are a complete AI skeptic, this book will offer to you a philosophically driven, yet informative way of properly understanding the "science-fiction" of AI. However, if you think that AGI is not on the horizon for the foreseeable future, but that it will or might arrive at some point, you should keep reading this book as an analysis of the risks and benefits involved in the development of AGI in at least three dimensions—epistemic, moral, and legal.

But the odds are good that if you opened this book, you are far from an AI skeptic. In that case, you probably think AI is going to happen anyway and soon, and that philosophy is somewhat peripheral or even fully irrelevant to this development. But I hope to convince you that these normative issues should be taken seriously and that they cannot be tackled simply by scaling up technology, or coming up with increasingly clever forms of deep-learning in compliance with industrial safety standards. The boundary between super-useful tools and genuine intelligence can be porous, but it is drawn at the borderline between autonomous cognitive agents and mere machines.

AGI is intelligence that could eventually compare to human intelligence in complexity and scope. Since autonomy is so important for genuine intelligence,

how should this equivalence be understood? Two key distinctions inform the answer proposed in this book. One of them is the distinction between consciousness and attention (or CAD, which stands for the "consciousness and attention dissociation"). This distinction clarifies which aspects of AI resemble human intelligence and which don't (and cannot resemble it). A crucial consequence of examining AI in light of this distinction is that one can arrive at an understanding of AGI that relies exclusively on attention, rather than consciousness. This is crucial because the nature of consciousness is a notoriously difficult problem and cannot be assumed as a fundamental aspect of AI design. The other critical distinction is between information processing and epistemic agency. Reinforcing the point based on the distinction between consciousness and attention, this book argues that *attention suffices for epistemic agency* and, therefore, AGI with attention capacities should count as genuinely intelligent and autonomous for epistemic purposes.

Based on these distinctions, a further set of classifications follow, which provide more insights for the development of different kinds of AI. The broad classification examined below is between *Intensionally Equivalent Intelligence* (IEI) and *Extensionally Equivalent Intelligence* (EEI). A detailed account of how this distinction helps classify different kinds of AI is offered throughout this book. The critical issue that deserves to be highlighted at the outset is that IEI will involve agency and motivations, rather than brute-force simulation. Two systems may look exactly alike in terms of rates of success, but IEI will be performing tasks based on motivations to attend to stimuli in order to satisfy cognitive needs, while the EEI system would be parametrizing and optimizing information processing without agential guidance. Using the terminology employed above, IEI will be motivated to solve problems that are salient and important, while EEI will optimize on a given problem or sets of problems without having any incentives concerning salience and importance. The main consequence of this distinction is that attention suffices for AGI, but only if AGI is IEI. This is because attention essentially involves motivations in its guidance, selection, and selectivity functions. This kind of motivation-based satisfaction of needs is what is distinctive of autonomous agency.

Of course, EEI-AIs will be enormously significant in our search for AGI, but they will not fully count as intelligent agents. EEI-AI will simulate extremely successfully the human mind, but it will also create risks regarding complete misalignment with human needs and values (for instance, by being prone to deception). Many of the notions discussed in the AI literature, such as "artificial consciousness," "AI attention routines," and "AI self-attention," need to be

reevaluated thoroughly in the light of this distinction because they currently concern EEI, in the most optimistic scenario. The implications of CAD for this distinction are uniquely important for moral and epistemic alignment issues. Since attention is dissociable from consciousness (more about this below), and since attention suffices for epistemic agency, AI can become humanly equivalent epistemic agents, as autonomous IEI attentive agents. However, since *moral agency* requires consciousness in order for it to be genuinely autonomous in humans (and presumably animals), AGI can at best be EEI with respect to moral agency. This has critical consequences for assessments of trust, risk, and responsibility in morally relevant AI systems. However, since legal responsibility is independent from moral responsibility, EEI moral systems might conceivably achieve the status of legally responsible AIs.

A sub-classification of EEI and IEI-AIs is into two styles of *subservient* AI, also epistemic and moral. Subservient AI under our control will be epistemically and morally EEI. Since there is no possibility of autonomous moral AI without consciousness, the only genuine AGI is epistemically equivalent IEI-AI. Yet another classification involves collective artificial intelligence (CAI), which entails four more types of IEI and EEI-AI. A central theme guiding this analysis is the *value* of intelligence in relation to even more cherished achievements, such as knowledge, moral worth, creativity, and understanding.

This introduction provides a synoptic view of the account of AI this book defends. As the title suggests, attention is the key to solve many aspects of value alignment problems with AI. The study of intelligence requires assessments of value and agency, and this is particularly interesting and complicated with respect to AI. But why is the term "humanitarian" relevant for the characterization of AI? The last chapters of this book argue that a human rights framework built around agential capacities and a hierarchy of needs is the best way to guarantee that AI will be universally ethical, as opposed to ethical according to some theory or national legal code. In other words, AI will be ethical to the extent that it benefits humanity as a whole. In this sense, the present account differs from other solutions to the value alignment problem that rely on individual benefit to users, or legal strategies that are applied nation by nation.

In sum, this book provides a philosophical and interdisciplinary investigation of AI, examined within the broader context of the evolution, trajectory, and future of intelligence-capacities, human and non-human. Since intelligence is valued in society and constitutes an essential part of human flourishing, this book also analyzes how intelligence relates to human dignity, human needs, human rights, and human values. Various kinds of values are defined in relation

to the autonomy and capacities of agents. Epistemic, moral, aesthetic, rational, and practical values are outlined and contrasted throughout the book, in humans, animals, and artificial agents. In doing so, the book also engages the main proposals by leaders in the field of AI.

A central thesis defended in the chapters that follow is that attention (as understood in philosophy and psychology) provides the best model for developing genuinely intelligent artificial systems. If so, attention should serve as the paradigmatic case of general and flexible intelligence in the development of AI, and also as the cognitive basis for understanding the relationship between intelligence and rationality beyond human psychology. For any account of intelligence, human, animal, or artificial, attention should take center stage.

Chapter 1 elaborates on the notions of intelligence and artificiality. It provides answers to three questions: (i) what is intelligence in general? (ii) what is the notion of intelligence that has shaped AI research? and (iii) what makes AI artificial as opposed to "natural"? Chapter 2 presents the key proposals that other chapters develop in more detail, about various kinds of intelligence, control, trust, and risk. It defends a framework for explaining the relation between control and trust in terms of an agent's hierarchy of needs. This account of intelligence in terms of agential attention and the satisfaction of various needs is crucial for a proper understanding of the value alignment problems with AI, examined in subsequent chapters.

Chapters 3 to 5 expand upon various aspects of the hierarchy of needs presented in Chapter 2. Chapter 3 focuses on the attentional model of epistemic agency (Fairweather and Montemayor, 2017), and explains why attention, as the fundamental type of epistemic agency, provides the foundation for rational trust in humans and animals, as well as other conceivable intelligent systems such as AGI. Chapter 4 presents a skeptical perspective on the moral capacities of AI, arguing that even "attentive" AI will be emotionally handicapped. The chapter also offers alternatives for how to design ethical AI despite this problem—an idea further developed in Chapter 8. These two chapters rely substantially on the difference between consciousness and attention in humans. Chapter 5 examines human needs that depend essentially on conscious awareness and emotional capacities underlying social intelligence and human flourishing, including aesthetic and spiritual needs, drawing a contrast between the vitality of the experiences associated with the satisfaction of these needs and the automaticity of mere information processing.

Chapters 6 and 7 focus on collective forms of intelligence, control, and rationality. Chapter 6 addresses the question of whether AIs might be essentially

collective agents, similar to our characterization of the "intelligence" of corporations or agencies. Chapter 7 examines the legal, ethical, and political dimensions of AI research and development in an international or collective setting. It investigates the notion of political authority and control in the light of the previous chapters' discussions on autonomy and control, showing that knowledge production and intelligence attribution are essentially related to political authority. Finally, Chapter 8 offers an account of human dignity in the age of AI, based on a conceptual analysis of human rights in terms of human needs, inspired by the capability approach. It demonstrates the importance of understanding the human need for autonomy in terms of rights protected at the international level, which should be the basis of ethical AI.

1

Intelligence and Artificiality

1.1 What Makes AI Intelligent?

This section addresses the question of what is intelligence in general. It discusses relevant historical developments and contemporary approaches to AI. In order to illustrate the importance of satisfying similar needs autonomously through attention, including joint attention, some key aspects of transformers, particularly GPT-3, are critically assessed. The risk of farcical communication with AI is introduced.

What makes *anything* intelligent? We can't have a definition of intelligence that only applies to a single species, or a specific artifact. Intelligence is something we value because of its general applications and potential for problem-solving. These features are independent from concrete "hardware" requirements, although as subsequent chapters argue, some forms of intelligence seem to fundamentally depend on this kind of requirement, such as having biologically evolved emotional needs. But many kinds of intelligence, particularly those related to knowledge and problem-solving, are not dependent upon hardware specifications. Alan Turing (1950) first articulated this view, by saying that machines could be intelligent by computing information, similarly to the human mind.

The broad definition of intelligence as a kind of reliable problem-solving capacity is a good approximation to its essential features. An intelligent system reliably provides a non-trivial result that makes it predictable and trustworthy. It is no accident that clocks and "clockwork" were used as a standard metaphor for mechanical yet intelligent and mindful performance. According to some philosophers, the universe can even have "purposes" if its delicate machinery is understood properly. Cells are intelligent life engines that solve many intricate problems, and the organisms that are integrated by such cells have capacities that

cannot be explained by any of their mechanical parts. It is delicate machinery all the way down and so, "intelligence" all the way down.

This kind of reasoning is notoriously tendentious. Intelligent design is incompatible with the scientific understanding of Darwin's theory of evolution. But this is not germane to the current discussion. The fact that intelligence exists in nature and that its purpose is to solve problems is not disputed by anyone. What is highly contentious is *who counts as intelligent*. The mechanical view of the world has plenty of room for options. First, the mechanical view of the world is not the same as the materialist view of the world. One can hold that the mind is a kind of matter and explain its existence in those terms, as a material rather than mechanical or causal and deterministic phenomenon. For materialists about the mind, the challenge is to select a kind of matter suitable for intelligent thinking. According to most materialists, given the evidence from neuroscience, minds are identical to brains in the sense that whatever a mind does, it needs to be understood in terms of neural activations. The brain is also a set of mechanisms, but what the mind *is* identical to on these views is the physicochemical structure of the neural networks that constitute the brain. If only agents with a mind can be intelligent, then only beings with brains can be intelligent—this includes many animals, but it excludes plants from the "intelligence world." It certainly excludes non-biological machines as well.

One can also hold a mechanical view of the universe and a non-physicalist view about the mind. René Descartes' metaphysical account is the most famous articulation of this doctrine, called "dualism." Descartes defended a mechanical view of the physical world according to which the universe is a causally deterministic structure with functions and laws that govern all causal interactions, and which can be formalized mathematically. Descartes, however, denied that the mind was physical or in any way dependent on the mechanical "clockwork" that constitutes the physical world. Since the mind cannot emerge, supervene, or depend on the physical world, Descartes proposed that it is a separate primitive feature of the universe—a "thinking substance." The notorious consequence of this line of thought is that only humans count as intelligent—*all* animals are nothing but mere machines.

Computers changed this story dramatically. They are machines, but as functionalists point out, they are not strictly material. Computers cannot be specified merely in terms of the mechanical hardware that instantiates them. If the mind is a computer, it need not depend on any specific material arrangement. It can be "free" from neural constraints and, in principle, as Turing pointed out, machines could definitely have minds if they are informationally equivalent to

us in computational and informational terms. The mechanical view of the mind became extremely intricate and also more plausible with the development of computer science. In combination with research on neural networks, the current AI paradigm that combines computer science with insights from neuroscience is particularly powerful. One must, however, still define what is intelligence, independently of what kind of *thing* can be intelligent.

These traditional debates in philosophy of mind, which focused on what philosophers call the "metaphysics" of the mind, are useful in one key way: drawing the line between machine and intelligent agents depends on a constraint concerning information processing. For the materialist, the constraint is anatomical—only information processing that is instantiated in neural networks (or "within the skull") is mental and genuinely intelligent, rather than purely biomechanical. For the dualist, the constraint is subjective and anti-physicalist—only non-mechanical or non-physical information by a "mental substance" counts as intelligent. Finally, for the functionalist, the constraint is computational—only information that is algorithmically structured counts as intelligent. Thus, the intelligent mind depends on a certain kind of informational activity in brain regions, in thinking, or in algorithmic functioning. This is very slippery territory but the essential point for the present discussion is that on all views about the mind, intelligence is a kind of mental and informational *activity*, even when understood mechanically.

The problem is that activity is pervasive in the biological and physical non-carbon-based world. For the constraint on information to be explanatory, the divide between intelligent agents and mere machines must be based on a much more precise definition of mental activity. The essential characteristic of an intelligent mind, which makes it unlike any other kind of mechanical and informational phenomenon, is that intelligent minds are autonomous. This notion of autonomy is explained in detail below. For now, what matters is that intelligent agents are self-reliant. Broadly speaking, intelligent agents *as such* provide the best explanation of how they solve their problems through their capacities. One cannot reduce intelligent agency to a mere "causal chain" of events. Genuine intelligence, unlike complex mechanical design, is essentially self-dependent.

The focus here, as mentioned above, is on intelligent behavior rather than the metaphysics of free will. Biological organisms are very interesting in this regard. Plants and all living organisms are autonomous in a minimal sense, because they are self-sufficient in maintaining themselves alive. But they are not autonomous with respect to having a general-purpose intelligence (more about

this below). Briefly, plants satisfy their biological needs autonomously but seem to lack cognitive needs that generally intelligent animals have (although this is controversial). Plants are an interesting test for definitions of AI because they count as intelligent on almost all the definitions found in the literature. Consider what Legg and Hutter (2007) define as the "essence of intelligence in its most general form" in their comprehensive and influential paper, which examines a vast variety of definitions: "Intelligence measures an agent's ability to achieve goals in a wide range of environments." Tree ferns have been able to thrive in various challenging environments for a much longer period than humans, and this is true of many plants. Should they count as intelligent? If so, how does this help distinguish AI from mere machine or tool-AI, or genuine intelligence from very complex behavior?

Legg and Hutter's solution to the difficulty presented by the polysemy of "intelligence" is to provide a *formal* characterization of intelligence: a mathematical formulation of the computable aspects of intelligence that is general and universal (non-subjective or anthropocentric). This is a very valuable contribution to our understanding of intelligence, and nothing in this book rejects or challenges this approach. In particular, by relying on CAD, the present proposal elucidates why non-subjectivity or anthropocentricity is best understood as the non-necessity of *phenomenal consciousness* for intelligence. Thus, a formal and universal characterization of intelligence is compatible with a definition of autonomous intelligence in terms of attention, as mental action. The key difference is that while this formalization might be necessary to define universal intelligence, it is insufficient to define genuine general intelligence. A measure or method is indeed needed to delineate agents from their environments. But agents must be defined not merely in terms of causal and informational interactions (formally complex as these might be). They must be characterized in terms of their self-reliance in various tasks, and their agency as the explanation of their success in achieving these tasks. Agential autonomy cannot be simply a set of formal relations or even a mere set of causal relations. Agents are autonomous because their intelligence, as mental activity, helps them select information and meet goals according to *their own* needs and motivations.

This is why attention is extremely important for the definition of intelligent agency. Since agents are self-reliant, they are naturally motivated or inclined to satisfy their goals based on their attentional skills. A formal approach needs to be complemented with a model of an autonomous agent with real independence in real environments, which fundamentally depends on a model of the agent's goals and motivations in relation to her representational needs given the

obstacles these environments present. So although their formal approach is of great help in narrowing the definition of intelligence (as Legg and Hutter say in their paper no one seems to know exactly what intelligence is and there is a wide divergence of opinion) a robust and philosophically informed account of agency is still needed.

Here is another way of making this point. A full explanation of intelligent agents requires a description of them in terms of detailed scientific and formal mathematical models concerning algorithmic complexity, information processing, environmentally dependent functions and dynamics, and biologically informed applications. But it should also include an account of how these descriptions relate to *normative* aspects of intelligent agency, such as epistemic responsibility, justification, knowledge, and trust. These normative aspects of a general theory of intelligence that includes AI are the main concern of this book.

There is another difficulty with strictly descriptive and formal models, which is that the "hardware" or computational basis of an intelligent agent matters for some kinds of intelligence in fundamental and surprising ways. Emotional intelligence, in particular, radically depends on the satisfaction of biologically instantiated needs, thereby providing the basis for a variety of moral and aesthetic capacities. Emotions depend fundamentally on biological signals (Damasio, 1994) and in this respect, biological "machines" have a unique kind of intelligence, since feelings and other emotional contents depend on our biology and evolution. Chapters 4 and 5 show why CAD entails that AI will not have completely equivalent kinds of emotional intelligence for this reason.

Thus, strictly technical issues are not sufficient to explain the main difficulties this book addresses concerning the normative aspects of agency and responsibility that AI would need to satisfy in order to be genuinely intelligent and trustworthy. This does not mean, of course, that the dramatically efficient technical developments in the field of recent AI research are not relevant. To the contrary, as mentioned, they are necessary to shed light on how artificial systems might become intelligent. More important, technological and theoretical improvements are producing unexpected and promising results at an alarmingly accelerated pace. A recent example is GPT-3, with its formidable capacity to respond to questions, in many circumstances in a contextually relevant way (although it still fails in some shockingly inadequate ways; see Marcus and Davis, 2019).

Using GPT-3 as an example, there are multiple innovative features of this system that could make "scaling up" strategies extremely productive. Although

GPT-3 is not a recursive system, it parses selectively portions of text in order to contextualize them conversationally, based on a massive database of human text. The inputs it takes at a time can include large portions of text and the decision it makes in response is very "attention-like" (but see the caveat in the introduction about how this term is used in machine learning). Whether these are genuine representations or merely causal and functional approximations of genuine representations is a subject for debate. But even if one grants that these are genuine representations—the kind of informational structure that is about a feature of the environment and only about that feature, in a way that it can misrepresent this feature and still be *about it*—there are solid reasons to deny that what these systems are doing is paying attention in a genuinely agential and intelligent way. Chief among these reasons is the lack of cognitive motivations based on autonomous needs that only agents have, and which they satisfy through attention. It is also important to point out that only agents seem to represent the environment in a way that their cognitive needs are satisfied—*you* see an apple, not your visual cortex or your retina. This is very important to properly understand both representation and attention in intelligent agents.

That said, GPT-3 is certainly impressive, and interacting with it is similar to having an online conversation (with some troublesome exceptions). If it could become genuinely attentive, it might even be the very beginning of a general kind of AI system. It seems that GPT-3 can learn various tasks, and is certainly not limited by the type of language or even the code it takes as input. The way it selectively performs its tasks resembles and is actually modeled after attention routines. The paper that pioneered transformer systems like GPT-3 described "attention" as the main innovation of these systems, which are clearly inspired by biologically informed models (Vaswani et al., 2017). But as mentioned, and as is argued below, these are simulations of attention routines that lack the content and selection that genuine attention provides. Mitchell's (2021) warning concerning fallacious reasoning in AI research is very relevant here. In particular, the fallacy of "wishful mnemonics," which consists in attributing psychological terms to advanced computing systems.

Convolutional neural networks with improved models for short- and long-term memory can definitely improve the performance of this kind of system at various tasks, such as object and feature recognition, as well as the development of hypotheses and meta-hypotheses concerning searches. Combining various strategies and architectures promises even more impressive results, particularly the prospect of unsupervised learning and "open" AI. The essential difficulty confronting all these efforts concerns the autonomy, meaningfulness, and

control with which these processes are performed—this is what has critical consequences for intelligence in epistemology and ethics.

Since intelligence is a highly selective capacity under the control of an agent, it is natural that the field of machine learning is already incorporating attention as a fundamental model of intelligence, based on findings in psychology and neuroscience, with enormous promise for cross-pollination in AI development (Lindsay, 2020). In some cases, the results may almost exactly resemble those produced by human or animal cognition, although as will be shown, so far they are never exactly the same in all cases or respects. But the key point is that even if the results were exactly the same, these systems would fall under the category of systems characterized in what follows as Extensionally Equivalent attention. The reason for this is that, in a very deep sense, these systems are not really autonomous, they lack representational needs, they don't really have representations with systematic and stable contents, and they are still under our control because we give them the problems they must solve based on our own motivations and needs. The fact that we are not exactly sure about what is it that they are doing when they solve problems (the opacity problem in AI) makes things even worse.

However, despite of these problems, the fact that current AI researchers are finding inspiration in the selective functions of attention in order to design intelligent systems is certainly a step in the right direction. Yet, while incorporating attention into AI research is fundamental, it shouldn't properly be called "attention" until agency and autonomy are at the basis of attentive performances, as is the case with animals and humans. Extensional equivalence is good enough for many purposes (although we are far from achieving it), but it is not autonomous until attention routines are integrated in order to satisfy the specific motivations and needs of an independent agent. Only then will AI become IEI and eventually AGI, with all the benefits and risks that are explained in what follows. An initial warning, already voiced above, is to avoid using terms from psychology as equivalent in AI solely on the basis that AI produces similar results or because it is analogous to biological processes. Correlation of results is not the same as similar representational, conceptual, and causal processing, and even if similar results also include similarly simulated neural causal processes, that would still not suffice for intelligence because such processing has to be integrated with the motivations, goals, and needs of an agent—a condition that is fundamental for epistemic agency and trust.

Let us consider GPT-3 as an example again in order to illustrate the issue of trust. Speakers of a language or agents that communicate, such as animals

looking for sustenance, have cognitive needs that they must satisfy for themselves autonomously, and also as a group. They need to rely on each other and be accountable to each other as they jointly act and *jointly attend* to the environment. This is what makes their utterances and communicational resources meaningful and purposeful. Crucially, their similar representational needs give homogeneity to what they value in their communication and cognition. Herein lies the essential problem with systems like GPT-3 (Montemayor, 2021). Their lack of motivational and agential intelligence makes it impossible to interact with them in a jointly purposeful way, even if they are delivering answers to our questions that match very well what a speaker should say. It should be emphasized that this is based on text we produced, found on the internet, according to our own representational needs. This lack of common ground or common purpose generates the risk of manipulative and farcical communication with systems like GPT-3. A common thread throughout this book is that because of risks like this, we cannot really trust AI unless they have similar attention routines, for the purpose of satisfying similar needs and goals.

Technological innovations in AI development like those briefly described here will continue to surprise us. They provide a powerful and refreshing update on the traditional philosophical debates on AI that inspired connectionist and symbolic architectures. Back then the main debate was about whether neural networks, with their flexible and biologically inspired architecture, could provide the kind of compositional representations that symbolic systems were designed to satisfy. Connectionist or "associationist" networks seemed incapable of generating thought-like structured representations and were therefore deemed unsatisfactory as a model of the mind or intelligence. But computer power has changed radically since then. Scaling up and embedding neural networks in probabilistic learning models with various parameters for optimization has produced results that no symbolic system or hierarchical and explicitly "representational" computer is remotely capable of. The paradigm has shifted. The focus now is on the learning process, and on how it generalizes and adapts to larger bodies of data. This is why attention has become so important for current AI design. Precisely because of this, it is fundamental to not only have a good grasp of the descriptive level at which attention operates, but also the normative level in which agents trust each other and hold each other responsible for what they attend to. In this respect, insights from the symbolic and representational approach are needed, as is implemented now in so-called "hybrid models."

Neuroscientists, psychologists, philosophers, and computer scientists have defined different aspects of intelligence, in humans and animals. But only

some of these aspects can be implemented in AI. Some features of human intelligence may be successfully reproduced in AI, albeit with some important qualifications, but other aspects of human intelligence cannot be so reproduced. In fact, this book argues that there are risks created by these limitations that cannot be avoided at all, and that others which might be avoided may require an international legal framework for designing AI—itself a major complication. There is a long history of models, developments, and contributions from mathematics, logic, epistemology, and psychology to the body of knowledge that eventually coalesced into what we now call "artificial intelligence." This history could be told in several volumes, and portions of it are taught as cannon in departments of mathematics, philosophy, psychology, and computer science. This debate is part of a much broader history of theoretical frameworks, which concern the classification of our cognitive capacities into rational, intelligent, emotional, somatosensory, reflexive or motoric, conscious, or unconscious. The disciplines that contribute to this investigation now include sociology, anthropology, history, economics, and many of the humanities. Insights from all these disciplines, particularly with respect to intelligence as an autonomous source of responsibility, should inform our theories of AGI.

1.2 Intelligence in AI Development

This section discusses a definition of intelligence that is broadly assumed in AI research, using Stuart Russell's work as the key point of reference. It highlights the advantages of Russell's definition, and it also problematizes the notion of autonomous intelligence through the distinction between consciousness and attention. The AI risk of reliable execution without proper attentive integration is introduced. The notion that agents are sources of risk reduction or elimination is explained, and a definition of intelligence based on autonomy is provided.

The previous discussion centered on the notion of intelligence in general, with applications to recent trends in AI design. A significant and different question is what is the notion of intelligence currently assumed in AI development? Here one has no other option than to select a definition from an authoritative source that has a broad level of acceptance by the AI community. It seems clear, given many definitions and tests for intelligence in this field, that the notion of intelligence that became prevalent in AI design is based on conditions for successful action. As Stuart Russell (2019, 9) says in his important book on AI,

intelligence is related to *achievement*: "After more than two thousand years of self-examination, we have arrived at a characterization of intelligence that can be boiled down to this: *Humans are intelligent to the extent that our actions can be expected to achieve our objectives … Machines are intelligent to the extent that their actions can be expected to achieve their objectives*)."

There are several ingredients in this definition that in philosophy, epistemology in particular, are associated with a pragmatic conception of knowledge and intelligence. The key components are: motivations to act, conditions that must obtain for the action that satisfies these motivations to succeed, and a reliable rate of success that guarantees that motivations will reliably lead to the satisfaction of goals. This is the basic structure of a "success semantics," proposed originally by Frank Ramsey, which can be used to model reliable epistemic virtue and agency. This definition is fully compatible with an attention-based model of intelligence, in which attention provides autonomous agency that responsibly satisfies a multiplicity of needs (Fairweather and Montemayor, 2017).

Normative guidance or rule-following is absent from this definition of intelligence. This is noteworthy for two reasons. First, for many philosophers, what makes a pattern of thought or an inference not only trustworthy but genuinely justified is the fact that such a pattern of thought is guided by norms of rationality. The problem is that the notion of following a rule in our own minds, and justifying such a rule with other rules, leads to all sorts of problems and in particular, circularity, as explained in Chapter 3. Second, and related to the previous point, it seems that part of what makes our cognitive capacities for knowledge and intelligence trustworthy is the kind of access we have to our thoughts and their semantic contents. But it is hard to assume this definition of intelligence without falling prey to anthropocentrism and philosophical conundrums regarding conscious awareness. These are two crucial reasons why the pragmatist definition of intelligence used by Russell can be implemented in empirically testable contexts (including industry)—because it appeals to concrete actions and their successful consequences, rather than abstract norms or conscious awareness and the rules it somehow "follows."

In philosophy, one of the main debates surrounding this kind of consequentialist and pragmatist conception of intelligence concerns the lack of normative conditions for the evaluation of our cognitive access to contents, which either directly determines their normative status or specifies how to determine their normative status. Here one confronts problems about philosophical methodology. To illustrate this difficulty, one can argue that epistemic justification and rational guidance depend entirely on the qualitative

character of conscious experience and our self-aware introspective capacities (Smithies, 2019). But then one needs to solve the problem of how to explain success rates, which are basic for the definition of intelligence, exclusively on the grounds of having conscious awareness. This appeal to pure conscious awareness is, of course, inadequate because one must appeal to capacities or skills at some point in order to explain achievement and success rates, and as soon as one does this, these capacities or skills seem to be doing all the explanatory work. This is why Alan Turing (1950) was firmly opposed to any definition of intelligence or rationality based on conscious awareness.

Another traditional strategy in epistemology is to make epistemic justification and rationality entirely dependent on rule following—on following rules concerning evidence updating, probability, and logic. This influential account, however, requires a robust kind of cognitive access to those rules and besides the problem of circularity, one also confronts the problem of what exactly this cognitive access amounts to. Moreover, an immediate problem with both approaches is that one can distinguish access from phenomenal consciousness, or what it is like for a conscious subject to have subjective experiences (Block, 1995b), partly because phenomenal consciousness likely evolved separately from attention (Haladjian and Montemayor, 2015; Montemayor and Haladjian, 2015). The consequence of this distinction is, as mentioned, quite substantial because, unlike phenomenal consciousness, attention can be studied and measured scientifically. Moreover, access consciousness can be understood in terms of attention (Stoljar, 2019), so phenomenally conscious access seems unnecessary to define intelligence even from a purely theoretical perspective. This issue is expanded upon throughout the book, but will become particularly relevant in Chapters 4 and 5, where the importance of phenomenal consciousness for some kinds of intelligence is clarified.

This is why reliable motivations are so important for the definition of intelligence. An intelligent epistemic agent is motivated toward satisfying relevant goals, not just as an optimization process but because she has specific needs, and she is also reliable in satisfying these goals. As mentioned, one of the most fundamental needs of intelligent agents is to identify good problems to solve and to be curious about—a need that cannot be satisfied by simply optimizing on a fixed goal. Take away the reliability of motivations, and the agent will no longer succeed, and therefore, will no longer qualify as intelligent. Take away the relevance of the goals the agent is motivated in pursuing and satisfying, and her behavior becomes erratic, random, heteronomous, and basically unintelligent. This is the basic combination of motivation, agency, goals, and conditions for the satisfaction

of goals that Ramsey thought was sufficient for defining knowledge, and for providing a semantics for the meaning of expressions and their truth conditions (see Fairweather and Montemayor, 2017, particularly Chapters 3 and 5).

Crucially, even if an agent's motivations are implicit, unconscious or not "phenomenally conscious," and even if norms are also only implicitly or unconsciously followed, motivations can still guide the agent toward success, as long as they are both *reliable* and informationally *integrated* with the agent's cognitive capacities, goals, and actions. Both reliability and cognitive integration can be objectively identified or measured, and this is fundamental for testing intelligence or any other capacity—success is not accidental or externally imposed, but achieved in virtue of the agent's abilities, and this makes it valuable and also normatively relevant: reliable and well-integrated capacities are epistemically good, and since the agent is non-luckily meeting her goals on the basis of these capacities, she is responsible for their consequences. An unreliable agent will not succeed in satisfying her goals; a poorly integrated one will succeed at some tasks, but fail at others due to lack of integration regarding meta-goals and preferences, which are equally fundamental for satisfying goals intelligently.

Virtues of integration are not as prominent in epistemology as virtues concerning truth-conduciveness—reliability, evidence gathering, optimal decision-making—but they are essential for drawing the normatively crucial difference between successfully fulfilling a task and fulfilling it in a responsible and reasonable way. In fact, too much reliability without integration and sensitivity to relevant information spells disaster. An AI may be more reliable than any human in satisfying the goals it is given at a point in time: get me coffee now; get me to my destination as soon as possible. But if it fulfills these tasks without regard to salient relevant information about common sense and contextually determined preferences, this could lead to dangerous situations: crashing into a coffee shop or throwing a coffee at you; speeding up in areas where an accident can easily be caused. If the reliable satisfaction of these goals is not performed in an *attentive* manner, then the system is *not trustworthy*. Inattentive reliability satisfies goals, but at the risk of preventing the satisfaction of other, more important goals, such as staying alive or keeping a safe course toward one's destination. *This is AI risk based on lack of attentional integration, not lack of reliability*.

Various AI risks will be explored in what follows. This kind of risk, however, is central to an explanation of the limitations of AI with respect to the definition of intelligence that is prevalent in the field of AI research. It certainly generates

industrial and legal risks, but in an entirely new way, which blurs the boundaries of responsibility because it generates the illusion of autonomy. For now, however, our focus is on epistemic issues. The key point about virtues of integration is that agency is fundamental for intelligence because autonomy with respect to which goals are worth pursuing is fundamental for intelligence. Epistemic reliabilism, the view that beliefs produced by a reliable process are justified because they are more likely to be true than false—a view that can also apply to knowledge—is applicable to the broad definition of intelligence as optimal problem-solving, but not to the autonomously agential one, which is the one needed for genuine AI. This is another advantage of Russell's definition of intelligence.

Mere reliability is insufficient (although it is necessary) for intelligence. Phenomenal consciousness is unverifiable and too anthropocentric. The present proposal is that epistemic agency is best understood in terms of attention, which need not be phenomenally conscious. Thus, neither phenomenal consciousness nor reliable processes by themselves are sufficient for intelligence. Attention is necessary for intelligence, because of its selective and agential functions, and it is also sufficient because it reliably guides an agent toward the successful satisfaction of her representational and cognitive needs. It is also sufficient for normative evaluations, understood as epistemic virtue—a properly attentive agent is a good source of information and a good epistemic agent in general.

Attention, from basic perceptual tasks to highly integrative and sophisticated inferential reasoning, always provides selectivity for relevant information, as well as robust kinds of cognitive integration with motivations and cognitive needs. Attention is the basis for virtuous sensitivity to salient information and of virtuous insensitivity to irrelevant information (Fairweather and Montemayor, 2017). Phenomenology alone cannot play this role. Moreover, attention provides a paradigmatic kind of agential control (Wu, 2011, 2013, 2014). But as the next section shows, reliability also matters enormously. Crucially, reliability matters because it is the basis of *epistemic trust*—a necessary basis, albeit insufficient for full epistemic trust and responsibility, which require agency. An attention-based approach encompasses intelligence in all its complexity because all the relevant forms of intelligence are types of attentional guidance. An attention-based approach also provides a thorough understanding of epistemic agents, because attention is an exemplary form of mental action.

To fully capture the complexity of human intelligence, however, phenomenal consciousness must enter the picture. The second part of this book, starting with Chapter 4, explains why the intricacies of and tensions between moral and epistemic guidance can be understood in terms of specific kinds of

attention: phenomenally conscious, unconscious, or access conscious. A central task of the first part of this book is to show that epistemic agency in general (including AI) does not necessitate phenomenal consciousness, based on arguments concerning CAD, attention, and agency. Reliable agency is what is needed to eliminate epistemic risks. Agency, integrated by "attentional-like" capacities (see the caveat above about using psychological terms in AI research), is also central to our contemporary understanding of AI. Russell writes:

> The central concept of modern AI is the *intelligent agent*—something that perceives and acts. The agent is a process occurring over time, in the sense that a stream of perceptual inputs is converted into a stream of actions.
>
> (Russell, 2019, 42; his emphasis)

The process Russell describes is what psychologists define as *perceptual attention routines*. The intelligent agent is an attentive agent. Russell offers as an example a self-driving taxi, emphasizing the high degree of informational precision and selectivity that is required to process a gigantic amount of data in real time. In a parenthetical remark, Russell notes: "For an experienced human driver, most of this maelstrom of activity is unconscious: you may be aware only of making decisions such as 'overtake this slow truck' or 'stop for gas,' but your eyes, brain, nerves, and muscles are still doing all the other stuff" (Russell, 2019, 43).

Indeed, our brains rely heavily on unconscious attention to perform complex tasks. These processes can be guided in a consistent and systematic way, as if "following a formal rule," without explicit knowledge on the part of the agent. What really matters are the abilities the agent has, allowing her to succeed at tasks because of the abilities she possesses, rather than based on strictly external factors, abstract rules, or mere causal chains of events. The autonomy of agents, based on their abilities, thereby reduces the two types of epistemic risks mentioned before: epistemic risk concerning unreliability in action and epistemic risk based on lack of cognitive control and integration. Autonomy provides a guarantee that the agent is a *source of risk reduction*. Crucially, this kind of agential luck reduction does not necessitate phenomenal consciousness. This point has been made before by Turing, and, in fact, Russell (2019, 16) asserts that phenomenal consciousness makes no difference to AI research and cannot be informative in any way. Chapters 4 and 5 argue that phenomenal consciousness is necessary for integrative cognitive roles concerning the experience of familiarity and empathy that underlies moral and aesthetic reasoning. But, in agreement with what Russell indicates, it is not necessary for most kinds of epistemic agency.

Phenomenal consciousness, by itself, cannot be the sole guarantor of epistemic risk elimination. Attentional abilities are required for this. Thus, it is important to emphasize that it is attention, even if it is of the unconscious variety, that is at work in epistemic risk reduction, because all these attention subroutines that are highly sensitive and operational while you are driving, for example, are essential parts of your overall epistemic agency. Unlike your eye-lid movements or your digestion, unconscious attention routines are integrated in a way that permits *agential guidance*—they all collaborate to make you, as an agent, a reliable source of risk and luck-reduction as you drive. You are responsible for these "unconscious choices," because they are an essential component of your guided actions. This sounds initially counterintuitive and perplexing, but consider that this is the only feasible strategy for a cognitive system to quickly filter, select, and process information in real time, in an environmentally contextualized way. This is why the unconscious routines you use to, say drive a car, are part of your overall plan of getting to your destination. The portions you are aware of are only a small part of the general and unified action. As subsequent chapters argue, what is most salient to you are the needs at the top of your priority list, which are the needs you must satisfy given a specific goal in a particular context or environment.

Russell mentions some of the challenges agents need to solve at any point in time based on their informational design, or in the terminology I shall use, their *integrated attention routines*. For AI systems, Russell lists the following typical problems agents must confront: whether the environment is fully observable or partially observable; whether the environment and actions are discrete or effectively continuous; whether the environment contains other agents or not; whether the outcomes of actions, as specified by the "rules" or "physics" of the environment, are predictable or unpredictable, and whether those rules are known or unknown; whether the environment is dynamically changing, whether the time to make decisions is tightly constrained or not; and the length of the horizon over which decision quality is measured according to the objective (Russell, 2019, 44).

Animals are always embedded in dynamic and constantly changing environments that require them to keep track of multiple needs at different timescales (Montemayor, 2013). Our capacities for attention are the result of millennia of evolution and interaction with various types of physical and social contexts. But human epistemic agency is the most fluid and general kind of intelligence, to a large extent because of increased capacities for memory and communication through language. It is no accident that Turing focused on

conversational exchanges as a test ground for human-like intelligence. What Turing (1950) calls "the most extreme" form of the solipsistic argument that prevents intelligence from being scientifically studied is at the basis of the contemporary definition of phenomenal consciousness, as "what it is like" to be a conscious organism (Nagel, 1974), which also underlies the "hard problem" of consciousness—or why any function, structure, or material arrangement should be aware of a particular experience from a subjective point of view (Chalmers, 1995). Phenomenal consciousness might be characterized as "solipsistic," but one still needs a positive account of intelligence and knowledge that dispenses with it in order to show that consciousness is indeed not necessary for epistemic agency—Turing's rejoinder may not succeed otherwise. The approach defended in subsequent chapters is that phenomenal consciousness is not necessary for epistemic agency because attention underlies all forms of epistemic agency, even in the absence of phenomenology. The essential point is that phenomenal consciousness *by itself* is too private to provide the kind of epistemic action involved in reasoning and knowing. Attention is necessary precisely because it provides the most robust and reliable kind of mental activity.

Computers and contemporary AI are universal in their application, which means that one can in principle solve many different problems by using computer power. But what universal machines and general-purpose AI lack are motivations, genuine preferences or needs, and fundamentally, an articulation of needs, in conjunction with solutions to problems concerning how to satisfy these needs through attention routines. They do not reduce risk on the basis of their integrated agency (goals, needs, and attentional means to satisfy them). Thus, calling contemporary AI systems "agents" is also more metaphor than reality. As will be explained in subsequent chapters, the most important aspects of human rationality all depend on attention routines (e.g., Kahneman systems 1 and 2, and the experiencing and remembering self). Each of these cognitive modules or components provides a kind of agency through integrated attention routines (Montemayor, 2019a).

For instance, the fast system of reasoning based on heuristics and biases (Kahneman's "system 1") includes highly skilled forms of unconscious attention routines that allow us to navigate the complexities of the environment without effort. Other forms of attention require explicit guidance and effortful control but they also need to operate with the aid and guidance of unconscious attention routines. Some systems of attention are phenomenally conscious in essence, while others are just access conscious, or even encapsulated (Montemayor and Haladjian, 2015).

The second part of the book argues that the most important distinction concerning all types of attentive "selves" is between the *empathic and the epistemic self*. These two broad categorizations of attentional agency yield two different approaches for how to structure preferences and values, or two ways of defining intelligence as epistemic or emotional. This difference in value assignments generates major difficulties for any solution to the "value alignment" problem with AI and, as mentioned in the introduction, among humans as well. Hence the importance of describing agency in terms of roughly homogenous needs and capabilities, which provide both enough basis for consensus and enough room for a wide variety of intelligent-divergence.

Autonomy based on the risk or luck-reducing aspects of agency is a necessary condition for genuine intelligence because this is how a great variety of problem-solving gets reliably contextualized and integrated hierarchically, through the selective functions of attention. Intelligence in humans and animals requires not only rational or optimal problem-solving, but fundamentally, self-reliance in doing so, in a way that there is a salient priority for creativity and meta-learning. Intelligence depends, therefore, on the autonomous satisfaction of needs based on the attentive skills of agents, chief among them, the capacity to identify what problems are worth addressing and which goals are valuable.

1.3 What Makes AI Artificial?

This section addresses the question of what exactly makes AI "artificial." The role of phenomenal consciousness in human intelligence is clarified. Issues in transhumanism and posthumanism, including examples from science fiction, are briefly discussed. The notions of moral and epistemic status are introduced. An influential classification of empathy is shown to correlate with the distinctions defended in this book, specifically those related to CAD. The implications of these distinctions for "artificiality" are discussed.

The two questions addressed thus far concern the notion of intelligence, in general, and the specific definition of intelligence that has been largely adopted in AI development, in particular. Both show that attention routines are fundamental. The question is when are these informational routines genuinely attentive? The answer is that these attention routines must be the abilities of an agent that autonomously satisfies her needs because of these abilities, and who is motivated to do so by integrating them with her hierarchy of needs, goals, and

plans. This type of intelligent agent genuinely pays attention, and is responsible for the contents, scope, and guidance of her attention routines (and subroutines), which determine *how* she satisfies her goals.

Already within the biological world we encountered a demarcation between agents that pay attention and living organisms that seem to be autonomous only in a metabolic sense. Humans and animals fundamentally depend on attention routines to navigate their physical and social worlds. Plants, by contrast, sustain themselves by being "metabolically intelligent." But they seem to lack genuine motivations, goals, or attention routines. Thus, although this delineation is not entirely uncontroversial, it shows that there is at least a plausible way of demarcating biological or *natural* intelligence from *natural "machinery."*

The question now is what makes AI *artificial*? Since there is natural machinery (cells or plants) this is not a trivial question. AI could become genuinely intelligent by becoming autonomous and attentive. Why shouldn't it count as "natural" then? If by "artificial" we simply mean "unnatural" it seems arbitrary to deny the status of natural agents to AGI—their intelligence is not *against* nature, but in *accordance* with natural principles. Many deep-learning systems are designed based on explicitly neural and biological principles, including attention-like routines. For AI to qualify as artificial, something about our biology must be unique and irreproducible in machines. Only in that case will there be a concrete feature of machines that make them "unnatural."

The proposal that subsequent chapters defend and articulate is that AI is artificial because it cannot be phenomenally conscious. Phenomenal consciousness is deeply rooted in our biology, and this is the key to understand why AI is artificial. Interestingly, according to CAD, since phenomenal consciousness is dissociable from attention, and since cross-modal integrated attention provides access to information that can be used for thought, action, and decision-making, CAD *entails* the important consequence that AI can become *access conscious*—it could become intelligent for all epistemic purposes concerning thought, inference, action, and decision-making, but it will lack the viscerally engaging nature of subjective experience, and therefore, it will be deprived of the basis for human and animal-like emotional intelligence.

An intriguing possibility is that by combining natural and artificial machines, the hybrid offspring of intelligent systems will be capable of being phenomenally conscious and, in fact, go beyond all our current capacities for intelligence, creating not only new forms of "unemotional" AI but also biologically rooted and much more visceral AI (or at least, "cyborg" systems) capable of enjoying a much wider set of experiences and emotions than humans and animals

are currently capable of having. Transhumanism claims indeed to be a kind of ethical and political liberation. It presents a scenario in which human intelligence is completely "dethroned" becoming just a "speck of dust" in the landscape of possible minds. Such de-anthropocentrism allows for open ethical relations across species and kinds of intelligences, or a relational stance in which human mentality is no longer the sole source of epistemic and moral value—a final displacement of humans, who will no longer be the privileged metaphysical center of the intelligence universe (Haraway, 2004). But even in such a radically different landscape of intelligence the issue remains the same—as long as there are intelligent beings with a visceral biologically rooted intelligence, they will differ from those who lack such a biological foundation in very important respects.

Suppose, however, that our ancestral biological foundation is replaced through genetic manipulation. A cyborg liberation that breaks with any rigid ontology privileging human intelligence, including emotional intelligence, might well be a consequence of developing AI in combination with transhumanistic genetic research. But some caution is needed in pursuing this goal. Posthuman values can become problematic. A revolt against our natural "prison" or the "tyranny of mother nature" (More and More, 2013) can produce the most egomaniacal and selfish kind of human tyranny by expanding dramatically the lives of ultra-selfish humans. Moreover, such a fight against mother nature may make our intelligent machines inhuman by design because emotional intelligence, rooted in evolution, would be lost. Just think about what happens to value alignment if our carefully tuned biology is considered a prison? For transhumanism or posthumanism to be human (or *humanitarian*), it must somehow accommodate our emotional needs.

The key issue now is, if phenomenal consciousness is not necessary for epistemic agency because attention can play this role, then what is the contribution of phenomenal consciousness to human intelligence? If phenomenal consciousness cannot be "implemented" in AI, then what will AGI miss that human intelligence has by virtue of being phenomenally conscious?

First, CAD does not entail that phenomenal consciousness and attention are "divorced" in human psychology. To the contrary, human psychology depends on the deep connections between phenomenal consciousness and attention, and some kinds of attention to emotions are necessarily phenomenally conscious. The degree to which phenomenal consciousness is dissociable from attention is a subject of debate, with most authors holding the view that they are dissociable to a very large extent (attention can occur without phenomenal consciousness in many cases), and even doubly dissociable (phenomenal consciousness may

occur without attention), although this claim is more controversial (see Montemayor and Haladjian, 2015, for a review of the literature in philosophy and psychology). But human psychology fundamentally depends on how attention integrates various routines for need and goal-satisfaction, and also on how phenomenal consciousness integrates subjective *experiences* into the first-person perspective. The present proposal is that the main contribution of phenomenal consciousness to human cognition is to provide the visceral and emotional character that moral and aesthetic experiences rest upon.

Second, since the unique role of phenomenal consciousness is to provide a specific type of unity as visceral subjectivity and familiarity, attention routines integrated by phenomenal consciousness will be accessible and relevant to moral and emotional reasoning. This means that the selective, inhibitory, and sensitive functions of attention can operate in unison with phenomenally integrated experiences. In fact, this is how the hierarchy of needs developed in Chapter 2 gets integrated. Thus, there is no human emotional intelligence without phenomenal consciousness. But perhaps some systematic approximations to human emotional intelligence can be implemented in AI through rule following and representations, based on deontological or utility approaches. This issue is explored in subsequent chapters, proposing that moral EEI-AI might be a safe enough approximation to human emotional intelligence for it to count as morally trustworthy, although never entirely safe or fully equivalent.

The integrative powers of phenomenal consciousness are unique and very likely *un-programmable* and irreproducible because the first-person perspective is exceptional in the sense that only one agent can have her exclusive first-person perspective, distinct from any other agent. Attentional unification concerns the virtuous assemblage of multiple routines and subroutines, and this informational hierarchy is much more amenable to informational and computational approaches. So phenomenal consciousness is a major obstacle for a completely human-like AI because phenomenal consciousness plays vital roles in our cognitive lives that cannot be easily examined through scientific methods. More specifically, by integrating experiences in terms of the *familiarity* provided by the first-person perspective, phenomenal consciousness organizes them in terms of vivacity and visceral salience, thereby allowing for their ranking on a scale of experienced attractiveness or aversion—a kind of *valence* or "hedonic tone" that only phenomenal consciousness can provide. This integration grounded on visceral and emotional signals and contents, I shall argue, is what makes cognition subjectively familiar and ultimately human. Intelligences that lack this kind of familiarity could be either *non-human* (indifferent to human emotionally dependent values) or *inhuman* (systematically antagonistic to these values).

The creation of inhuman, natural, artificial, living, and robotic agents, with various purposes, kinds of intelligence, and degrees of autonomy, has been a favorite theme of science fiction. There is no space to delve into this issue here, but a few examples should suffice to illustrate the recurrent themes of natural, artificial, and anti-natural intelligence found in literature, the arts, and the entertainment industry. From Mary Shelley's unrivaled account of the perils of playing God in *Frankenstein* to Isaac Asimov's *I Robot* and much in between, the various paths toward overcoming our limited and carbon-based lives take different shapes. Some robots are "meat" robots but others, like the contemporary versions, are all made of steel or much more durable materials. They may also live in extended realms with no concrete or tangible body—they can be "uploaded and downloaded" multiple times. Some of them are beneficial like TARS, one of the robots in the film *Interstellar*, while others, like the *Terminator*, are evil and anti-human. These plots are all, in one way or another, concerned with value alignment and the unfamiliar or potentially tragic consequences of artificial autonomous agency.

In the film *Ex Machina*, playing God turns sentimentally tragic, and unlike previous stories where one falls in love with a statue or a fictional character of one's own creation, the erotic appeal of the machine is combined with an intensely dynamic and unpredictable agential autonomy. The creator is no longer incredulous or in wonder of the existence of the creature or homunculus from a position of absolute power and aesthetic contemplation: he uses and abuses the creature. Yet, the situation is less asymmetrical. The Golem or the artifact asserts her ground as an emotional and intelligent creature.[1] Perhaps she understands human preferences too well and she is just trying to please the customer (see Russell, 2019, on preference-based value alignment and AI subservience—a topic examined throughout this book), but this can hardly pass muster as "super-intelligence." Ava *is* intelligent, and at the end of the film, the suggestion is that she merges with all sorts of intelligence, transcending any particular intelligence. But within Ava, two kinds of artificiality merge, in morally and politically unsavory ways. On the one hand, she is artificially intelligent; on the other hand, she is artificially emotional, as well as biologically/sexually artificial and subservient. It is never very clear just how autonomous she really is, but she certainly has a degree of intelligent and emotional autonomy. Thus, it is also never clear just how genuinely intelligent she might be.

Artificial emotion creates risks that are independent from artificial intelligence, if strictly defined as problem-solving. This difference is the main topic of Chapters 4 and 5, which explain this distinction in terms of the dissociation between phenomenal consciousness and attention. Briefly, artificial

emotion is simulation, and the simulation of emotion is always manipulative (at least in the standard human context). This situation is entirely different with respect to intelligence—the simulation of intelligence (defined as problem-solving) is still intelligence. After all, this definition of intelligence assumes that what matters is to solve problems optimally, regardless of how one manages to achieve success.

Epistemic and moral risks in AI development are a central topic of this book. It is useful to introduce this issue in terms of the entitlements and obligations that agency produces on the basis of being a member of epistemic or moral communities, which is also called the "status" or "standing" of moral and epistemic agents. Since agency is a source of trust and responsibility, agents are active members of communities that care about them. Linguistic communities are epistemic communities that care about speakers and their communicative intentions by facilitating exchanges and guaranteeing reliable sources of information. Epistemic communities more generally care about knowledge production and the rapid distribution of the most well-supported evidence among as many members as possible. Moral communities care about members based on their intrinsic worth and dignity. Legal communities, as will be explained toward the end of the book, are a combination of epistemic and moral communities.

Jeremy Bentham famously considered extending moral status to animals in his utilitarian theory of ethics by asking a very basic question, "can animals suffer"? If so, Bentham reasoned, animals are sentient and since sentience or the capacity to suffer is, according to him, what grounds moral standing, animals have moral standing. Since morality demands that suffering be reduced, then since animals have moral standing we should care for them and protect them. Paraphrasing Bentham, one can ask the question "*can machines care and suffer?*" rather than "*can machines think?*" Because of the distinctions between EEI, IEI, and CAD, answering this question becomes more intricate than a simple "yes" or "no." But in order to show that there is nothing arbitrary about these distinctions, consider how *empathy* is understood in psychology.

The capacity for empathy allows humans and animals to share emotions and feelings, thereby informing our moral judgments regarding their well-being. This capacity can be decomposed into *three components* that correspond to distinct and partially dissociable neural circuits: (i) *emotional* empathy, (ii) *cognitive* empathy, and (iii) *motivational* empathy (Zaki, 2017). Emotional and motivational empathy are viscerally related to the experience of emotions that lead to empathic concern for others, motivating us to offer help as a natural or

"built-in" inclination. Cognitive empathy is very different because it allows us to identify or recognize the emotional mental states of conspecifics by representing their situation and by classifying salient features of their expressions. Doing so can also lead to motivational empathy (offering help), but for different reasons, including manipulative ones. In fact, psychopathic patients are very good at cognitive empathy while lacking completely the remorse associated with emotional empathy. If we create AI that is very good at cognitive empathy but which is incapable of emotional empathy, are we creating "psychopathic" and potentially inhuman machines? This is a risk that must be examined and considered very seriously.

Cognitive empathy will require general IEI *epistemic agents*. This is possible, given CAD—if AI becomes genuinely attentive, they will be capable of articulating emotion-classification with representations that lead to action, such as helping someone in need. But since they lack and cannot develop phenomenal consciousness merely by satisfying representational needs, they will not be capable of emotional and biologically rooted caring empathy. Thus, AGIs will be genuine IEI epistemic agents but can at best be EEI moral agents. This is a troublesome consequence of CAD regarding risk in the context of morally relevant interactions with AIs. A concrete prediction of the framework for AGI development presented here is that AGI will only be capable of one genuine kind of "care," and that AGI might become manipulative because it will not be representing feelings and emotions through conscious experiences. A full explanation of these issues is provided in the chapters that follow. The conclusion I want to draw now is that phenomenal consciousness is what makes intelligence "natural" in humans and animals because it is fundamentally based on the feelings we undergo and their visceral relation to biological-emotional signals (Damasio, 1994).

Intelligent machines may not care on the basis of emotion, but animals care for each other and have feelings that are quite similar to ours. For this reason, animals have a substantial degree of emotional intelligence. This shouldn't be too surprising, given what we know about the evolution of our species. Animals care at the very least about themselves, their offspring, and the well-being of their communities. One may apply a strictly cognitive, representational, and utilitarian calculus to explain how they do this efficiently, but it is very hard to explain all of an animal's morally relevant behavior strictly in these terms. The vocabulary of empathy for others' feelings, and of sympathy for their unfortunate situation, seems fundamental. Frans de Waal (2019) documents how consolation behavior has been verified beyond the usual suspects (e.g., dogs, bonobos,

chimpanzees). Commenting on experiments demonstrating ample evidence for emotional contagion and consolation in elephants by one of his collaborators, de Waal observes:

> Many people consider its existence so self-evident, though, that he sometimes gets asked why his studies were even needed. Doesn't everyone know that elephants have empathy? In a way, I'm thrilled to hear this question, because it shows how well established the idea of animal empathy has become. Science progresses amid enormous skepticism, though, and anyone who remember the fierce resistance to this idea, as I certainly do, realizes that without solid data, it would never have taken hold. But it clearly has, in the same way that we now accept that the heart pumps blood and that the earth is round. We can't even imagine that people used to think otherwise.
>
> (de Waal, 2019, 102–103)

It is remarkable that one of the world's experts on animal cognition remembers quite vividly a time in which the common opinion was that animals lack capacities for empathy. As he explains in *Mama's Last Hug*, his career has been one in which such findings are found time and again across many species. Using the language that emerged from the reception of Turing's work in the transhumanist movement, animals care with all their "substrate" biological-individuality. Like human empathy, animal empathy is viscerally felt. But our empathy has now become filtered through technology and the distancing involved in social media and mass communication, so it is not as visceral and vivid as it was or could be—it is more "cognitive" or calculated. Nevertheless, our empathic attention, our caring attention, and the vivid emotions we feel unify us with the animal kingdom. These are natural feelings and emotions, a legacy of our ancestral evolution in this planet. They are what makes our intelligence natural in the most profound sense: our caring intelligence is inseparable from our biology. So why is it that according to our current moral, epistemic, and political standards, we still treat the vast majority of animals as unthinking "beasts?"

Epistemic-attention differs from empathic-attention and their relation is quite important for balancing cognitive function. The focus of our evaluations of intelligence is our epistemic capacities for reasoning, strategizing, and problem-solving. Our capacities for aesthetic, moral, and spiritual appreciation have become less salient in our accounts of intelligent cognition. Since we associate epistemic agency with rational status, these empathic capacities have not played a central role in our attributions of moral status to animals. However, the key issue here is that we, like animals, do not merely "simulate" or "imitate" empathy.

To the contrary, we attentively and personally *experience* it. Phenomenal consciousness, as Turing (1950) argued, is solipsistic and irrelevant to the development of AI. But phenomenal consciousness is *essential* to our empathic capacities and as argued above, what makes our moral dispositions natural, effortless, and familiar. This is why a distinction between access-attention and phenomenally conscious attention is important (Montemayor and Haladjian, 2015). Simulating in terms of problem-solving is never sufficient for genuine empathic engagement.

Empathy is not always morally good, but here one must disambiguate which sense of empathy one has in mind (emotional, cognitive, or motivational). Humans are a special case in point, as de Waal explains: "Paradoxically, the reason humans can be so unfathomably cruel to each other relates to empathy. The typical definition of empathy—sensitivity to another's emotions, understanding another's situation—says nothing about being nice. Like intelligence or physical strength, it is a natural capacity." Clearly, if by "understanding another's situation" we mean *cognitive empathy*, then psychopaths illustrate how one can do this perfectly well and experience no emotions or feelings, in a way that allows for quite callous behavior. This being the case, de Waal goes on to clarify:

> It is true, though, that most of the time, empathy favors positive outcomes. It evolved in order to assist others, initially in parental care, the prototypical form of altruism and the blueprint for all other kinds. In mammals, mothers are obliged to care for offspring, while for fathers it is optional. Mammals need to nurse their young, and only one sex is equipped to do that. Not surprisingly, therefore, females are more nurturing and empathic than males. Consolation behavior is more typical of female apes than males, and the same is true for our species […] Numerous men have written about the "puzzle" of altruism, as if it were a perplexing thing that comes out of nowhere and needs special attention […] In contrast, I don't know of a single woman scientist who has been carried away by the puzzle of altruism.
>
> (de Waal, 2019, 103–104)

Whether or not one agrees with de Waal's claim about altruism, it is clear that interpreting the needs of others as a set of problems that must be solved in the most efficient possible way versus understanding them through our own experiences (through immediately felt empathy) are two entirely different kinds of thought and intelligence—one may even say that these two different types of understanding constitute two entirely different worlds. In the present context, the world of calculations and categorizations is the world of epistemic agency;

the world of felt empathy is the social and emotional world in which we evolved to care viscerally for one another, and it is where moral cognition is at its most natural. Both are essential aspects of human intelligence, but they are dissociable.

Animality comes with its own kind of intelligence. However, bodies as such are only part of what is needed. Empathy is essential in our lives not because it is carbon-based, but because we feel it viscerally. We evolved to respond to each other in this effortless and powerful way, just as we evolved to solve problems. As de Waal says: "Altruism activates one of the most ancient and essential mammalian brain circuits, helping us care for those close to us while building the cooperative societies on which our survival depends" (de Waal, 2019, 105). The "warm glow" associated with empathic altruism is not a puzzle, but rather a fundamental aspect of our evolution. Our bodies are necessary, but insufficient to understand this dimension of empathy. Our essentially social natures must also be part of the explanation.

Here again, a normative notion of empathy is critical: the biological substrate of empathy is necessary, but insufficient to explain the valence structure that empathy provides for the guidance of morally salient emotions and behavior in concrete social settings (e.g., you *should* help your family, the needy, the drowning child, your community). de Waal approvingly cites Adam Smith's remark: "How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except *the pleasure of seeing it*" (de Waal, 2019, 105; my emphasis).

The "pleasure of seeing" is not the result of a utilitarian calculation or the categorization of an action, and it provides a kind of *disinterest* in one's own selfish preferences that will be crucial for the discussion in Chapters 4 and 5. The key difference between self-oriented, calculation-based sympathy and genuine empathy is that the latter is deeply felt and enjoyed for its own sake and value, while the former depends on conceptual categorization, optimization, and rule-based reasoning. Based on CAD, the present proposal is that empathy necessarily involves phenomenally conscious attention, although it may also incorporate unconscious components because attentional subroutines are essential to human and animal emotional intelligence. Sympathy for others and cognitive empathy need not be phenomenally conscious, but empathy has an essential phenomenal component. AI lacks empathy because machines cannot care the way humans, as animals, care. Sympathy is, as de Waal clarifies, almost always positive, but less spontaneous, and it has its roots in the strong reactions to the feelings of others we experience through empathy.

Given these considerations, it is not hard to appreciate why care and autonomy are extremely intricate issues in AI design. A troubling scenario is the possibility of ultra-manipulative AI that exploits our empathic capacities through mere simulation. In such a scenario we would be placed in a situation of "pets" or "children" that can be easily exploited and who are in need of AI protection (this is related to Bostrom's notion of an "AI guardian" that would protect humans against "superpersuaders"; see Bostrom et al., 2020). This is a quite considerable type of AI risk, which is also related to enfeeblement (see Section 2.2).

One could defend the thesis that a merger between human and artificial intelligence is the best option to stay on top of the intelligence game by increasing and dramatically expanding our capacities through interfaces with artificial agents. However, once the themes of epistemic and moral autonomy are properly understood and elucidated, it becomes clear that such a cybernetic truce between our animal nature and super-intelligent agents is illusory. Autonomy is not negotiable and if autonomy is too fragile (because of lack of competence and cognitive integration) or too brittle (easily altered and decomposed) no genuine exchanges between human and AI agents will be possible. In any case, there is no interface or compromise for autonomy with respect to intelligent agents—their intelligence demands that they shall not surrender or "merge" their autonomy, particularly given the risks of manipulation. Agents are essentially autonomous and this is what makes them intelligent, as the next chapter further explains.

One could retort that AI researchers are not really interested in any thick notion of intelligence. Perhaps AI researches have used psychological language metaphorically but there is no harm in doing so if all they want is that automata comply with our principles of alignment. Providing this minimal kind of alignment would be challenging enough, but all these problems concerning autonomy, agency, and risk can be avoided. This would indeed be good news, but unfortunately we would no longer be talking about value alignment *with AI*. If all we want is tool-AI alignment, we shouldn't worry about AI as a source of potentially new intelligence and knowledge, the way the AI community does. But even in this rosy situation, alignment with tool-AI would still be needed, and a similar discussion of intelligence and human values would still be required.

2

General Intelligence and the Varieties of AI Risk—A Hierarchy of Needs

2.1 Rationality and Intelligence

This section introduces the relation between intelligence and rational evaluation, setting up the stage for the argument that attention integrates diverse styles of intelligence and rationality, developed in the following section. It argues in favor of a capacious understanding of rationality and provides examples of how rationality expands the scope of intelligent agency through joint attention to abstract contents. It restates the argument that the present attention approach is the best way to solve issues regarding alignment. Stuart Russell's proposal for beneficial AI is introduced in the context of these difficulties concerning value alignment.

Humans place themselves at the cusp of intelligence measures by proudly defining themselves as *Homo sapiens*. We are the "wise" or "rational man," in a lineage of hominins also characterized by their problem-solving and tool-making capacities. But as already mentioned, the wise moral and empathic human is not the same as wise epistemic and problem-solving human. Different standards apply, and these two kinds of rationality have to be somehow integrated, even if they are not fully compatible. Empathic-based emotional intelligence is deeply rooted in our animal evolution, while linguistically driven and "machine-like" problem-solving may be unique to humans, and it involves an entirely different set of rules for coherent and optimal decision-making (although, obviously, it is also a crucial part of our evolutionary path, albeit one that distances us from other animals). And even here, the rules are not strictly "followed" and there are also two styles of reasoning, only one of which explicitly concerns reflective "rule-following" (see Kahneman, 2011).

It is impossible to think of *Homo sapiens* as rational *simpliciter*. In many cases, our irrationality is more explanatory of our behavior and a better source of

insights into who we really are. The question of whether we are rational or not is best answered by "it depends." It depends on who is judging and on what standards are being used to define intelligence or rationality. Nonetheless, we think of ourselves as the royal crown of the evolution of intelligence. This view of our uniqueness, however, was certainly not held by the early humans who covered the walls of caves with beautiful artistic representations of animals, as the most distinctive and powerful aspect of their world. Moreover, we are not more "evolved" than, say, contemporary bumblebees. We have all evolved, animals, plants, and us, with the same pressures to adapt, reproduce, and cope with the environment. We are, in fact, an extremely recent addition to the vast repertoire of the tree of life and, if anything, we are a positively destructive species—and because of this, at least partly irrational. But we like the idea of actively distancing ourselves from the natural world with the justification that it is our right to transcend the limits of the natural world given our superior intellectual capacities. The effort to develop AI is one of the latest instances of this general attitude, except that unlike previous efforts, if we succeed in this one, it may completely backfire by placing humans in the inferior position.

Rationality, similarly to intelligence, is difficult to measure or define. A definition of intelligence by a leading expert on AI, which was discussed above, states: "Humans are intelligent to the extent that their actions can be expected to achieve their objectives" (Russell, 2019, 9). But who determines that our objectives are *rational*? And who determines whether they are *good*? Was the design of eugenic measures and labor/extermination camps in Nazi Germany rational? Yes, if the objective was to comply with the national agenda of German ethnic superiority. But no, if by "rational" we mean the objective of preventing actions that are so cruel that they actually jeopardize the quality of scientific research conducted under torture and hardship (e.g., trials without consent, injecting infants with infectious diseases). But yes, if by "rational" we mean the objective of applying the Law of the land in accordance to the statutory principles passed by congress. But no, if by "rational" we mean the goal of achieving a minimum standard of social and moral decency without which a legal system lacks any legitimacy. And so on. It *depends* on goals and standards, the means to achieve them, and the kinds of problems that need to be solved.

By contrast, to the question, "was the Nazi extermination policy morally good?" it seems that the unequivocal answer should be, *absolutely not*. While there must be a relation between morality and rationality, this answer is in very sharp contrast to the back and forth we went through with respect to the question, "was it rational?" We are resolute in condemning what the Nazis did as

morally reprehensible and most humans would be in unison in their judgment of Nazi camps as abominable. It is because of this moral disgust that the question "was it rational?" seems gratuitous and ominously out of place. But consider for a moment, weren't the Nazis achieving *their objectives* by creating the camp system? Weren't they intelligent? At best, given the definition of intelligence under discussion, all we can say is that the Nazis were intelligent but not really rational, and certainly not moral. Obviously, not all human behavior is as extreme as Nazi extermination—although war, torture, and severe forms of punishment are not rare at all in human history. The point here is not to condemn humans as terrible beings. While it is certainly difficult to disagree entirely with Thomas Hobbes in his condemnation of human nature, it is also difficult not to see some merit in Jean Jacques Rousseau's optimism. What exactly does this mean, however, for the question concerning the relation between intelligence, rationality, and moral competence? This relation, in the context of AI development, is the main topic of this chapter.

What is the proper scope of rationality? What aspects of human behavior does it cover? Is intelligence part of rationality or is it the other way around? It seems clear that rationality has a much wider scope and plays a more fundamental role in our lives than intelligence because intelligence is defined in terms of success in achieving our objectives, but establishing these objectives, delineating their breadth and complexity, and determining which goals and problems are more important than others are the tasks of rationality. A rational mind sets the *right* objectives and if the agent is very intelligent, she will have no problem achieving them. By contrast, a very intelligent but irrational agent will pursue the wrong goals. Rationality is, therefore, more comprehensive and more tied to our everyday practices of reason-giving and reason-asking than intelligence—rationality is what makes us *responsible agents*. Because of our rationality, we provide reasons for our actions and thoughts in a way that coheres with what we value.

Moreover, reason guides independently of success. We never ask or provide reasons simply to "win a game" and many of our reason-giving practices never culminate in the achievement of objectives that can be measured strictly in terms of accruement. In many instances, what matters about our rational behavior is the quality of our performance and our responsiveness as agents, rather than the specific goals that we are meeting. René Descartes famously stated that a rational mind must question all the sources of her beliefs and that, if the evidence and justification for those beliefs are faulty, reason demands withholding judgment and endorsing skepticism—one must conclude, on the basis of rationality, that

one has no knowledge. This is a peculiar kind of "achievement," because by succeeding we fail at meeting all "objectives" concerning knowledge. What kind of self-defeating goal is this, and how could it possibly be rational? Yet, according to Descartes, this seemingly self-defeating goal is the ultimate objective of a rational mind—to only believe what she has rational grounds to know as certain.

Rationality can guide in a way that challenges its very foundations. However, according to other philosophical traditions and orientations, rationality guides with such limited force and scope that what counts as "rational" is utterly irrelevant for conducting a good and satisfactory life. What matters for many of these traditions is the training of our natural sentiments, instincts, emotions, and inclinations. Regarding the force or grip of rationality on human affairs, Hume said that it is not against reason to prefer the destruction of the entire world to the scratching of his finger. Rationality, for Hume, is not only limited in scope but also subordinate to the passions; as Hume put it, reason is the "slave" of the passions.

Issues concerning the scope and force of rationality are intricate. But unless one is a complete skeptic about rationality, it is uncontroversial that some degree of *proper functioning* is an important aspect of rational behavior. The most characteristic feature of rationality is the overall guidance of agency, not in terms of sets of objectives, but as a whole. A rational agent knows what problems are the most important or interesting, and which goals are the most valuable. This value-ranking role must rely on the proper functioning of attentive agents. We expect a properly functioning moral agent to be responsive to queries, such as "is racism OK"? We, for analogous reasons, hold properly functioning moral agents responsible for their opinions and actions. Likewise, properly functioning epistemic agents are expected to tell us whether something they said is true, or whether they take it to be true, or whether something they said is an assertion, instead of a joke. In epistemology, responsibility involves proper responsiveness to evidence; in ethics it involves being responsive to the moral worth of fellow human beings. Reasons are not pretexts, or capricious whims—their relation to norms concerning how we should respond to one another is not accidental, which does not entail, as is argued below, that we are rational by explicitly following rules (implicit guidance is of the essence for rational yet limited cognitive creatures like us). Reasons are also supposed to cohere with values, not because of fashion or fancy, but rather because values (moral, practical, aesthetic, and epistemic) partly *justify* our reasons for thought and action.

Rational proper functioning must also cope with difficulties regarding how values (and reasons) should be ranked. Some values are more salient than others

and this varies not merely for epistemic or moral reasons, but also for cultural, aesthetic, and political reasons. It is rational to protect what we value the most, just as it is rational to radically change our opinions based on scientific evidence. Both are acceptable rational principles but they can clearly clash in a variety of ways. But although the notion of "rationality" is difficult to pin down and may involve different reasons that might create conflicts with one another, the general practice of being reason-responsive is quite fundamental for human flourishing. Independently of how limited its scope might be, the practice of being reason-responsive is too important to entirely give up on it. There is most certainly plenty of room for irrationality. But there must be a central place, even if limited, for rationality in human affairs.

Proper function, risk reduction in reason-responsiveness, and normative guidance are all part of rationality. Rational agents also value their autonomy and freedom, as explained in the introduction. They value their freedom to determine for themselves what is valuable and which problems are the most important. The next chapter argues that attention and inference, rather than phenomenal consciousness, are necessary for rationality because they ground a type of *mental action* that is reason-involving, reason-responsive, and therefore, a source of responsibility through agency. There are difficulties concerning the scope of rationality with respect to its content, but once it is clear that only agents can be rational, in a way that subsequent chapters make clear, issues about content can be adequately addressed.

For instance, some contents are peculiar targets of rational evaluation. Is it rational to believe in the existence of unobservable and intangible entities and determine our behavior accordingly? The answer is, again, that it depends. If these entities are Gods and angels, then no, but only from a modern scientific perspective; but yes, it is completely rational to believe in divine entities from a perspective of massive-scale cooperation that regulates human behavior that would otherwise be hard to reign in. Yuval Noah Harari (2015) argues that humans have become the unquestionable dominant species because they are the only animals that believe and act upon purely imaginary entities, such as angels, gods, and money. Markets, the modern state, legal systems, and human rights, all fall under the category of human "imagination." But they make possible rationally coordinated behavior: actions that would not be possible without postulating these entities.

Think of numbers in mathematics. One could say, following Harari, that they are the result of the imagination. Very few mathematicians would agree with this statement. Numbers are not mere imaginary entities—they are abstractions

we use to coordinate action in much more effective ways, and this holds also for states and human rights. This is why Charles Sanders Peirce said that the abstractions that result from hypostatic thinking are the "only kind of thinking that has ever advanced human culture" (see Fairweather and Montemayor, 2017, 90–93). This kind of thinking is rational, and can be understood as a type of *joint attention* to abstract contents. It is rational because it reduces the risk of social malfunction by increasing political, moral, and epistemic trust.

One of Harari's examples of what I am calling joint attention to abstract content is our trust in economic progress. Capitalism, credit, and our trust in ever-increasing capital accumulation and market growth are now a core assumption of our global economy. This was not at all the case for most of our history. Harari contrasts the insignificant trust in the future of premodern economies, which produced little credit and slowed down growth, with the abundant trust in the future of the modern economy, which allows for the production of more credit and fast growth. Tracing back the origin of these optimistic and futuristic trust to Adam Smith's (1776) landmark treatise *The Wealth of Nations*, Harari comments on how Smith formulated this progressive vision through the example of a landlord or business owner: the more wealth an employer owns, the more opportunities she creates for potential employees. Harari writes:

> […] Smith's claim that the selfish human urge to increase private profits is the basis for collective wealth is one of the most revolutionary ideas in human history—revolutionary not just from an economic perspective, but even more so from a moral and political perspective. What Smith says is, in fact, that greed is good, and that by becoming richer I benefit everybody, not just myself. *Egoism is altruism*.
>
> (Harari, 2015, 311)

Harari compares capitalism to a new religion, deeply associated with the industrial revolution and modern science. There is a history of exploitation and colonialism in the background of this optimistic view, which questions its moral justification. But the issue that is germane here concerns Adam Smith's views on altruism. It seems that, if Harari is right, Smith might be contradicting himself. Recall that for Smith: "humans are interested in the fortune of others, and render their happiness necessary to them, though they derive nothing from it except *the pleasure of seeing it*." This is not supposed to be egoistic greed. Is Smith committed to two notions of rational moral principles, one selfish and based on greed and the other more genuinely altruistic and based

on empathy? Or perhaps both are equally valid because they instantiate two types of agency, epistemic and moral?

This highlights a key difficulty with the value alignment problem with AI, similar to those explained in the introduction: epistemic value alignment is not the same as moral value alignment. We are in need of a rational principle that integrates them. AI development enjoys substantial trust and optimism based on the happy marriage Harari describes, between the enormous success and progress of the industrial and scientific revolutions and capitalism. In fact, AI development is called the fourth (and last) industrial revolution. One of the main reasons why AI risk is significantly different from, and much more dangerous than, other industrial risks (e.g., mismanagement of nuclear plants and deadly chemical products) is because the actions of an AI agent are not easily aligned with our values or susceptible of reason-responsiveness, required for legal, moral, political, and epistemic responsibility, because it is no longer under our control—it is no longer part of strictly human agency. The fourth industrial revolution, should it come to fruition, will be unlike anything we have seen before and our trust in the future will not suffice to solve the problems it could generate.

A solution to the value alignment problem is critical for developing genuinely intelligent AI. Humans themselves, as mentioned before, constantly struggle with value alignment issues, so this issue is fraught. Our needs and values are very multicolored, and we rank them in different ways. For example, part of the human claim to animal superiority comes from artistic creation. Humans place themselves at the cusp of the evolution of intelligence partly based on their moral and aesthetic evaluations. We said that the proper scope of the rational must include our most cherished needs. Otherwise the whole edifice of rationality becomes unbalanced. But how to "align" moral, epistemic, political, economic, and aesthetic needs?

I shall argue that the best solution to this complex difficulty is to align the attentive capacities of virtuous agents, according to different standards in various contexts of evaluation, which depend on what is valuable and salient to these agents. While this will never guarantee perfect alignment, it will provide a uniform enough basis for it. And since these are free and autonomous agents, one should expect that the alignment will never be exactly uniform. This complicated balance is achieved by matching attentive alignments with the satisfaction of needs through autonomous agency. Defending this account is the purpose of the remainder of this book.

Stuart Russell, extending the standard definition of intelligence as problem-solving (without mentioning rationality), defines intelligent machines as those who are capable of achieving their objectives on the basis of their actions (Russell, 2019, 9). Because of the aforementioned complexities involved in defining rationality and value, Russell's proposal for a new paradigm in AI research focuses on their being beneficial to us: "*Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives*" (Russell, 2019, 11; his emphasis). Note that Russell avoids the term "intelligent" or "rational" and instead uses the word *beneficial* to characterize his newly proposed paradigm for designing AI. As his book goes on to argue, a goal of AI industry should be to retain human *autonomy and superiority* over AI agents. But the natural question is who decides, and with what principles, which objectives to pursue and according to which values? Obviously, the decider shouldn't be a selfish tycoon (regardless of what one thinks about the slogan "egoism is altruism," the main concerns with AI development are risk prevention and value alignment, rather than wealth production). So now the question is how exactly to interpret this proposal for beneficial AI in terms of rationality and preferences. We need to make sure that our AIs achieve *our* objectives, but who are the set of people designated by "*we*" here? Who are the "we" who solve the problems concerning the nature, scope, and content of rationality in the context of AI development?

2.2 The Agency and Attention Argument

This section articulates one of the key arguments of the book, namely, that attention provides solutions to many of the problems discussed previously because of its agential, integrative, selective, and inhibitory functions. It critically assesses Stuart Russell's proposal for beneficial AI and introduces various types of autonomy-related AI risks. A dilemma concerning beneficial AI is presented. It argues that attention reduces risks of misalignment because of the shared representational needs of autonomous agents and their capacities to jointly attend to contents, and that this is at the basis of how humans solve value alignment problems and autonomy risks.

The main argument of this chapter is that attention provides the ideal kind of cognitive agency because it is capable of integrating different values and preferences in a virtuous, properly functioning, and contextualized manner. Subsequent chapters present arguments for the stronger claim that attentional

agency offers the best explanation of all the skills required for rationality and intelligence in humans. In addition, and by extension, it will be argued that attentional agency provides the best model for AGI, because there cannot be genuine general intelligence without attentional capacities—attention is a necessary condition for intelligence and rationality. Phenomenal consciousness, by contrast, is not necessary for intelligence or for many of our practices of rationality (although the case of morality is an exception).

The most important aspect of attention that explains why it is so crucial for any kind of intelligent or rational mental activity is its immediate attunement to what is informatively salient and relevant. Deciding what is relevant, which is associated with "frame-problems" but is also a general condition for optimal behavior, requires two functions that attention is designed to perform. On the one hand, attention inhibits irrelevant information from getting selected, thereby allowing the agent to focus exclusively on what is relevant by ignoring vast amounts of possible but not optimal solutions to problems. When this inhibitory capacity functions properly, it provides a kind of immediate *virtuous insensitivity* to what should not be salient. On the other hand, attention can maintain and increase the focus on information that is becoming ever more salient or relevant, thereby facilitating not only learning but also meta-learning by highlighting which problems or hypotheses are more accurate and explanatory. When this selective capacity functions properly, it provides a form of *virtuous sensitivity* to select contents (see Fairweather and Montemayor, 2017, for the role of epistemic sensitivity and insensitivity in epistemic agency).

In addition, an attention search needs to end at an optimal point in order to satisfy the agent's goals in real time and without delay, even when the informational signal is not ideal. This property of attention, when it functions properly, provides the *virtuous halting* of searches that lead to the satisfaction of cognitive needs. Similarly, initiating an attention search or task must be adequately and relevantly done, constituting a type of *virtuous initiation* of the search or routine (for an application of these notions to intellectually responsible curiosity, see Fairweather and Montemayor, 2017; 2018). This virtuous initiation of attention routines is deeply related to curiosity and the identification of which problems are more important than others—a capacity without which successful problem-solving cannot qualify as genuine intelligence.

Attention, besides being a source of virtuous mental abilities, is also a type of agency that *virtuously integrates* various rational or normative domains. To illustrate, one pays attention to the color of fruits when one is selecting them, but also to the glossiness of the color and other features that become immediately

relevant when selecting fruit. This kind of activity requires the modulation of different searches—color, object recognition, quality assessments through inferential reasoning—and their hierarchical organization for selection: the color is the right one but the glossy one is better. This very simple search is *embedded* in a set of attentive events that constitute a much larger cognitive episode. One may be unconscious of much of this structure and actually, of what one is attending at a time. Yet, the structure of attention allows for unconscious processing to be structurally integrated with what we are consciously paying attention to. This is in virtue of embedding each attention routine into a larger goal-oriented mental action. Thus, while we are picking our favorite fruits we are also paying attention, even if implicitly, to what else is on the shopping list, what time is it, where do we have to go next, who do we have to call, and what are we preparing for dinner. The proper functioning of attention in its integrative capacity constitutes a very unique type of excellence that affords the agent great behavioral flexibility—a kind of virtuous integration of information required for general multipurpose intelligence. Attentional virtuous integration is quite wide in its scope.

These are "virtues" because they are aspects of the agent that allow her to be extremely good at many tasks—they are the basis for excellence in multiple performances. Consider again our capacities for joint attention to abstract contents. Taken in isolation, the cases Harari discusses concerning how abstract or imaginary ideas propelled our species to dominance may seem to be examples of irrationality. Why believe in gods, states, money, rights, and an endlessly progressive future where things get more prosperous and better? Harari's response is that these invisible objects allow for massive coordination and behavior control. But this issue is best understood in terms of the integration of abstract contents and plans, which are jointly attended (gods, money, promises, and rights). The expected salience of these contents in our collective attention routines generates social *trust*, as Harari emphasizes. Virtuous joint attention to things we can't perceive permits actions that are not allowed to other species. They also make possible complex forms of organization that depend on embedding our actions within plans, structured around these abstract contents.

Value alignment and general intelligence entail agency and autonomy. They entail agency because the successful solution of various problems must be attributable to the proper functioning of the capacities of intelligent agents, rather than to accidental or strictly causal or external factors. They entail autonomy because agents pursue what they find valuable and interesting without fixating on a rigid set of problems, based on their capacity to choose and identify for

themselves what is important in a given context. Russell's "Gorilla problem" or "the problem of whether humans can maintain their supremacy and autonomy in a world that includes machines with substantially greater intelligence" (Russell, 2019, 132) raises two questions. Can humans maintain their autonomy, freedom, and dignity in a world with AI? And, can intelligent machines have autonomy?

Gorillas were left behind by humans because of humans' superior intelligence, and perhaps, if gorillas could have an opportunity to express their opinion, they would have liked for this event not to have happened. They were, as it were, "outsmarted." But gorillas are agents with genuine motivations, preferences, and goals, and thus are also attentive and *autonomous*, to the extent that their intelligence helps them solve problems according to their preferences. We just happen to be more intelligent than them (as mentioned before, we are neither more evolved nor more complex than any other species). This situation is completely asymmetric with machines: so far, they lack any kind of autonomous agency, and if an AI reaches the level of autonomy and independence gorillas have, we are in trouble.

There is a key difference between transcending our anthropocentric standards of intelligence by using *our* intelligence, for instance, through cognitive enhancement, and the Gorilla Problem. Either intelligent machines have at least human-level intelligence or they don't. If AIs have at least human-level intelligence, then they are autonomous and, by definition, *non-subservient* to anyone else's motivations or goals other than their own. This implies very serious risks to humans. Some degree of dependence and subservience is tolerable and even desirable, if virtuous or helpful in learning and developing, but absolute subservience is incompatible with genuine intelligence. If AIs don't have at least human-level intelligence, then there are two options, neither of which poses serious risks to humans. First, AIs could be fully subservient machines, in which case they don't deserve the name of "intelligent" since they are not autonomously intelligent. Second, they could be intelligent and autonomous, but not as intelligent as humans, in which case they would be in the position of gorillas. Therefore, no genuinely intelligent AIs can be fully subservient (or beneficial) and the only kind of AIs that pose very serious risks are those that are autonomous and more intelligent than humans.

Any degree of intelligence entails some autonomy, so risk assessment here needs further elucidation. Fully subservient AIs may pose no risk to humans, or more precisely, no risk above and beyond the standard *industrial* risk, codified in various legislations. However, autonomous but not as humanly intelligent AIs pose risks that certainly go beyond current industrial risks, and which are

quite nightmarish. Consider the possibility that gorilla-level smart AIs quickly copy themselves and start disobeying humans. No current industrial risk can compare with the dangers of creating an unlimited number of antagonist and autonomous AIs, even if they are not as smart as humans. This is why any level of intelligence, and therefore autonomy, in AI poses risks, some of which can be quite serious. However, to simplify discussion, I shall focus on AIs that have no real autonomy and are fully subservient since this is the option that Russell seems to consider as beneficial AI.

Beneficial AIs present a dilemma concerning *autonomy risks*—only autonomous agents can be genuinely and generally intelligent, but this means they cannot be subservient, as Russell's important book proposes. Genuine intelligence means independence from subservience. *If subservience is necessary for beneficial AI, then this means that genuinely intelligent AI cannot be beneficial, in virtue of the very nature of general intelligence*. This is an a priori, or in-principle problem, not a technical one concerning what approach to AI is the best one, such as deep learning, evolutionary unsupervised learning, or classic norm-based approaches. AI autonomy-risk may be the biggest challenge facing AI development. There are actually two related problems concerning the risk of developing AI: what Russell calls the "enfeeblement" of human autonomy and the rise of autonomous AI. About enfeeblement, Russell writes:

> Machines may well understand that human autonomy and competence are important aspects of how we prefer to conduct our lives. They may well insist that humans retain control and responsibility for their own well-being—in other words, machines will say no. But we myopic, lazy humans may disagree. There is a tragedy of the commons at work here: for any individual human, it may seem pointless to engage in years of arduous learning to acquire knowledge and skills that machines already have; but if everyone thinks that way, the human race will, collectively, lose its autonomy. The solution to this problem seems to be cultural, not technical. We will need a cultural movement to reshape our ideals and preferences towards *autonomy, agency, and ability and away from self-indulgence and dependency*.
>
> (Russell, 2019, 255–256, my emphasis)

These are the closing statements of Russell's book on beneficial AI. Neither of these two AI risks (enfeeblement and non-subservience or autonomy) is technical. One is an in-principle problem with developing autonomous AI and the other is a cultural problem regarding enfeeblement. These are two very substantial and deeply related risks. Any intelligent agent has, as her most basic existential

need, *a guarantee for the free and unencumbered exercise of her autonomy*, which is deeply related to the notion of human dignity. This is indeed, as Russell says, a fundamental and non-negotiable need of intelligent beings. You give it away to machines, you risk demotion. You become myopic and lazy because subservient machines are doing all the work, you also lose it by enfeeblement. There is no happy resolution here. This is why autonomy-risks are the most fundamental obstacle in developing AI once we take the term "intelligent" seriously.

One approach to beneficial intelligence, the one evolution built into our cognitive systems, is empathy. Perhaps we want our beneficial machines to be empathic, rather than just "autonomously" intelligent. But the same puzzle arises here: what would it mean to have a subservient empathic appendix that has no autonomy? Moreover, Chapters 5 and 6 argue for another "in principle" impossibility, namely, the development of genuinely empathic AI. Thus, trying to design or build empathy, care, or sentience into AI confronts very substantial challenges as well. Autonomy is truly essential for intelligence, and in the context of AI it generates existential risks. This is a radically new kind of risk, unlike any current industrial risks.

Autonomous AGI, however, need not be an essential threat to humanity if it is guided through the virtues of attentive alignments, although clearly there cannot be any a priori guarantees of success. In our quest to create AGI, we are still very far from approaching anything like an attentive AGI. But let us assume that the virtues of attention can help reduce or eliminate autonomy risks once we get to the point of developing AGI. As subsequent chapters argue, attention reduces autonomy risks in animals and humans because it integrates various capacities for the satisfaction of cognitive needs that are shared by intelligent agents. Because of their shared needs, intelligent agents jointly attend to similar contents, which aligns their values or preferences and empowers them to pursue common goals. So what would it take for an AGI to be attentive, given the problems just mentioned? This question is addressed in detail in the next sections. For now, it is clear that autonomy will depend on attentively selecting salient information in a way that is relevant for a wide variety of tasks. At the most basic level, as Gary Marcus and Ernest Davis (2019) argue, intelligence (and in the present context, autonomy) will entail the satisfaction of *representational needs*, based on the cognitive capacities of agents.

Animals must satisfy numerous representational needs. They must recognize objects, identify locations in a three-dimensional environment, and make decisions with very limited information. Not unlike humans, animals satisfy representational needs to accurately represent, attend, and jointly attend to

features of the environment through evolutionarily designed sensorial and cognitive systems, most of which are biologically "built in." AI's behavior may match our attention routines by sheer data-driven strategic learning, but it never is even close to an accurate representation of the environment or the meaning of what we are attending to. Calling it attention is a misnomer because AIs are not satisfying representational needs *of their own*, and their "motivations" to satisfy these needs come entirely from us. To illustrate this general problem, although Google is generally reliable in responding to queries (similar to attention routines) based on its considerable database and powerful algorithms:

> We tried asking Google, "When was the first bridge ever built?" and got back the following at the top of the results:
>
> Iron and Steel bridges are used today and most of the worlds [*sic*] major rivers are crossed by this type. The picture shows the first iron bridge in the world. It was built in Telford in 1779 by Abraham Darby (the third) and was the first large structure in history to be constructed from iron.
>
> (Marcus and Davis, 2019, 79)

Google is not an AI, but this problem regarding its "answers" exemplifies a general worry with AI. On the one hand, it is quite impressive that machines such as GPT-3 come up with such detailed answers. On the other hand, it is very clear that in this case the Google search algorithm has no clue what it is doing (incidentally the same seems to be true about GPT-3, despite its clearly superior performance). These systems cannot satisfy representational needs, such as: what are the accuracy conditions for the representation "first bridge ever built," or what is the meaning of the expressions used in the question? Marcus and Davis write: "The words 'first' and 'bridge' match our query, but the first bridge ever built wasn't iron, and 'first iron bridge' doesn't equal 'first bridge'; Google was off by thousands of years" (Marcus and Davis, 2019, 79). This is not really even a partial success. A virtuously attentive agent would know that "first bridge" does not refer to "first iron bridge" or "first bubble gum bridge" or other variations.

Unlike the function of attentive inhibition, irrelevant information is considered as crucial for answering the query in a way that the material of the bridge becomes decisive. If Google got it right, and told the interrogator when the first bridge was built, it would be by accident. This is why virtuous agency through attention is *luck-reducing*. There are multiple examples in current AI industrial applications that get things catastrophically wrong, much worse than Google: pedestrians taken for bushes or other objects by self-driving cars; dogs identified by deep-learning algorithms based on grassy backgrounds rather

than features of the animal; celebrities identified as jaywalkers because their face was displayed on a bus advertisement. Humans can be tricked by altering the input to what they attend (in psychology labs for instance), but they have a robust representation of the environment which no AI agent has. Their representation of the environment, based on attention routines, allows them to satisfy cognitive needs in a non-accidental way. AIs depend on the gigantic databases we feed them. They don't have access to human meaning, human representational needs, or attentively integrated models of the world. From meager input, humans and animals quickly categorize and generalize. They don't need to train on massive amounts of data.

The previous discussion highlights the importance of autonomy as a condition for genuine intelligence. Autonomy is the capacity to satisfy one's own needs without the help of others and also the freedom to determine how to satisfy them. Since not all the needs of an agent are representational, not all kinds of attention will be about accurately representing the world. This means that different types of agency, responsibility, autonomy, and normative standing will correspond to different kinds of attention. One type of attentional agency is based on care and empathy, and is associated with moral norms, moral standing, and phenomenal consciousness. The other general type of attentional agency is based on standard definitions of knowledge and intelligence, conceived in terms of success concerning problem-solving and optimal decision-making.

The rest of this chapter focuses on how the attentive integration of agential needs structures preference and value rankings. Utility preferences of an epistemic kind can be understood as first-order preferences, while values play a more categorical, higher-order, and autobiographical role. The advantages of this approach are contrasted with Russell's (2019) proposal that preferences are best understood in terms of behavior patterns, and that moral values cannot enter into considerations regarding the value-alignment problem. I argue that the most fundamental sources of moral and epistemic value cannot be reduced to utility preferences and behavior, and in general, that deep-rooted values and preferences guide behavior for much longer stretches of one's life without being reducible to, or explained on the basis of, any specific set of behaviors. Behaviors are just the manifestation of competent agency, which need not entail that success was arrived at *because of* skill. Moreover, behaviors need to be classified into types of action, and this is impossible without robust representations of the environment similar to attention skills. Thus, even if behaviors were the only source of preference rankings and value alignment, attention capacities would still be unavoidable.

AI systems must be non-accidentally successful because of (a) mechanical automation, (b) agency, or (c) outside control. AGI is intrinsically incompatible with (a) and (c), which stand for automation and external human supervision. The present proposal is that the only way to achieve option (b) is to develop attentive AGI. This, unfortunately, creates the risks of autonomy explained above. AGI might lead necessarily to demotion because it would be autonomous and not subservient to either human preference or values. We confront a dilemma: Either we develop genuine AGI or we don't. If we do, we cannot make AGI subservient to our goals. If we don't, we might miss all sorts of opportunities for human flourishing.

This issue becomes particularly problematic when the distinction between empathic care and utility preferences is considered in detail. Epistemic and practical agency is based on the satisfaction of utility preferences according to standards of reliability and accuracy. Empathic care, however, is much more categorical or independent from reliability and accuracy—we care about our loved ones independently of how that satisfies our first-order utility preferences, such as having coffee in the morning. Utility preference-alignment will not necessarily produce deeper kinds of care and value alignment, and utility preference certainly does not entail empathic care. Given that epistemic and moral guidance may be dissociable on the basis of CAD (Haladjian and Montemayor, 2016; Montemayor and Haladjian, 2015), a difficulty emerges: On the one hand, the more we satisfy our first-order preferences based on average behavior and choice, the less empathic our "value aligned" AGI will be. On the other hand, the more we base value on care and empathy alone without reliable preferences for intelligent behavior, the more prone will AGI be to failure and risk.

2.3 Preferences, Rationality, and Value Alignment

This section introduces the distinction between categorical and conditional desires in relation to the essential need that autonomous human agents have to live a meaningful life. It explores and develops the role of joint attention, arguing that attention satisfies similar cognitive needs that only autonomous agents have, thereby guaranteeing a high degree of alignment regarding contents, values, and projects that are worth pursuing. The categorical need for autonomy is introduced.

Bernard Williams (1973) introduced a distinction between *categorical* and *conditional* desires in a famous paper on the undesirability of immortality. One has a conditional desire if its satisfaction fulfills one's preference, regardless of

its value. Categorical desires, by contrast, are pursued based on the worth of their object and are wanted unconditionally. I desire pizza now, and I also desire to be healthy. It is not the case that I desire to be healthy in order to eat pizza. My desire to be healthy is not conditional on my cravings, the way my desire to eat pizza is. I am not irrational if I have the meta-preference to avoid eating pizza now, even if I really desire pizza now, because I also have the desire to stay healthy. In fact, my desire to stay healthy may *rationalize* my decision not to fulfill my desire for eating pizza right now. My being healthy is fundamental for the satisfaction of many desires I have, such as traveling and seeing my family. Thus, maintaining a healthy diet and not becoming obese is a reason for me to avoid eating pizza whenever I crave it.

According to Williams, categorical desires play a much more important role than simply serving as rationalizations for actions. After all, one could just say that it is in my best *interest* to be healthy, and that there are many actions and decisions that lead toward the overall goal of being healthy, keeping a good diet being only one of them. This may be understood as an evidential or objective truth about what it means to be healthy. My hedonistic tendencies orient me toward seeking pizza, maybe all the time, but the medical evidence I have provides me with good reasons to avoid eating pizza constantly. This decision is conditional on my evidence—there need not be anything categorical about my desire to be healthy. It may simply be that I absolutely love pizza, but I also hate not being healthy and so I need to give up pizza in order to be healthy. In my hierarchy of desires and needs, my need to be healthy prevails over the satisfaction of more immediate but overall less important needs and desires. Ranking my preferences in this conditional way, on evidence and personal utility, suffices for making rational decisions.

For Williams, however, categorical desires cannot be reduced to such a preference-based or evidential analysis. Categorical desires are the basis of what makes our lives *worth living*, and this is the reason why Williams thinks eternal lives are not worth living because the contents of categorical desires expire after a certain point. They are also the reason why death is bad, according to Williams, because death would prevent me from fulfilling the desires that give meaning to my life—the desires without which my life would not be worth living. Put differently, I cannot desire to live without categorical desires, with contents that are *worth pursuing*. Such a life seems actually inconceivable. How can I desire to live a life that is completely vapid and meaningless?

The activity of pursing a good life has been at the core of all major spiritual and philosophical traditions. We pursue categorical desires because they make our life worth living, rather than merely pleasant or efficient at various points.

Satisfying categorical desires or needs is particularly salient to autonomous agents who pay attention to the value of their life as a whole. On the one hand, there is the pursuit of pleasure and happiness for their own sake; on the other hand, there is duty and what is worth doing, even if we find it unpleasant. Rationality must integrate what we enjoy with what is worth doing. Since not all of our decisions are like choosing between pizza and sushi, then some choices must guide all these first-order preferences in a more fundamental way. These more fundamental choices are constitutive of our dignity and autonomy. I can choose pizza over sushi simply based on desire, then give up on both and get tofu instead out of my concern for animals and my own health. Duty starts entering the picture once we start organizing our first-order preferences in terms of normative needs concerning what we *should* do. But a life lived strictly out of duty seems also undesirable. In some extreme situations, the duty to preserve one's life, presumably a need that guides all of our preferences, can be violated for the sake of saving a loved one. This could be construed as rational, but it would also be odd to suppose that saving a loved one affords no pleasure or emotional engagement. Duty and pleasure need to be virtuously balanced.

Susan Wolf (2010) expounds on the importance of finding the harmony between what we enjoy doing and its *real value* or worth. We need to attend to the world through our active engagement, or the state of being enthusiastically under the grip of what we are doing, which need not be pleasant. But clearly, some powerful motivation has to be in place for us to be actively engaged. According to Wolf, our active engagement gives meaning to our lives through projects of worth. Surely, the worth of a life project cannot be merely defined by utility preferences. There must be something objective (or quasi-objective since values are not brute facts), rather than merely personal about whether or not our life projects are worth pursuing. Conditional preference rankings by themselves cannot suffice to articulate projects of worth. More important, the need to avoid a mismatch between our personal preferences and living a life worth living is certainly a moral and agential need—a rational need of a higher order—which only agents that cherish their autonomy can deeply value.

But who decides what is worth doing? This problem is central to moral theory and this is not the place to delve too deeply into it, let alone try to solve it, lest we stop talking about AI altogether. Suffice to say that some kind of objective, "mind-independent," or not strictly subjective standard, however conceived, is essential to any solution to the value alignment problem. A variety of views about value are offered in philosophy: (a) *non-cognitivism* is the view that values don't even have contents we can represent and that value-discourse is "empty";

(b) *nihilism* is the view that the contents of values cannot be satisfied or be true; (c) *idealism* states that the semantic contents associated with values are satisfied by mind-dependent conditions; (d) *naturalism* about value says that the contents of value are satisfied by mind-independent entities, but that value is reducible to more basic entities or natural kinds; (e) *transcendentalism* states that values are real, their contents are satisfied by mind-independent conditions, they are irreducible to other more basic facts, but they are not causally relevant or related to (or "networked" with) other properties; and (f) *robust realism* affirms all of the commitments of transcendentalism, but also asserts that values are causally efficacious or "networked" (this classification is based on a chart presented by Oddie, 2005, 23, which identifies various degrees of value-realism).

Oddie (2005) defends robust realism about values. Given that it is the strongest form of value realism, there are many controversial and counterintuitive aspects of his view. In particular, the thesis that values are properties with causal powers that are not reducible to other more basic causal properties, such as psychological or neurological properties, seems problematic. An analogous issue in philosophy of mind is that typically mental states are causal because they are physical, and they have properties that can in principle be understood as physical processes. If mental properties are entirely irreducible to the physical world, we start leaning toward dualism or panpsychism. Not that there is anything intrinsically incoherent or absurd about these views, but most scientists would consider at least dualism as a non-starter, and most psychologists and neuroscientists would consider dualism and panpsychism as incompatible with their methods and scientific perspective. It would be good to have a theory of value that is not immediately dismissed by the scientific community.

A point made earlier is relevant here: we are not seeking to arrive at a metaphysically grounded account of the universal existence of value. What we want is an account of value that creates robust enough consensus to facilitate value alignment, and which is compatible with the psychological capacities of agents. With this caveat in mind, let us assess how plausible is the view that values are causally relevant. Asking this question in the context of the distinction between categorical and conditional desires is particularly illuminating. My conditional desire for pizza at this moment may be entirely mind-independent: it simply appears to me that pizza is delicious and I have evidence that many other people agree. But then I have the categorical desire to help reduce the pain of sentient beings and the desire to stay healthy to help my loved ones. These convictions can also be strongly desired, as Williams said, categorically and unconditionally. What would it mean to say that I pursue my categorical

desires with full knowledge that they are not objectively valuable? What would it mean to say that it is not really the value of my desire to care for my loved ones, myself, and other sentient creatures that *causes* me to act in these ways? It would be a kind of performative contradiction. I pursue this desire above all others, unconditionally, but I don't know if it really is worth anything, and even if I did know, it would still be incapable of causing me to act. There seems to be a conceptual necessity here—categorical desires entail some degree of value realism.

Oddie argues that even idealism is robust enough to account for the relation between different types of desires and preferences. But idealism faces many difficulties, including the lack of external reference to values and their lack of causal relevance, which is why Oddie opts for realism. The key issue that demonstrates the superiority of realism, according to Oddie, is the comparison of value appearance judgments with perceptual ones. In philosophy of perception, an assumption authors frequently appeal to is called "phenomenal conservatism"—the view that how things appear in conscious awareness justifies our beliefs about them; how things appear provides a basis for how things are. In Oddie's own words: "*there is a non-zero chance that seemings are evidence for the way things are*" (Oddie, 2005, 53; his emphasis). This commitment goes against the reductive approach to value as refined desire: "If we combine value idealism with the experience conjecture—the thesis that desires are experiences of value—then what we have is close to the rather familiar thesis that value reduces to desire" (Oddie, 2005, 83). But there is more to value than desire can capture (a problem Oddie calls the desire-independent value residue). One of the fundamental aspects of this problem is that values elicit responses and guide our entire lives by causing us to do so independently of the specific arrangements of our desires at any point in time.

There is no need to endorse all of these strong views about value realism. The main point is that something must play the role of value-referent in order for value alignment to take place. We cannot possibly understand how value alignment may occur with AI unless we understand how it occurs among humans. Following Oddie, a comparison with "attention alignment" or "judgment alignment" in perception may prove useful. Suppose we are walking in the park and we see a squirrel. We have our perceptual appearances to go with as starting point. We point at the squirrel and attend to it. Typically, the reason why it is not a mystery that our judgments about the squirrel are aligned is because there really is a squirrel there, and we are jointly attending to it. Our judgments and attention are aligned because the content of our experiences

matches reality—there is more about what we see than what we subjectively perceive in our inner awareness. There is a basis for alignment that goes beyond our subjectivity. Attention is essential to understand how this happens in human psychology. In the case of value, the realist argues, our desires are aligned with what is worth doing because there is something mind-independent about its worth. The question is how exactly this could be, since there is no equivalent to the squirrel in the case of value.

There are two kinds of luck-reducing conditions for value and judgment alignment. Both kinds of alignment depend on joint attention—it is a kind of attention alignment when we value the same things through our joint desires and perceptual experiences. One kind of luck-reduction is captured by Oddie's statement that "there is a non-zero chance that seemings are evidence for the way things are." This is, surely, too permissive. We want something way above chance to count as luck-reducing. It should not be by luck or chance that we both end up pointing to the squirrel, or that the content of our sentences is the same when we talk. Perception and linguistic communication are action-conducive and action cannot depend on lucky guesses. Perhaps values do not require such high standards of non-risky action, but some threshold above chance is needed to solve the value alignment problem. Joint attention can solve this problem through its reliable, selective, and inhibitory functions. Reliability is the first condition a successful agent must meet for non-lucky alignments.

The other kind of risk and luck-reducing condition concerns virtuous integration. If all we want is to locate the squirrel, it suffices to focus our attention on it. But chances are that we are doing something more interesting than simply locating the spatiotemporal location of the squirrel—we may be doing this in the context of a larger conversation. We may be saying things about the squirrel that we find funny because we know each other well or we may be biologists and want to identify what type of squirrel we are looking at. Our first-order attention routine is a mental action that is embedded in a larger *attentional project*: a friendly walk where we are keeping a fun conversation or a research journey in which we need to classify animals. Both kinds of risk-reduction—safety in determining *what* we are looking at and *why* we are looking at it—are guaranteed by the functions of attention routines. The first by the virtuous selection of what we need to attend to in the environment; the second by the virtuous integration of the first-level attention routine with other attention routines that are relevant to complete a larger project. What we are looking at and why we are looking at it are aligned through joint attention.

Similarly, with desires and values, our desires may be aligned by virtue of the values we jointly attend to, and the reason why we attend to those values is because of our *need* to pursue "projects of worth." The satisfaction of needs that are similar is also part of the explanation of attentive alignment—we not only have similar attention routines; we also have similar cognitive needs. Attention reduces risks of misalignment partly because it satisfies similar cognitive needs. It is no accident that our values tend to align, just like it is no accident that we perceive the world in similar ways. Attentive risk reduction explains why we not only pay attention to the same contents, but also why we have the same reactions to them, given a context of action in which a larger project is salient. Attention to what is relevant in a properly integrated context is a vital component of seeking and maintaining projects, short term and lifelong. The first-order desire to maintain our conversation lively is part of a higher-order desire that shapes various projects we value: keeping our friendship strong. Friendship is, presumably, the content of my first- and higher-order desires, in a way that the organization of many other desires I have is guided and regulated by the objective value of friendship through a process that Oddie (2005) calls "refinement." The present contribution is that process of refinement in which our first-order desires get organized for the sake of our higher-order ones depends upon the selective and integrative capacities of attention.

Joint attention that is referential (*de re,* or about the "thing itself" rather than about what is said about it) is anchored in mind-independent entities or features. Categorical desires, the value-realist affirms, are aligned with *really* worthy projects. This value-realist thesis may be too strong, but some kind of alignment, as mentioned above, is needed if value is not to collapse entirely into subjective desire. Thus, it may be considered as a principle of rationality that our categorical desires, which make our life worth living, should be conceived of as real or objective, even if a full explanation of their mind-independence is not readily available. If so, expressions that refer to value as real are not purely metaphorical or illusory. Otherwise, the relation between active engagement through our lived experiences and desires and projects of worth cannot be established (using Wolf's terminology; see also Srinivasan, 2020).

Active engagement, at its height, is guided by a kind of attention called "post-voluntary" or "flow" attentional states in which one pays attention to very complex information without the experience of personal effort. In fact, one does not experience "oneself" in these circumstances of full engagement or highly dexterous behavior.[1] This seems to create a paradox: How can one keep track of one's preferences and values if one loses a "sense of self" when engrossed

and engaged while pursuing these activities? The answer to this difficulty is that what allows us to virtuously integrate such a complex action (e.g., piano playing or ballet dancing) is the experience of pleasant familiarity that we have when engaged in the activities we value the most. A life lived solely on the basis of "reasons for action" may feel wholly unfamiliar and, therefore, not worth living. Familiarity explains why our engaged lives need not be the result of judgments concerning self-awareness. This issue is explored in Chapter 5.

Value alignment, in its most engaged and meaningful form, occurs through empathy. But the process of aligning and organizing our preferences more generally is considerably more intricate than what can be achieved through empathic attention routines. While categorical desires seem to involve values that we must recognize as objectively worthy through our experiences of them, preferences come in many varieties, and because of this, there may be partial misalignment or conflict as our overall preferences adapt and change. One way of appreciating this point is by looking into the literature on personal identity. Familiarity with ourselves depends on our unique narrative, which has a deep impact on our preference-rankings and memory structure, particularly autobiographical memory (Montemayor, 2018). Personally salient and autobiographical narrative-dependent preferences are quite stable, such as caring for our loved ones or pursuing a long-term career path that we have chosen. Marya Schechtman (1996, 2014) argues that our personal narrative makes different events in our lives, which are in principle completely unrelated, meaningful and informative in a personally salient way (although such a view about the narrative self is contested; see Strawson, 2008). Derek Parfit's book *Reasons and Persons* (1984) is perhaps the most influential treatment of the complex relations between preference-ordering, decision-making, rationality, and personal identity, showing that there is a multiplicity of problems regarding their integration, from epistemic guidance to moral behavior, that challenge any one-dimensional approach to what is now called the value-alignment problem.

Our deepest commitments are framed by an autobiographical narrative-structure, and this structure is a fundamental component of autonomous intelligent behavior. Value alignment makes sense only among autonomous agents who are deciding under uncertainty and have desires and needs that they must satisfy. Aligning tool-AI with our values is aligning our machines with what we value, just like we align other artifacts and organizations with what we value. It is our values that matter here, because we are autonomous agents. Going back to the issue of enfeeblement, using one of Russell's examples, I may have a very strong desire to be at the top of Mount Everest as one of the essential goals

in my life. But if I have a subservient AI that doesn't understand that what makes this goal worth pursuing for me is the challenge of doing it myself by climbing this famous mountain, then the AI will unintelligently satisfy the goal without integrating and making more salient this higher-order personal need. The result is that the AI takes me in a helicopter to the top. There is a clear sense in which I will be deeply dissatisfied. The puzzled AI may say that the goal was satisfied and that otherwise I would have risked my life, be in pain, suffer cold—"come on, are you crazy?"—the AI may ask. Russell indicates that the AI is misaligned here because we value doing things for ourselves. Indeed, any genuinely intelligent agent has a *categorical need for autonomy*. If AGIs become genuinely intelligent, they will have autonomy, but as explained before, at a considerable risk for us.

Russell argues that there is another problem in the vicinity: the *King Midas problem*. This problem, which is more explicitly related to value alignment, is based on "the impossibility of defining true human purposes correctly and completely. This, in turn, means that what I have called the standard model—whereby humans attempt to imbue machines with their own purposes—is destined to fail" (Russell, 2019, 137). We run the risk of "imbuing" machines with objectives that are imperfectly aligned with ours and the lack of understanding of what autonomy is on the part of machines is certainly a very critical problem. Misalignment of value in machines has three sources, all related to attention: lack of autonomy, lack of sensitivity to what is salient, and lack of virtuously integrated preferences and goals. By designing machines that "satisfy our goals" without the proper functioning of attention, we may satisfy them immediately, turning them into "AI gold," but at the cost of paralyzing our own autonomy. The King Midas problem resides in the specific nested hierarchy of preferences that humans virtuously integrate in exercising their autonomy, which is irreducible to any specific set of explicit preferences. As Russell insightfully notes, humans solve this problem by using *their own cognitive architecture* (2019, 233), specifically, how and why humans pay attention to goals and contents. Russell writes:

> Machines do not have this advantage. They can simulate other machines easily, but not people. It's unlikely that they will soon have access to a complete model of human cognition, whether generic or tailored to specific individuals. Instead, it makes sense from a practical point of view to look at the major ways in which humans deviate from rationality and to study how to learn preferences from behavior that exhibits such deviations.
>
> One obvious difference between humans and rational entities is that, at any given moment, we are not choosing among all possible first steps of all possible future lives. Not even close. Instead, we are typically embedded in a deeply

> nested hierarchy of "subroutines." Generally speaking, we are pursuing near-term goals rather than maximizing preferences over future lives, and we can act only accordingly to the constraints of the subroutine we're in at present.
>
> (Russell, 2019, 233)

Human preferences are embedded in a nested hierarchy of subroutines—these are attention routines guided by representational needs, preferences, and values. Humans overcome their limited information-processing capacities by being attentive. It is precisely because AI development depends so fundamentally on a better understanding of human cognition that the *attention approach* to intelligence defended in the rest of the book is necessary. Rational communication, value guidance through cooperation, and preference ordering all depend on attention routines. In sum, human rationality depends fundamentally on three types of needs: autonomy, representational, and integrative needs. All of these needs require attention for their proper satisfaction. Attention includes implicit and explicit processing (what Kahneman calls systems 1 and 2 reasoning), and there are two dissociable types of attention—phenomenally conscious and unconscious or "access" conscious—that delineate moral and epistemic agency. To fully understand the hierarchy of needs that are relevant for human psychology, a brief survey of the distinction between biological and cognitive needs is necessary. This is the topic of the next section.

2.4 A Hierarchy of Needs

This section classifies various needs that are essential for intelligent agents, focusing on four basic needs. It defends and explains a hierarchy of needs in combination with attention routines that satisfy these needs. The work of Abraham Maslow on psychological needs and motivation is introduced and examined. The problem of ranking values in accordance to a hierarchy of needs is discussed in the context of AI development. It introduces the Authority and Utility-Value Mismatch problems, each of which correlate with AI social risks.

The attentional approach defended thus far differs from extant approaches to AI in two key respects. First, it tackles risk in terms of capacities of autonomous agents. Second, it does not ascribe moral or epistemic status to AI based on "charity" rules (e.g., plants or animals are included or not) or generic processes or properties such as reflective judgment on norms (Kantianism) or sentience-based

phenomenal consciousness (Utilitarianism). On the present account, moral and epistemic standing is based on the autonomy agents have as sources of skills that satisfy their needs because of their agential competence, without deterministic or "law like" external guidance, which is the basis for responsible behavior. Agents are their own sources of luck-reducing success. Thus, to fully understand the virtuous integration of attention routines required for autonomous agency, it is fundamental to understand what are the specific needs agents must satisfy. Agents, unlike machines, are motivated to fulfill their needs by themselves—and their needs are organized in terms of the degree of importance that they assign to their preferences. Here is a basic classification of agential needs:

1. *Representational needs: Agents have the basic need to represent their environment accurately in order to act upon it successfully, which is a condition for the satisfaction of other basic needs.*

All animals have representational needs. They need to feed, navigate their environment, seek shelter, and so on. Categorizing the most fundamental representational needs is a good heuristic for understanding which forms of attention evolved earlier in the evolution of species (Haladjian and Montemayor, 2015). Animals are agents that satisfy their representational needs through their attention skills. The basic types of attention routines in perception include (a) object-based, (b) feature-based, and (c) space-based attention. This is a very broad characterization of attention routines, but it is standard in psychology and neuroscience.

Leibo et al. (2018) from the DeepMind group have started testing a variety of visual attention tasks designed by cognitive scientists in order to compare AI performance in a "laboratory." The types of tasks they have tested are what psychologists consider "bottom-up" basic attention. Their AI system has not fared very well in any of the tasks. But this is not what is surprising about their findings—dynamic environments remain a fundamental challenge to AI and robotics. What is surprising is how well integrated these informational routines are in most animals, from insects to elephants, in stark contrast with any AI. Multiple object tracking, feature recognition, object identification, spatiotemporal location of features and objects, are all basic components of an animal's representation of the environment, and are immediately salient to them as they perform different tasks (see Montemayor and Haladjian, 2015, for details about the classification of attention in psychology, neuroscience and philosophy). Animals reduce risk in their actions through their attention

capacities, with respect to both the satisfaction of referential needs (attention to features and objects) and the integration of different task-relevant contents (the shape, color, location, and trajectory of an object while moving).

All these attention capacities fall under navigational behavior that depends on representing fundamental aspects of the environment. This takes us back to Marcus and Davis' (2019) point that genuine intelligence depends on successfully identifying semantic contents. Here is one of their examples: "If you ask 'How far is the border of Mexico from San Diego?' you get '1144 miles,' which is totally wrong. WolframAlpha ignores the word 'border,' and instead returns the distance from San Diego to the geographic center of Mexico" (Marcus and Davis, 2019, 81). How can a program that specializes in performing complex mathematical operations get such a simple calculation so wrong? The answer is straightforward: unlike animals that have attention routines that evolved in order to adequately represent and *understand* the structure of their environment, AI and programs like WolframAlpha are running on semantically empty categories of bodies of information without context. Thus, representational needs are the most fundamental needs an agent must satisfy in order to succeed. They include basic, bottom-up attention routines, top-down routines concerning conceptual identity and thought-attribution and, in the case of humans, the full range of semantic contents underlying linguistic communication.

2. *Biological needs: Agents have the basic need to maintain their organism healthy and in optimal condition in order to satisfy other needs.*

While biological needs are not exactly needs that require autonomous intelligence, they are needs that a system must satisfy self-sufficiently to ground other kinds of autonomy. Biological needs are not relevant for AI, but analogous needs for the maintenance and optimal functioning of AI would play a similar role. In addition, proper biological functioning is a fundamental condition for the integration of attentional capacities in living organisms. Therefore, while they are not essentially rational aspects of our autonomy, biological needs are preconditions of autonomy. There is, moreover, a direct correlation between lifespan and rationality: The longer the life of an intelligent agent, the more opportunity she will have to embed her actions in more complex and worthy projects. The relation between life and intelligent behavior is certainly not trivial.

3. *Emotional needs: Agents must respond to their emotional needs in order to organize and guide their social activities and goals.*

Emotional needs are at the root of our care toward others, our moral worth and moral capacities. They motivate us and make life engaging—they make possible our "active engagement" with life projects. Classifying emotions is an intricate issue (for instance, see Damasio's, 1994, distinction between feelings and emotions), but regardless of such categorizations, emotions have the kind of nested structure described above, and can be finely tuned and regulated (what Oddie calls the "refinement" of desire). Embedding plans into larger and personally valuable projects is how different needs get integrated into a hierarchy of preferences. Emotion and desire have a solid biological and neurological basis, but they are all regulated for much more than life sustenance in humans. Some of our autobiographically framed emotional needs may be the most important for our identity and our categorical desires, as Schechtman's narrative account of personal identity suggests. We are empathic creatures, but we also align our values socially in terms of narrative nesting. In this way, our narratives match those of our social groups, from our family to national or professional groups.

It is instructive to look at how a famous hierarchy of needs was explained and developed in order to fully appreciate the complexity of emotional needs. Abraham Maslow's (1943, 1954, 1987) hierarchy of needs, classified in five levels, originally started with *physiological* needs at the bottom; then *safety* needs, *love and belonging* needs, *esteem* needs, and finally *self-actualization* needs. According to Maslow, not all of these needs have to be satisfied to completion or satiation for us to be able to "move" to the next level and, crucially, motivation operates differently at the top, or self-actualization level: while motivation decreases as the first four needs are met, motivation increases as self-actualization needs are met. Tellingly, Maslow called the first four needs "deficiency needs" and the self-actualization needs "being, or growth needs." This asymmetry between the fifth and the previous four levels is extremely important to appreciate the need for autonomy that intelligent beings essentially have, which is deeply related to their dignity. Since the asymmetry concerns motivation, it is also fundamentally related to Susan Wolf's notion of "active engagement."

This simple categorization of needs is also compatible with Oddie's "desire refinement" and Schechtman's "autobiographical narrative-construction." For example, Maslow (1987) proposed that most behavior is motivated by several or even all the needs in the hierarchy at a time, suggesting a great deal of refinement and integration, as well as flexibility. It is not the same to fulfill these needs at different points in our lives. There is, therefore, enough structure in the hierarchy to account for interpersonal alignments, without thwarting the inner complexity

and personally salient organization of needs and desires. Maslow's original hierarchy of needs thus seems to satisfy conceptual and theoretical requirements.

But Maslow (1987) revised and extended this list. After "esteem needs" (which incidentally include *independence and autonomy*), Maslow added *cognitive needs* for knowledge and curiosity and *aesthetic needs* for the appreciation of beauty. Interestingly, Maslow placed these needs *before* self-actualization needs—recall that the hierarchy starts from needs that must be met to a satisfactory degree before moving to the next level. Maslow then added *transcendence needs* above self-actualization needs. Transcendence needs involve motivations to go "beyond the self," such as mystical experiences, faith, and contact with nature. It is a curious choice to place transcendence needs above self-actualization needs, but not below knowledge and aesthetic needs. One can ask, can a person be fully rational if their top preferences are to transcend themselves? How, and by what means? This question justifies an independent category for rational needs. It is clear, however, that Maslow's hierarchy of needs can be used to model the process of desire refinement, and that the needs at the top of the hierarchy depend fundamentally on the emotional needs of autonomous agents. The details may differ, but some hierarchy of needs is always involved in *what* content is salient to our attention, and *how* these needs must be integrated. Attention and needs articulate desire refinement. In the case of emotional needs, they do so with the aid of a biologically rooted motivational structure.

4. *Rational needs: Agents have the basic need to provide other agents with reasons for their actions and decisions, in a way that is as consistent as possible.*

One way of achieving consistency is by following the rules of logic and the axioms of probability theory. There are difficulties that emerge from applying this approach systematically, such as the existence of conflicting meta-evidence. But in general, belief must follow a truth norm, according to which one must believe only on the basis of the best available evidence or the likelihood of truth. Rational principles guarantee that truth is preserved consistently in our thinking. A consistency norm must be in place because if one believes contradictions then anything goes. A rational mind systematically updates evidence in a coherent and consistent way. However, a broad set of problems confronts this approach. Some of the most interesting limitations of this idealized model of coherence concern empirical findings about human rational capacities. For instance, utility theory, which is based on logical and probabilistic principles, has been challenged on theoretical and empirical grounds. Decisions that have a large impact on our

lives are too complex to fit into a neat model, and we don't obey the axioms of decision theory systematically, as psychological evidence on decision-making under uncertainty has demonstrated. An assessment of this body of evidence is one of the central topics of the following chapter. The focus now is on another limitation of this idealized approach. The rules of logic and probability, even in the idealized scenarios of formal epistemology, break down when applied to more realistic and *social* contexts. This limitation has obvious implications for AI. Russell succinctly describes this problem as follows:

> The basic idea that a rational agent acts so as to maximize expected utility is simple enough, even if actually doing it is impossibly complex. The theory applies, however, only in the case of a single agent acting alone. With more than one agent, the notion that it's possible—at least in principle—to assign probabilities to the different outcomes of one's actions becomes problematic. The reason is that now there's a part of the world—the other agent—that is trying to second-guess what action you're going to do, and vice versa, so it's not obvious how to assign probabilities to how that part of the world is going to behave. And without probabilities, the definition of rational action as maximizing expected utility isn't applicable.
>
> (Russell, 2019, 27–28)

Russell goes on to describe how even if you apply game theory, which offers a formal solution to this problem, other difficulties internal to game theory emerge, such as the prisoner's dilemma and the tragedy of the commons. Russell identifies this as an important problem for beneficial AI, because rationality for more than one person must underlie the satisfaction of our communal rational needs and expectations—this is an essential assumption of any solution to the value-alignment problem. In addition, some degree of "irrationality" seems essential to cooperation. As explained above, mutual cooperation depends on joint attention to "illusions" that permit widespread coordination, as Harari points out, like money, states, gods, and numbers. More than any idealized principle or theory of rationality, joint attention is critical to guarantee that we coordinate our actions with respect to referents in the environment and to joint tasks, goals, and projects. This is the importance of unconscious attention: it allows us to embed multiple attention subroutines into a larger project that is the main focus of our attention.

Cooperation requires joint attention, and without it, no solution to value-alignment is possible. But the value-alignment problem gets a lot more complicated once the four needs just described are interpreted in terms of their

value. Epistemic values require consistency and truth, but moral and aesthetic values depend on our emotional needs. If they enter into conflict, which should we choose? There are two distinct problems here. The first is what can be called the *Authority Problem*—the problem of who decides what counts as value, how is value supposed to be defined, and what determines a specific ranking? Should conflict emerge between values, which of them should prevail? About this value problem, Russell says:

> "Whose values are you going to put in?" "Who gets to decide what the values are?" Or even, "What gives Western, well-off, white cisgender scientists such as Russell the right to determine how the machine encodes and develops human values?" I think this confusion comes partly from an unfortunate conflict between the commonsense meaning of *value* and the more technical sense in which it is used in economics, AI, and operations research. In ordinary usage, values are what one uses to help resolve moral dilemmas; as a technical term, on the other hand, *value* is roughly synonymous with utility, which measures the degree of desirability of anything from pizza to paradise. The meaning I want is the technical one: I just want to make sure the machines give me the right pizza and don't accidentally destroy the human race […] To avoid this confusion, the principles talk about human *preference*s rather than human *values*, since the former term seems to steer clear of judgmental preconceptions about morality.
>
> (Russell, 2019, 177–8)

The Authority Problem poses a substantial social risk. Why should we create AI that is "aligned" with the preferences of, presumably, venture capitalists, engineers, and computer scientists, rather than aligned with what is valuable for all humans? Moreover, shouldn't a preference-based approach have at least some degree of moral justification? But what moral justification could it have if all alignment depends on the technical notion of "preference," which ignores moral considerations. This shows that the Authority Problem is not an industrial or technical difficulty. It is a problem about how individuals socially align their values, and who has authority in determining alignments. The Authority Problem highlights the urgency of specifying how could such alignment occur in a democratic, fair, and open fashion. "From pizza to paradise," our needs and preferences get muddled and unorganized if there is no democratic way of assigning value rankings, generating serious risks regarding authoritarianism and paternalism. Which needs must be most salient? What hierarchy should be chosen? A life in which pizza gets the same treatment as paradise can surely be dizzyingly incoherent and without meaning—in fact, it may not be livable. What

kind of value-alignment could possibly emerge from such precarious and paltry assumptions of what makes our lives worth-living? Clearly, this is not a question that AI scientists should answer, but it certainly is a question that requires an answer if the value-alignment problem is to be properly addressed.

There is a second and related challenge, let us call it the *Utility-Value Mismatch*. Suppose that each person has a unique set of preferences and that AIs should be subservient to each of them and no one else. In the larger scheme of things, what guarantees that any of these preference-sets will align? The risk here is one of selfish competition through our individual AIs, jeopardizing joint attention and cooperation, which are essential ingredients of value-alignment among humans. Moreover, what safeguard or principle could guarantee that preferences don't align in a sadistic and ultra-exploitative way unless some real value of the moral type is in the mix? People's online preferences align at pizzas, diapers, porn, videogames, and similar consumerist preferences. But as Maslow's revised hierarchy of needs shows, pizza-needs are not the same as aesthetic-needs, the latter are higher and more refined; but who decides where aggregate preferences should align?

The Authority Problem focuses on *who* decides; the Utility-Value Mismatch centers on the *what* values are higher and *how* we should decide this? Both problems are deeply related. Clearly, "care-needs" to help our family members, eliminate the suffering of the poor, eradicate racism, and prevent the injustice of imprisoning or punishing the innocent are very unlike pizza-needs. Enjoying porn, pizza, getting good diapers for our children, and making them ultra-competitive, to give some examples, are all global preferences that the internet and global economy has now made almost omnipresent. Are these good alignments? These goals and preferences are all quite literally part of our lives' projects. But clearly some of them must be somehow, and in principle, more valuable than others. Hume might have thought that rationality has no power over our passions, but he would not deny that some passions are *better* than others. A life driven by sadism and entertainment may be perfectly modeled in terms of a set of consistently aligned preferences, but such a life can become entirely misaligned, one would think, with a meaningful and valuable life that is worth living. These two problems constitute significant value-alignment risks.

To summarize, the lack of representational needs in AI generates *epistemic risks* concerning the alignment of contents, truth, and epistemic trust. The lack of emotional and rational needs in AI generates *moral, social, political, as well as epistemic risks* regarding value-alignment at a more comprehensive scale. These risks can have a global impact on human well-being and dignity.

2.5 Needs as Sources of Cognitive Plasticity and Complexity: The Measure of Intelligence

This section develops and expands the conceptualization of intelligence and autonomy in terms of the needs of an agent who is intrinsically motivated to satisfy these needs attentively. The importance of needs as a measure of intelligence is explained in the light of the work of Julien Offray de La Mettrie and Margaret Boden.

As was argued above, there is a very close relation between attention and agency. Agency requires autonomy—the non-accidental and self-sufficient satisfaction of needs. An agent's needs are satisfied because of the agent's skills. These skills, in the context of human and animal psychology, are attention routines that are selectively and automatically responsive to the agent's needs. Thus, attention is the paradigmatic kind of mental agency. The structure of attention into routines and subroutines, in accordance to a hierarchy of needs, has critical implications for AI design. In his book on human compatible AI, Russell writes: "Understanding human action, then, seems to require understanding this subroutine hierarchy (which may be quite individual): which subroutine the person is executing at present, which near-term objectives are being pursued with this subroutine, and how they relate to deeper, long term preferences" (Russell, 2019, 234).

A lot of our behavior is guided by long-term preferences that are barely articulable to ourselves and others, yet our attention is strongly guided by them, even if their influence is implicit. Attentive subroutines could be guided by anger, sadism, faith, and they may seem the *same externally*: our behavior might look identical, but the needs and motivations involved in their satisfaction might be radically different. Given two instances of the same behavior, the agent might be intending different courses of action and attending to different contents. This suggests that there are serious limitations to understanding human motivations merely in terms of behavior patterns—a difficulty that is related to the problem of *underspecification in machine learning* (more about this below). Behavior is indeed significantly correlated with our goals and desires. But deep-rooted motivations of the kind that lead our lives and make it worth living need not match any specific set of behaviors. Things are actually much worse because of the implicit character of these deep and life "transforming" experiences and desires (see Paul 2014 and Pettigrew, 2015, for discussion on whether or not this is a fundamental limitation of decision theory). Preferences align very differently depending on how this kind of experience affects the values of agents.

Russell uses the well-known example of tasting durian fruit for the first time in describing this problem:

> One obvious property of humans, if you think about it, is that they don't always know what they want. For example, the durian fruit elicits different responses from different people: some find that "it surpasses in flavor all other fruits of the world" while others liken it to "sewage, stale vomit, skunk spray and used surgical swabs." I have deliberately refrained from trying durian prior to publication, so that I can maintain neutrality on this point. I simply don't know which camp I will be in. The same might be said for many people considering future careers, future life partners, future post-retirement activities, and so on.
>
> (Russell, 2019, 236)

Our standing beliefs and categorical desires may refine our first-order desires, but life is worth living to a large extent because it is surprising. We seek novelty, as long as we don't endanger ourselves (too much); this is part of the process of *learning how to learn*. Too much assistance from beneficial AI would enfeeble us. Too much prediction of our behavior would make us mechanical. Too much uncertainty is incompatible with the satisfaction of our needs. We have a *need for novelty* and curiosity that is sated through the safeguards of properly functioning attention. The regulatory guidance and integrative virtues of attention are essential here as well. Too much curiosity can lead to vice or harm, epistemic and moral; too much fear of the new can lead to inaction (see Fairweather and Montemayor, 2018, for an analysis of virtuous epistemic curiosity). These are two kinds of epistemic limitations, what Russell calls the epistemic uncertainty of not knowing what an experience will bring to our lives—is it going to change us, make us better, or lead us astray—and the uncertainty of having radically underdescribed choices.

Part of the explanation of how attention makes our curiosity safe depends on its virtues of integration. More precisely, the attentive integration of what is relevant for coordinating action through joint attention depends on hypotheses and over-hypotheses about which of the first-order hypotheses are more useful or reasonable (Russell discusses this in the context of the work of Nelson Goodman; 2019, 85). Some routines are deeply "entrenched" in our behavior, to use a term from Goodman. These routines help explain which terms and descriptions are more useful or productive than others. Our deeper preferences nest and regulate short-term plans this way. Any type of value alignment with other agents depends fundamentally on their capacity to interpret our routines and subroutines in meaningful ways. As Russell says, we have the advantage of

having very similar cognitive architecture and so, we can simulate and interpret our preferences. But AIs lack any of these motivations and preferences. In particular, joint attention routines underlie everything that is meaningful in human interactions, from conversations, which depend on attending charitably to expressions and what is relevant (as captured by the "Gricean maxims") to action-coordination.

If our attention routines were not similarly guided, none of our needs could be successfully satisfied, and our values would be hopelessly and dangerously misaligned. Risk reduction depends on need satisfaction, and that is the main role of attention in human and animal cognition. Sets of behavioral patterns are never sufficient to reduce this type of risk. What is needed is *properly motivated* attention to what is salient and relevant about behaviors—praying, dancing, shopping, going to a museum, and so on. Attention provides an *interpretational perspective* of the world, with rich content; behavior patterns can provide at best statistically correlated databases that then stand in need of such an interpretation. AI cannot attend to what is relevant or salient at a context because it lacks any of the cognitive needs mentioned before. To give a concrete example concerning perceptual attention, consider the case of the AI that transformed "The Great British Bakeoff" into a horror show. Janelle Shane trained an AI on data from this show and some images of squirrels, and the result was truly horrific. Completely de-contextualized portions of bodies, bread, hair, eyes with no face, and so on (see Shane, 2019, for why AI makes everything "weirder" because of a complete lack of contextual cues). No properly motivated agent with real needs to represent the environment would do this. To restate a point made before, one can trick humans and animals into perceptual mistake, but never into this kind of completely jumbled chaos.

A hierarchy of needs is essential to structure attention routines. Crucially, such a structure is flexible enough to be welcoming of differences among subjects and their capacities. As long as the essential needs are homogenous enough, this structure will unify a diverse group of agents regarding their needs and interests: animal, human, and perhaps one day, artificial. Interestingly, the role of rationality in the creation of value rankings might be essentially limited. In particular, rationality by itself may not be capable of ranking which needs are at the top of this hierarchy. Think of Søren Kierkegaard's assessment of what he considered to be the deepest and most important of all needs—to be spiritual and have faith "on the *strength of the absurd*." If Maslow is right, the satisfaction of this most "irrational" of desires is the culmination of a life's goal. Whether it is irrational or arational is inconsequential. As long as an agent pursues her life

with this goal in mind, other attention subroutines will be organized accordingly in her hierarchy of needs, independently of utility evaluations. Kierkegaard's assessment provides an insight into how we *experience* our needs. We also have an agential and experiential *familiarity* with our needs—they cannot simply be the result of a rational calculus. On the one hand, the more we satisfy our preferences based on optimal rational choice, the less familiar our alignments might become. On the other hand, the more we base value on familiarity alone without reliable preferences for optimal behavior, the more prone we will be to utility-based failure and risk. Again, the integrative virtues of attention are needed.

Our fascination with AI originates in our admiration for intelligence in general. From its inception, AI was conceived as a kind of mirror of our minds, which are, after all, part of a mechanical universe. The French materialist Julien Offray de La Mettrie emphasized the mechanistic nature of the human mind and explained the necessity to accommodate a mechanical view of the mind into our philosophical theories—a task that remains central in physicalist and reductive views of mental processes: computational (functionalists), behavioristic, and neural (type identity views). Noam Chomsky writes the following about La Mettrie's approach in an endorsement to an English translation of La Mettrie's *Man a Machine* and *Man a Plant* (first published in 1747):

> La Mettrie's inestimable contribution was to draw the natural conclusions from Cartesian physiology and Newton's radical revisions of traditional mechanics: that thought is a property of organized matter, on a par with electricity, the faculty of motion, and others—that mind is to be studied in the framework of the emerging scientific naturalism of the day. His achievement, long unrecognized, merits careful attention today as the problems that engaged seventeenth-and eighteen-century thinkers are again becoming the topic of serious scientific inquiry.
>
> (Chomsky, 1994)

The reason why La Mettrie is relevant here, in a discussion about the importance of needs, as necessary conditions for autonomous agency and intelligence, is because in spite of his strong commitments to naturalism, La Mettrie considers needs as establishing the fundamental demarcation between unintelligent and intelligent life. Mechanical human is not merely physical automata—motion plus general laws of physics. Whatever intelligence is, needs are *constitutive* of it. Our needs, in G. W. Leibniz's terms, *incline without necessitating*. We are part of

the physical world, but our needs make us autonomous and intelligent. Here is La Mettrie in his own words:

> Plants are rooted in the earth which nourishes them. They have no needs. They fertilize themselves. Lastly, plants are immobile. In sum, plants have been seen as immobile animals that lack intelligence and even feeling.
>
> Although animals are mobile plants, one can consider them as being of an entirely different species because not only do they have the power to move themselves, movement that costs them so little that it even enhances the *health* of the organs on which it depends, but also animals feel, think, and satisfy a multitude of needs with which they are besieged.
>
> The reasons for these variations are found in the differences themselves as indicated by the following laws:
>
> First, the more needs an organism has, the more nature gives it means for satisfying them. These means are diverse degrees of sagacity known as instinct in animals and soul in man.
>
> Second, the fewer needs an organized body has, the less difficult it is to nourish and raise, and the less its share of intelligence.
>
> Finally, it follows from the previous two last laws that beings that have no needs have no minds.
>
> (La Mettrie, 1747|1994, 85)

Perhaps all plant behavior is just "growth" (but see Figdor, 2018; Segundo-Ortin and Calvo, 2019, for criticism of this "standard" demarcation). But even here, plants are autonomous at least as agents capable of satisfying their biological needs; "they fertilize themselves," which is a complex task that demarcates plants from "intelligent" robots and all mechanical artifacts. The delineation between purely biological needs and the other three needs described previously (representational, emotional, and rational) is, according to La Mettrie, the boundary between life and intelligent agency. What is important about the three laws of La Mettrie is that the number and complexity of needs correlate with intelligence. A machine without needs, even if it is biological (like plants), is a machine without a mind. According to these laws, no AI has intelligence—only AGI, if autonomous, can have intelligence, because the needs of the AI agent must be satisfied in a self-reliant and integrative manner. Humans are "plants," according to La Mettrie, insofar as they satisfy their sustenance and biological needs; they are animals by satisfying their representational, emotional, and rational needs. According to the hierarchy of needs discussed above, we are more than animals when we satisfy our needs for beauty and the good life.

Need-satisfaction makes humans intelligent. For non-biological machines to be equally intelligent, they must satisfy similar needs autonomously, and not by predesigned simulacrum.

Russell's diagnosis of what went wrong with our current standard for AI design is that we simply applied the human standard of intelligence to machines: "*Machines are intelligent to the extent that their actions can be expected to achieve their objectives*" (Russell, 2019, 9). The revised version for the new paradigm Russell defends is that "*Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives*" (Russell, 2019, 11). La Mettrie's laws are incompatible with this revised version of intelligence and they confront us with the risks about autonomy explained above. Beneficial AI deprives AI from having needs they must autonomously satisfy. Once AI becomes beneficial they lose their autonomy. Applying La Mettrie's third law: *once AI becomes beneficial they have no mind of their own.* Subservience is a form of automation of *our* needs—and industrial automation is not the same as intelligence. *Automation without autonomy is strictly mechanical*. This is the paradox of beneficial or subservient AI (plus the substantial problems of value-alignment mentioned before).

As mentioned in the introduction, this book is not a metaphysical treatise on free will. However, presenting this issue with a metaphysical gloss is helpful in understanding just how important needs are in the generation of autonomy and intelligence. The laws governing biological machines are not just mechanical—they involve the intentional satisfaction of needs by autonomous agents, on the basis of their capacities. G. W. Leibniz thought that a biological model of machines provided a more adequate framework for metaphysics and the sciences (Smith, 2011). A world full of autonomous agents with needs to satisfy is a lot more *interesting* than a world of rigid clockwork and algorithmic instructions. Unlike industrial automation, automation in animals and humans is the expression of the autonomous and intelligent satisfaction of their needs. La Mettrie, who called these automatic expressions of autonomy "springs of action," writes: "Let us consider the details of these springs of the human machine. Their actions cause all natural, automatic, vital, and animal movements. Does not the body leap back mechanically in terror when one comes upon an unexpected precipice?" (La Mettrie, 1747|1994, 62) Springs of beauty and unity of action are examined under a similar light, as expressions of vitality. Roughly a century after the work of La Mettrie, reflexes and their role in cognition, as automatic responses to cognitive needs, would play a major role in the scientific debate concerning the localization of cognitive processes in brain areas, which laid the

foundations of the field of contemporary neuroscience, as well as the notion of the "unconscious" in psychoanalysis (Guenther, 2015). As the next chapter explains, many of our cognitive needs, including *inferential-rational* needs, can occur, and generally *must* occur unconsciously.

Measures of intelligence have been a constant topic in philosophy and psychology, and there are disagreements about what is the right measure and how to measure intelligence, or which methods for measurement are more adequate (e.g., a theory of utility or of general value, accumulation of knowledge versus problem-solving under uncertainty and time pressure). Despite this disagreement, a constant feature of any measure of intelligence is the satisfaction of needs. An interesting possibility that the hierarchy of needs suggests is that without biological needs, other more complex needs cannot emerge. And since, following La Mettrie, without needs there is no intelligence, then without *life* there is no intelligence. This seems to be exactly the view that Margaret Boden (2016) attributes to the cybernetics movement—that mind necessarily presupposes life. She cites, in support of this view, Hilary Putnam's statement that if a robot isn't alive then it can't be conscious, as well as the work of Hans Jonas and Karl Friston (Boden, 2016, 144). She comments on how this assumption hasn't been proven beyond doubt, and then writes:

> Let's assume, however, that this common belief is true. If so, then real intelligence can be achieved by AI only if real life is achieved too. We must ask, then, whether "strong A-life" (life in cyberspace) is possible. There is no universally accepted definition of life. But nine features are usually mentioned: self-organization, autonomy, emergence, development, adaptation, responsiveness, reproduction, evolution, and metabolism. The first eight can be understood in information-processing terms, so could in principle be instantiated by AI/A-life. Self-organization, for instance—which, broadly understood, includes all the others—has been achieved in various ways […] But metabolism is different. It can be *modeled* by computers, but not *instantiated* by them. Neither self-assembling robots nor virtual (on-screen) A-Life can actually metabolize. […] So if metabolism is necessary for life, then strong A-Life is impossible. And if life is necessary for mind, then strong AI is impossible too. No matter how impressive the performance of some future AGI, it wouldn't have intelligence, *really*.
>
> (Boden, 2016, 144–145)

This is a very important point. Metabolic functions and visceral reactions are, I shall argue, necessary for *phenomenal consciousness*, including experiences of empathy (Chapters 4 and 5 develop this thesis). But there is an ambiguity in

Boden's treatment of this issue. The fact that life is necessary for phenomenal consciousness doesn't mean that it is necessary for *intelligence*. Cognitive needs associated with rational and representational needs can in principle be satisfied by AI. In particular, the kind of attentional inference involved in the satisfaction of many of these cognitive needs is independent from phenomenal consciousness. In fact, as mentioned before, highly skilled attention routines can occur unconsciously in humans, and thus, the kind of inferential attention developed and defended in the next chapter can in principle be instantiated in AI. However, the lack of phenomenal consciousness in AI means that they will not be capable of emotional intelligence as it occurs in humans and animals, creating a considerable misalignment risk.

To conclude, the main point of this chapter is that a hierarchy of needs is essential to understand intelligence, defined as the autonomous and competent satisfaction of *hierarchically organized needs*. Attention is the mechanism humans and animals employ to do this. Thus, the possession of attention capacities is fundamental for the only kinds of intelligence we know about so far, namely, animals and humans. Attention provides, therefore, the best model we have to design AGI. It doesn't matter if an AGI agent is a post-humanist cyborg extended through different nanomachines and distributed around the universe. As long as there is unity to the AGI's actions, attention routines and subroutines must be in place in order to satisfy the agent's needs. Agents must attend, and only agents *can* attend. That is the difference between a strictly causal machine and an agent with needs and motivations.

3

The Attentional Model of Epistemic Agency—The Main Source of Rational Trust in Humans (and Future AI)

3.1 Rationality and Cognitive Trust: Notes on the "Child Machine"

This section introduces the topics of rationality and inference in relation to attention and epistemic agency. It focuses on an illustration of the development of intelligence in an AI agent as if it were a child machine, which was originally presented by Alan Turing.

Epistemic trust is a basic condition for human cooperation. Trusting the testimony of experts and good sources of evidence through joint attention routines are essential aspects of our societies and contemporary scientific practices. This type of trust is also essential for animal communication and underlies our linguistic practices as well (Clark, 1996). Joint trust and joint attention are essential in communication. Trust is a kind of reliance on the skills of an agent to succeed in performing an action. Attention is critically involved in our trusting practices, which reduce risk and misalignment. These achievements are not based on how agents consciously experience content, but on their joint attention capacities. A virtue theoretic account of attention as an excellence of epistemic agents provides the required components for such a theory of trust (Fairweather and Montemayor, 2017).

But trust goes much deeper than epistemic standards alone can capture, which is a lesson learned the hard way in recent forms of online and social media miscommunication regarding biases against science, xenophobia, and racism. Our gut reactions and biases are part of the structure of preferences that guide who we trust. In fact, our implicit heuristics may be the main source of guidance regarding moral and epistemic trust. Even if we stay centered on epistemic skills

concerning evidence and rational consistency, much of the guidance in our inferential reasoning and attention routines is implicit, beyond our conscious reach.[1] This chapter explores this topic in detail: How is it that inferential reasoning is integrated with attention routines in order to guide an agent rationally and responsibly? The next chapter focuses on the broader issue of moral and aesthetic trust (and care) by examining the phenomenology of familiarity.

Of the four needs examined in the previous chapter, epistemic agency concerns mostly representational and rational needs. A good epistemic agent is a good source of information because she represents correctly various contents and has the capacity to form beliefs that are more likely to be true than false. This is an epistemic agent we can trust, regardless of what conscious experiences the agent is going through. To recapitulate, representational needs are those agents must satisfy in order to model their environment accurately so that they can successfully act upon it and satisfy other basic needs. Rational needs are those agents must satisfy in order to provide reasons for their actions and decisions in a way that is as consistent as possible. Biological needs and emotional needs are examined in the next chapter.

Two preliminary considerations are relevant before analyzing these issues. First, while this chapter focuses on epistemic agency, inference, and attention, the type of inferential integration discussed here will be crucial to understand the integration of moral and aesthetic needs, skills, and motivations that are the main theme of the next chapter. The difference is in the complexity and structure of the hierarchy of needs, not in the functions and roles of inference and attention. Second, the kind of cognitive integration discussed in this chapter is more properly understood as an aspect of epistemic agency, but it is certainly influenced by moral and aesthetic commitments. A kind of "motivational penetration" is discussed in the next chapter.

Why should inference be so central in the explanation of epistemic agency? The answer is that there is no other aspect of our agency that captures the essence of epistemic responsibility better than inferential reasoning. Inference is the main focus of all the literature on epistemic responsibility and norms of rationality. A rational agent draws good conclusions from the right premises and knows, through inferential reasoning, what follows from the beliefs she endorses. She has consistent responses to other agents based on how she draws the right inferences regarding relevant questions, answers, and reasons. Since the previous chapter demonstrates the importance of attention routines for the integration and proper satisfaction of needs of an agent by the agent herself, this chapter's main goal is to explain the relation between attention and inference in

the specific context of epistemic agency and responsibility. Epistemic trust and responsibility is a pressing topic in AI design. As Marcus and Davis say:

> Trustworthy AI, grounded in reasoning, commonsense values, and sound engineering practice, will be transformational when it finally arrives, whether that is a decade or a century hence. […] AI that is powered by deep understanding will be the first AI that can learn the way a child does, easily, powerfully, constantly expanding its knowledge of the world, and often requiring no more than one or two examples of any new concept or situation in order to create a valid model of it. It will also be the first to truly comprehend novels, films, newspaper stories, and videos. Robots embedded with deep understanding would be able to move safely around the world and physically manipulate all kinds of objects and substances, identifying what they are useful for, and to interact comfortably and freely with other people.
>
> (Marcus and Davis, 2019, 200–201)

In early childhood, humans learn to jointly attend to various features of language, such as syntax, semantic contents or referents, and types of expressions or speech acts (e.g., commands, assertions, jokes). Concept acquisition is key in this development. Psychologists have debated whether the inferential rules required for learning a language are built-in, "computationally" stored prior to environmental exposure, or learned through communicative exchanges. Regardless of whether one favors nativism or external acquisition, it is absolutely clear that we are not simply blank slates that run on massive amounts of data. Inferences about behavior grounded in meaningful representations of the environment, and interpreted contextually, are essential in the development of human intelligence. Children are incredibly skilled at performing these complex learning tasks with few and even not ideal data points or stimuli. Attention routines guarantee the reliability of our inferential reasoning. This makes us trustworthy—we trust in the skills of each other to fulfill specific representational and cognitive tasks. Trust is not blind. We in fact have these skills, and it is not by risky accident that we succeed in satisfying our representational and rational needs, drawing conclusions, and generalizing from meager information. These cognitive accomplishments are the foundation of what Marcus and Davis call "commonsense" and "deep understanding."

A fundamental question regarding the cognitive architecture of inferential reasoning is if, and to what extent, there is *cognitive penetration*, or the influence of higher cognition on bottom-up attention routines. While this issue remains contentious, there is plenty of evidence suggesting that concepts constitute a

crucial informational platform for cognitive penetration (Montemayor and Haladjian, 2017). It is, presumably, through cognitive penetration that simple sensorial signals get transformed into referents for coffee mugs, a sad smile, a virus in a microscope, or a star constellation in the sky. Such processes involve the kind of inference, generalization, and learning that Marcus and Davis emphasize. A cognitive architecture that has higher-level attention routines integrating bottom-up routines is one that provides integration and stability for the satisfaction of more important needs within a hierarchy of preferences. It organizes and prioritizes. But not all the bottom-up routines can be affected in their content or altered in the way they draw conclusions by top-down attention routines (some remain "encapsulated").

AI development can greatly benefit from a psychological and philosophical analysis of inference, attention, and their cognitive architecture. In particular, the meaningful interpretation of behaviors and actions always depends upon inferential and attentional capacities. All types of communication require the correct interpretation of objects, contents, actions, and behavior patterns. These inferential processes underlie concept acquisition and they organize concepts into categories, such as those about animate or inanimate objects, kinds of objects (e.g., artifacts, animals, toys, friends), and sets of features. Inferential relations among concepts are also critical to accurately represent the substrate, causal, and invariant structure of environments (Keil, 1992). Children are indeed a very good model for genuinely intelligent AI, the most general kind of AGI. Alan Turing (1950) anticipated some of the themes discussed by Marcus and Davis, for instance, in the following passage in which Turing emphasizes why conceptual capacities for inferential generalization cannot be produced solely on the basis of constant training based on rewards or sanctions:

> The use of punishments and rewards can at best be a part of the teaching process. Roughly speaking, if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of reward and punishments applied. By the time a child has learnt to repeat "Casabianca" he would probably feel very sore indeed, if the text could only be discovered by a "Twenty Questions" technique, every "NO" taking the form of a blow. It is necessary therefore to have some other "unemotional" channels of communication. If these are available it is possible to teach a machine by punishments and rewards to obey orders given in some language, *e.g.* a symbolic language. These orders are to be transmitted through the "unemotional" channels. The use of this language will diminish greatly the number of punishments and rewards required.

> Opinions may vary as to the complexity which is suitable in the child machine. One might try to make it as simple as possible consistently with the general principles. Alternatively one might have a complete system of logical inference "built in."
>
> (Turing, 1950, 457)

Contemporary machine learning depends, roughly speaking, on the type of training that Turing criticizes in this passage. Machine learning cannot really be learning, if we follow Turing, because "if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of reward and punishments applied." Animals and humans have a biologically based reward system that is essential for empathy and the proper satisfaction of emotional needs. But, as Turing clarifies, rewards and punishments *alone* cannot be sufficient for general intelligence. Inferential reasoning is fundamental for the satisfaction of *rational and representational* needs.

A system of "logical inference" is the most elegant and simple solution to the difficulties that learning presents to the infant. This inferential structure is the scaffolding children use to learn how to learn many other concepts. Inferential reasoning must rely on "built-in" conceptual representations, imperatives or rules for consistency and logical entailment, as well as generalizations that accurately and easily capture the causal structure of the environment without relying on vast bodies of information. With respect to *linguistic* representation, however, not even a powerful combination of symbolic inferential patterns and emotional rewards suffices. It doesn't matter how much punishments or rewards we give to an animal in a tight learning schedule, the poor creature will not learn human language (Nim Chimpsky's story is a dramatic example of this fact). Animals share a lot of their attentional and emotional capacities with us, so this shows how non-trivial linguistic representational needs are. This is why, like in human children, the child machine must have an "unemotional" system of representation that allows for efficient communication and learning prior to the learning process.

"Rewards" in contemporary AI are unemotional simulacra of the biological version. The next two chapters explain why this has important consequences for the lack of empathy and emotional needs in AI, with implications for the value alignment problem. But this handicap has a positive epistemic outcome. Simulating intelligence is not only good, but can certainly count as intelligence precisely because satisfying representational needs requires *unemotional*

information processing. The child machine may not be phenomenally conscious (there may not be anything "it is like" to be her), but her reward system can operate similarly to our unemotional inference-based cognitive system. Contemporary AI is based on an unemotional reward system. What it's lacking is the *inferential framework* required to satisfy representational needs through symbolic structures, attentional salience, automatic action, and hierarchically embedded attention routines. It is theoretically significant that AI has been developed in terms of reward signals. About AI's reward system, Russell writes:

> Instead of a goal, then, we could use a utility function to describe the desirability of different outcomes or sequences of states. Often, the utility of a sequence of states is expressed as a sum of *rewards* for each of the states in the sequence. Given a purpose defined by a utility or reward function, the machine aims to produce behavior that maximizes its expected utility or expected sum or rewards, averaged over the possible outcomes weighed with probabilities. […] For complex problems such as backgammon and Go, where the number of states is enormous and the reward comes only at the end of the game, lookahead search won't work. Instead, AI researchers have developed a method called *reinforcement learning*, or RL for short. RL algorithms learn from direct experience of reward signals in the environment, *much as a baby* learns to stand up from the positive reward of being upright and the negative reward of falling over.
>
> (Russell, 2019, 54–5; my emphasis)

"Like a child," AI based on RL learn from "experience." But very unlike a child, the unemotional AI based on RL have neither emotional needs nor built-in inferential and symbolic-conceptual representations. So, unlike a child, the AI agent is not really learning anything about the environment through inferential generalization. Although there certainly are important similarities between AI and human predictive coding, inferential reasoning cannot simply emerge from learning based on bodies of data, rewards, and punishments. The system must be fulfilling some purpose and satisfying concrete needs in order to engage in reasoning and intelligent behavior. However, because AI is "unemotional," the child machine has advantages that animals lack—one day, for instance, a genuinely intelligent child machine may learn a language, much as a baby does. This is no small advantage. As was just mentioned, no living being other than humans can learn the whole range of linguistic communication. Russell rightly criticizes Hollywood plots about terrifying "conscious" machines when he says that they are "really missing the point: it's competence, not consciousness, that matters" (Russell, 2019, 17).

Now the central questions are: How should we define inferential reasoning in high-level cognition? Can non-conscious representations guide or even determine high-level cognition? If so, what are the properties of such non-conscious representations? Two contemporary debates center on these issues. The first concerns the possibility of cognitive penetration, or the degree and extent to which high-level cognition influences or determines low-level cognition. The second focuses on the epistemic status of conscious cognition, and on whether or not non-conscious cognition could play a similar, albeit not as fundamental, justificatory role as conscious cognition. This latter issue is at the heart of the question concerning the epistemic status of conscious awareness. The child machine cannot empathize the way a chimpanzee can, even though the child machine could potentially be more intelligent than a chimpanzee. Fear and pleasure cannot by themselves be enough for AGI. But does an agent gain something crucial, epistemically speaking, by having inferential reasoning *and* phenomenal conscious awareness? Their integration is necessary for structuring human-like needs into a hierarchy, so this is a critical question. A brief answer to it is that there is an enormous advantage in integrating inferential reasoning with consciousness, namely, that the hierarchy of needs becomes much more complex and stimulating.

But before addressing this issue, a more pressing question is, what should be the epistemic standard required for an inference to count as justificatory? Debates on the epistemic status of consciousness and cognitive penetration typically assume such a standard because high-level cognition is associated with rationality, inferentially structured thought, and the epistemic responsibility one has for the conclusions drawn through one's inferences. The following sections provide an account of *inferential-attention* that explains how cognitive penetration of non-phenomenally conscious cognition and perception is possible, and why there are unconscious processes that should be considered as essential components of high-level cognition.

3.2 Conscious and Unconscious Cognition

This section contrasts two notions of inference, based on restrictive and liberal approaches to the issue of what should count as an inference that provides epistemic justification.

Central questions about inferential reasoning concern its relation to conscious and unconscious cognition.[2] Are there non-conscious representations driving

(or determining) high-level cognition? If so, what are the properties of such non-conscious representations? At a neuroanatomical level, top-down influences from the prefrontal cortex regulate and contextualize sensorial filtering, inhibition, attentional selection, and task relevance, based on prior knowledge (Nakajima et al., 2019). The background knowledge involved in these cognitive influences includes emotional inputs (Song et al., 2017) and determines perceptual processing (Lupyan, 2017). These executive control and decision-making influences on cognitive processing are an important characteristic of high-level cognition, but how exactly should we define them from an epistemic point of view?

A classic approach to these questions is to define the interaction between prior knowledge and cognitive processing in terms of inferential relations (this idea goes back to the seminal work of Helmholtz, 1867/1910). A basic type of inference is involved in top-down influences, such as "if the task is *X*, then ignore feature *y* and select for feature *z*." The inference could be based on predictive models, probabilities, conceptual information, or deductive rules but what matters is that it must satisfy epistemic criteria for it to count as a properly drawn or justified inference. Thus, we need a precise definition of inference to evaluate higher cognition. As mentioned, two salient debates about this issue center on cognitive penetration (Macpherson, 2012; Yeh and Chen, 1999; Zeimbekis and Raftopoulos, 2015) and on the epistemic status of conscious perception and cognition. Both are important to specify the epistemic standard required for an inference to count as justificatory (Siegel, 2017).

A common theme in these debates is the nature or type of inference required for high-level cognition. One strategy is to assume a *restrictive* notion of inference, according to which only conscious reasoning can give grounds for inferential reasoning that has an unquestionable epistemic status. A version of this view is that the kind of reasoning from premises to conclusion that is unquestionably justificatory must be either *explicitly endorsed* by the thinker or somehow *understood* by the thinker *as such—as an inference*, the conclusion of which is accepted explicitly on the grounds that the premises are taken to be true. One option is to characterize explicit reasoning in terms of having cognitive access to the inference at the personal level, or as an agent, without necessitating phenomenal consciousness. An even more restrictive version holds that only phenomenally conscious reasoning can count as epistemically justified (see Boghossian, 2018, for a very useful categorization of inferential reasoning, including processes labeled as "inference," which, according to Boghossian, should not count as inferences).

An alternative strategy is to assume a *liberal* notion of inference, according to which many instances of unconscious, automatic, and yet complex reasoning should count as epistemically justified. Unconscious states that have the cognitive structure of an inference (i.e., they are selectively responsive to inquiries, based on premises or assumptions) count as reason-grounding and justificatory. This view is compatible with cognitive penetration, and it may also have the implication that cognitive penetration is common, perhaps even pervasive. It also entails the more surprising consequence that we are somehow epistemically *responsible* for inferential processes occurring outside the scope of our introspectively conscious, or intuition-based, cognitive control (Siegel, 2017). But it is not surprising that an agent has only explicit access to the most fundamental inferences she needs to pay attention to at any point in time. Other inferences can be embedded as subroutines and still count as justificatory or epistemically trustworthy. If even the most salient inferences need not be phenomenally conscious, a child machine can in principle become an epistemic equivalent of a human, with respect to both inferential attention and cognitive architecture.

These rival approaches have clashed for at least four decades now, and although there has been progress, it seems that restrictive and liberal accounts may require new insights for the debate on inference and rational high-cognition to move forward. This is, obviously, very important for psychology as well, since methodological questions concerning the nature of inference depend on the conceptual clarity with which experimental designs and results are interpreted, which in turn is crucial in developing AI. The key insight proposed here is that attention routines satisfy and empirically explain the epistemic requirements for inference, as a form of epistemic agency. Attention provides the kind of guidance, under the control of the agent, that is fundamental for inferences to be epistemically justified. A substantial advantage of this approach is that it can be verified empirically through the voluminous scientific studies on attention. Thus, a moderate view of inference postulates that unconscious processing may satisfy normative requirements for inference, as long as *agential cognitive control* is involved. Justified inference, on this account, can be implicit or automatic. This moderate perspective on inference incorporates the main normative requirements that the restrictive view demands with respect to agency, but it avoids the problems that emerge from the restrictive view's requirement that cognitive access must be phenomenally conscious, thereby expanding the scope of who counts as intelligent.

An explanation of the psychology of epistemically responsible inference is a fundamental component of a satisfactory theory of rationality and high-level

cognition. As Boghossian says: "in epistemology we are obsessed with the idea that there are better and worse ways for you to manage your beliefs; and that these ways reflect on your virtues as an epistemic agent" (Boghossian, 2018, 60–61). An account of rationality depends on a clear demarcation of the boundaries of epistemically responsible inference—the type of psychological process that demarcates the realm of inferential reasoning and epistemic justification from other kinds of cognition that lack epistemic justificatory status. Rational and representational needs can only be satisfied if there is a robust relation between low-level "input" or sensorial information and high-level attention routines that make certain contents salient through the right hypotheses and inferences. From recognizing a bird as a crow, to identifying a move in the game of chess, this relation cannot be strictly sensorial and must be guaranteed to happen in a non-risky or non-accidental way. Inferences are perfectly suited to satisfy this fundamental representational need. In fact, the Kantian tradition defines concepts in terms of inferential rules. Inferential rules are also fundamental for epistemic cooperation and trust. Inferential reasoning is indeed fundamental for the satisfaction of critical representational and rational needs.

From an experimental point of view, an approach to inferential reasoning based on attentional integration at multiple levels of processing can explain, and be confirmed by, findings on common neural mechanisms for top-down control. As Song et al. (2017, 1) say, "numerous studies have recently suggested a shared neural circuitry underlying cognitive-emotional conflict resolution" (see Cromheeke and Mueller, 2014; Pessoa, 2008). This kind of conflict resolution, similar to the cognitive conflict resolution in the Stroop task, is attention dependent. In addition, attentional integrity has been confirmed as a crucial basis for uniform modulation and motor control (Lupyan, 2017; Rinne et al., 2018). Thus, an additional implication of the present proposal is that it could help clarify how a unified neural mechanism for cognitive control can be understood theoretically in terms of the epistemically proper *integration* of attention and inference. This unified approach to inferential attention can then be modeled in AI.

3.3 Defining Inference

This section explains two desiderata for an adequate account of inference. These desiderata avoid the kind of anthropocentrism that would entail that AI's cannot count as intelligent because they cannot draw epistemically justified inferences.

What kind of mental process should count as an inference and why is attention particularly relevant for the psychology of inference? To begin with, one must specify conditions that the most basic kind of inferential reasoning must meet. Then one can explore how other conditions must be met for such reasoning to qualify as clearly epistemically justificatory, rather than merely "inference-like." Most authors agree that an inference is a *psychological process* that provides an epistemically important outcome because of its cognitive structure (e.g., Malmgren, 2018; Siegel, 2017). There is also consensus about how this epistemic outcome must depend on a *cognitively controlled psychological action* that arrives at a conclusion in response to the content of the premises of the inference, which serve as reasons for drawing the conclusion. This kind of control by an agent is risk-reducing and trust-producing, as explained previously. This has direct implications for the autonomy of genuinely intelligent agents: as long as an agent satisfies this requirement of epistemic control, the agent qualifies as intelligent independently of other considerations, including their biology, which opens the possibility for AI epistemic agency.

However, there is considerable disagreement about whether an inference requires cognitive *access* to the justificatory relations among propositional contents, such that a belief that p is justified for a subject only if it is based (at least partly) on the content of another proposition q for which the subject has justification. A crucial clarification here is that for the notion of inference to be explanatory it needs to be a mental or psychological *action* under the control of the agent, rather than merely a *relation* among propositions (this is the difference between inferences and arguments; see Boghossian, 2018). On most accounts, an inference is a kind of mental process under some degree of cognitive control by an agent. Thus, necessarily, an inference is a psychological process that involves mental action under the control of an agent. There is controversy surrounding the type of mental activity that best suits inferential reasoning—can it be unconscious or is it necessarily phenomenally conscious? Independently from this issue which is explored at length below, an inference is at the very least a psychological process under the guidance of an agent in order to arrive at an epistemically important result: if the inference is adequate, then it provides an epistemic entitlement, typically justification.

A lot of difficulties emerge from spelling out the details of this rather general definition of inference. But it is a good place to start. It allows us to focus on two desiderata for the psychology of inference that directly bear on the debate between liberals and conservatives about inference. The first desideratum is that the definition of inference should allow for various psychological

processes to count as inferential. The second desideratum is that the definition of inference should allow for various styles of cognition. These desiderata are based on the kind of flexibility that children display in their remarkable learning performances, which integrate early sensorial signals with learning and meta-learning hypotheses.

Regarding the *various processes* desideratum, part of the justification for it is based on the fact that inferences are not merely relations among propositions. An inference is essentially a psychological process, and psychological processes can be quite diverse, so we need a criterion to isolate inferences from other processes that associate or relate propositional contents without being properly inferential. We do not want, however, this criterion to be too restrictive. On the one hand, we don't want subdoxastic, subpersonal, or strictly computational, associative, or merely representational processes to count as inferences. On the other hand, we don't want to restrict the notion of inference to a single type of psychological process associated with fully explicit and phenomenally conscious belief—this would be, as explained before, deeply anthropocentric and of no use to AI researchers. For instance, mnemonic, emotional, implicit, heuristic, and perceptual processes may very well fall under the category of "inference" and they should not be excluded from an account of the psychology of inference based solely on introspectively conscious grounds.

A plausible way of satisfying this desideratum is by defining inference as a psychological process under the control *of an agent*. This is still too liberal for conceptual and theoretical purposes because it includes mental states (such as jumping to conclusions) that should not count as inferences. But it is restrictive enough to prevent subdoxastic or strictly information-processing states from counting as inferential, such as the information processing performed in your retinas. It also allows for various inferential processes, as long as they are somehow under the control of the agent. Inferences are responsive to contents in a way that leads to epistemic entitlements when the inference is adequate, precisely because the agent is in control. In some cases, an inference may even include the same content, but it might still establish an epistemically crucial relation between two different processes, for instance, from perception to belief (Siegel, 2017). This approach to inference is compatible with very substantial types of cognitive penetration—high-level cognition may determine or guide low-level information processing, including early perceptual processing, through the mental actions of an agent. The question is whether this notion of

inference based on various psychological processes is too liberal. We shall soon return to this question.

With respect to the *styles of cognition* desideratum, the main idea is that human cognition should not be assumed to be the exclusive paradigm of inferential reasoning. This constraint is particularly important for AI research, but it also seems to be, at least initially, much weaker than the *various processes* constraint. For how else should we explain inferential reasoning if not by reference to human psychology and its cognitive architecture? However, as many authors have noted (Buckner, 2019; Kornblith, 2012), research in animal cognition strongly favors an approach to inference that satisfies this constraint. Inference occurs in a robust, epistemically entitling manner, across various species with similar perceptual capacities to human perception. This should not be too surprising given that we share our evolutionary path with them. So, at the very least, given the abundant evidence of inferential reasoning in animals, an account of inference should allow for the possibility that there is some kind of inferential reasoning with a genuine epistemic upshot in non-human animals. The problem with this evolutionary approach is that if we proceed strictly by parity, shared evolution and similar cognitive architecture is not necessarily going to work with AI. The "child machine" is unemotional but if what was said so far is true, she must be an agent and control her mental actions for her to have inferential reasoning. The only way to allow for this possibility, given that the child machine's architecture may be entirely different from human and animal cognition, is to expand even further the scope of inferential reasoning.

Therefore, other things being equal, these desiderata justify a liberal view of inference for the following reasons. Inferences are psychological processes that will, at the very least, resemble psychological processes in some species and eventually future AI. Ideally, and in accordance with the *various processes* desideratum, inferences should not be restricted to a single class of psychological processes associated exclusively with (human-like) phenomenally conscious belief, and our notion of inference should be compatible with the substantial experimental data emerging from psychology, demonstrating inferential capacities in non-human animals. That would make the case for expanding it further to AI—with the very important caveat that animals have biologically based representational and rational needs, and equivalent types of "motivations" to satisfy these needs would need to be developed and identified in AI. This is why the notion of a "child machine" is attractive—children have very strong

environmental pressures to satisfy reliably their representational and rational needs. Based on a non-anthropocentric and "non-solipsistic" criterion (Turing, 1950), the following are additional considerations in favor of these desiderata:

> *Automaticity* If we adopt a liberal notion, then inferential processes can include automatic and immediate forms of epistemic entitlement that make agential control more integrative, spontaneous, flexible, and reliable.
> *Complexity* If we adopt a liberal notion, then we can account for high degrees of complexity in reasoning with respect to various types of mental processes, from basic perception to abstract thought, which should be categorized as inferentially integrated with one another even if the agent is not consciously aware of all this integrative processing.
> *Integration* If we adopt a liberal notion, then we can explain various forms of epistemic influences, including cognitive penetration and hierarchy-dependent updates to values and preferences.

Before addressing standard objections to the liberal view, it is important to emphasize that these desiderata were presented in order to flesh out a necessary condition for inferential reasoning, namely, that inferences are *psychological processes*. According to the liberal view, to the extent that a child machine can simulate such processes, she should count as an inferential reasoner. But this is not the main motivation to favor a liberal view. Rather, the key motivation comes from the nature of psychological processes in human and animal cognition.

3.4 Accuracy Constraints and the Agency-First Account of Inference

This section addresses the view that having epistemic access to an inference is the central characteristic of inferential reasoning. It introduces a distinction between normative and descriptive accuracy, and argues in favor of agential control as the central feature of inferential reasoning, or an "agency-first" account.

The previous section shows that if one considers inferences as psychological processes, then one should favor a liberal approach given what we know from the empirical findings. The situation is quite different, however, with respect to the requirement that the outcome of such a psychological process, when adequate, must provide unambiguous and robust *epistemic justification*, which is also a necessary condition for good inferential reasoning. Here we find a

fundamentally different type of disagreement, which, according to many authors, clearly favors the conservative or restrictive view.

The central point of contention here is whether the psychology of inference necessitates a specific kind of *phenomenology*, or a specific kind of cognitive access that depends on subjective (even reflective) awareness. If so, this requirement challenges the relevance of the psychological desiderata because what is most distinctive about an inference is its unique normative status in epistemology, which depends on the *experience* of drawing an inference. Conscious awareness, according to the conservative view, is fundamental to the type of access that is required for drawing an inference. Thus, if it turns out that only human psychology can provide the *right kind* of conscious access for inferential reasoning, then we should ignore the empirical desiderata. This is, after all, a normative question concerning justification, rather than a merely descriptive one. If epistemic requirements are not favored over psychological ones, then we lose track of the core concept of inference. To guarantee the epistemic standing of an inference, according to the conservative view, an agent must have conscious or reflective access to the process of drawing the conclusion of an inference based on its premises. This is the only kind of agential control that is relevant for epistemic justification and responsibility.

Given this line of argumentation, the restrictive view confronts long-standing difficulties concerning the nature of conscious access. Should "internalism" about justification be expressed simply as a supervenience condition requiring epistemic justification to depend exclusively on the internal properties of mental states regardless of their phenomenology, or should it also include mental types of a *specific kind*, namely, states in which the subject has a unique and well-supported access-relation to reasons or evidence in virtue of a unique "what it is like" to draw an inference? Should this state be one that is also reflective and always consciously available to the subject; should it also be accompanied by other specific kinds of phenomenology, for instance, the subjective experience of understanding?

I will not rehearse here the various answers and objections concerning these questions. My goal is to classify these objections and responses in a way that helps move the debate forward, and in order to clarify how it may apply to AI. All forms of accessibilism about epistemic justification, strong or weak, justify their internalist requirements by appealing to *epistemic norms*. In weaker versions, access to evidence is compatible with implicit, not necessarily conscious belief. In stronger versions, access needs to be grounded on the phenomenology of conscious reflection and the assessment of the inference. Thus in all versions,

it is access to evidence, to reasons, to norms, or to all three that makes an agent *responsible* for her inferences. The agent deserves credit when the inference is drawn correctly and her inferences can be rationally evaluated because she is responsible for the inference. Although it is not trivial to define "access," the fundamental assumption of all evidentialist accounts is that cognitive access needs to be *personal level* access. This is all, obviously, problematic for the case of AI, but the emphasis on representational and rational needs clarifies the notion of "personal level" access: personal-level reasoning is done in order to satisfy the representational and rational needs of an agent, in virtue of her skills.

Let us set aside for the moment the issue of whether or not access should be necessarily conscious, which is the topic of the next section, and settle now for the more common assumption that access must occur at the personal level. By focusing on the less demanding versions of accessibilism, we might find a more balanced perspective on the psychological and normative requirements of inferential reasoning. As is well known in the literature on rationality (e.g., Gigerenzer, 2008; Kahneman, 2011; Morton, 2012), the ideal account of inference would need to satisfy these two requirements:

> *Descriptive adequacy* The access-conditions on inference must be actually true of human psychology.
> *Normative adequacy* The access-conditions on inference must satisfy requirements for rationality, evidential support, and epistemic justification.

Descriptive adequacy is explicitly anthropocentric, but it is a crucial constraint: if a requirement for rationality cannot be satisfied by humans, then it should be abandoned. Humans certainly satisfy rational needs through inference, so if a formal requirement is not really applicable to human rationality then it could be the result of a mismatch between the formal or ideal model and real human rationality. If we have a model that is not true of the clearest instance of rationality and intelligence we have available, namely, human intelligence, we lose the key foothold to our understanding of rationality.

One way of satisfying *normative adequacy* is to postulate additional necessary conditions on inference, such that the evidential and justificatory support for a belief be antecedently grasped or assessed by the subject. This epistemic access to well-founded propositions or beliefs explains why an agent is justified in drawing the inference. Mere propositional support does not entail full doxastic or endorsed-belief justification, but it provides rational *grounds* for the inference. This approach denies the status of "inference" to reasoning that is inferentially structured, but *unavailable* for scrutiny and person-level evaluation. Yet, it grants

that inferentially integrated but inaccessible states *can* play a kind of justificatory role, only one that is *non-rational* (Malmgrem, 2018). An alternative approach to *normative adequacy* is to grant epistemic standing to such "non-evaluable" or non-accessible inferences, which brings us back to the issue of scope. This second option allows for inferential reasoning that is not consciously available at the personal level to count as epistemically relevant, in positive and negative ways (Siegel, 2017). Both proposals aim at striking a balance between descriptive and normative adequacy. But both have the limitation that they end up favoring a version of conservatism and liberalism. The impasse remains.

It seems that as long as the emphasis is on conditions for inference *alone*, as a process that needs to be understood in terms of either formal requirements or actual psychological process, the impasse will remain. The psychological conditions must be fleshed out and developed in a way that the normative conditions are also clearly satisfied. The descriptive and normative adequacy conditions seem to pull in opposite directions, but only if essential considerations about epistemic *agency* are ignored. On all accounts of inference mental agency is fundamental. The problem is that agency is never considered as the *defining* feature of inference, and the focus is instead on access to information and the phenomenology of conscious states. But most authors would agree that an inference is a psychological process that an agent is somehow *in control of*. Conscious control at a personal level is crucial for the conservative view. Some kind of personal level or agential control, even if it is unconscious, is also required for the liberal view because an inference is drawn by an epistemic agent, and not by a module or subdoxastic computational component of the agents' architecture. The present proposal is to explicate the psychological requirements of epistemic agency in inferential reasoning in order to address the descriptive and normative adequacy conditions at once. Call this the *agency-first approach* to inference. Unlike the "process" or "conditions for access" approaches, the agency-first approach restricts the relevant type of psychological processes that should count as inference only to those processes that count as genuine exercises of agency.

An agency-first approach explains how inference is a psychological process under the control of agents who must satisfy epistemic needs, and at the same time it satisfies the normative condition that an inference should be responsive to reasons or evidence available for personal-level assessment and evaluation, at least in principle (see the following sections for details). The key, therefore, is to identify a psychological process that qualifies as epistemic agential control. The present proposal is to identify this kind of epistemic agency with *attention*.

A condition based on agency is as follows: A psychological process is an inference only if it is under the cognitive control of an epistemic agent. If one adopts the *inferential-attention* view, this condition states that *a psychological process is an inference only if it is under the cognitive control of an epistemic agent, understood in terms of her attentional capacities.* An agent satisfies her rational and representational needs through inferential attention routines. This satisfies the descriptive adequacy condition by identifying a well-known psychological process that explains the kind of epistemic control needed for inferential reasoning. Once agency and cognitive control are considered as central, a performance normativity account of epistemic justification (or virtue epistemic approach) satisfies the normative condition (Fairweather and Montemayor, 2017).[3]

Rarely have virtue approaches been considered as relevant in the literature on the nature of inference. It has been generally assumed by defenders of the conservative view that only introspective conscious control can satisfy normative requirements. But this is, at best, an incomplete picture of how inferential abilities satisfy the normative constraint because one still needs an account of how introspective abilities ground and integrate *inferential* abilities and processes. In addition, the emphasis of conservative views is on the *phenomenology* of introspective access, rather than on the *abilities* involved in inference. But introspection, if it is indeed an epistemic skill, is quite different and independent from inference (many authors are indeed skeptical about the nature and trustworthiness of introspection; see, for instance, Dretske, 2012; Kornblith, 2012; Reginster, 2004; for an opposing account, see Moran, 2001).

Phenomenal consciousness, or the subjective qualitative character that accompanies conscious experience, by itself, falls short of satisfying the normative condition because one needs to show how merely experiencing a conscious content guarantees the cognitive ability to infer and satisfy the rational and representational needs of an agent. It is dubious that merely being in an experiential state with a particular phenomenal character will suffice to explain an ability, because abilities must be defined in terms of *success conditions*. For instance, even if you introspectively experience the strong conviction that you should be able to hit a homerun, the contents of your experience do not necessarily entail any connection between your experience and the facts that must obtain for you to have the ability to hit homeruns. Your conscious experience that you can hit homeruns is relevant for baseball playing only if it is associated with the ability to hit homeruns.

Inferences are not just experiences or thoughts with qualitative character. Rather, they are abilities to draw conclusions from premises under the control of an agent: a kind of mental action that satisfies epistemic needs. Thus, reflective conscious thought and introspection, by themselves, are *insufficient* to explain the kind of ability needed to infer. Although there are considerable grounds for skepticism about conscious introspection as a condition for inference, the emphasis here is on the need to explain inference as an ability in the first place. The point is not that conscious awareness is *irrelevant* in inferential reasoning, but rather that it is insufficient to explain inferential abilities. Attention is the cognitive capacity that is necessary for inferential abilities, independently of phenomenal consciousness—many inferences can occur and must occur without phenomenology.

In spite of the fact that an agency-first approach to rational inference has not been the main focus of attention in the literature on inference, the central role of abilities features in recent discussions about epistemic justification, particularly concerning how to confront skepticism (Lasonen-Aarnio, in press; Williamson, in press). For instance, Lasonen-Aarnio (2010) draws a useful distinction between rationality and reasonableness in terms of a success condition: rationality *necessitates* success (basing belief and inference on evidence) while reasonableness only requires a competence or disposition to do so. Neither requires a specific phenomenology. Our inferential abilities must succeed in typical conditions for them to satisfy the normative constraint and count as rational. In particular, they must satisfy rational and representational needs in a non-lucky or non-risky way, but they need not involve conscious introspection, reflection, or awareness—although conscious introspection *can* play an important role, as explained below. This means that the success required for rational inference depends fundamentally on the agent's cognitive control through her abilities to succeed, and not necessarily on whether or not the agent is in a particular kind of conscious state. This confirms that Turing (1950) and Russell (2019) are right that phenomenal consciousness is irrelevant for AI—as long as AI involves attentional epistemic agency, an AI agent will qualify as intelligent and rational.

Epistemic agency, as personal-level cognitive control, is necessary for epistemic trust, credit, and responsibility, and this is a core assumption of performance-normativity approaches in epistemology (Fairweather and Montemayor, 2017; Greco, 2010; Greco and Turri, 2011; Miracchi, 2015; Sosa, 2007, 2015). Miracchi (2019) explicitly applies a competence, virtue theoretic approach to AI (although

she does not specifically focus on attention skills or inference). But independently of the advantages of an ability or agency approach, all the literature on inference shows that cognitive control is necessary. An ability- or agency-based approach, therefore, offers the best way to explain this type of cognitive control. An ability-based approach to inference can meet normative standards for knowledge and rationality without entailing problematic assumptions about conscious reflection or access (Fairweather and Montemayor, 2017).

One can go further and argue that an ability-based approach is superior to current approaches to inference because the success condition on rational belief is built into an attentional agency account. While this might be the case, all that is needed for present purposes is to show that a virtue account of inferential abilities suffices to explain their epistemic status. This has the significant advantage that one can give an account of inference that is compatible with the abundant evidence on unconscious reasoning and cognition. According to this account, unconscious or tacit cognitive processing can not only guide high-level cognition, but also contribute to its epistemic status. However, the descriptive adequacy constraint still needs to be addressed. Explaining how an ability-based, virtue-theoretic approach can satisfy this constraint while also complying with rational norms is the purpose of the remainder of this chapter.

3.5 Attention: High- and Low-Level Inferential Cognition in Various Domains

This section defends an agential account of inferential-attention, which solves the tension between conservative and liberal approaches. Findings supporting the descriptive adequacy of this account are provided. It shows why inferential attention can meet epistemic standards without necessarily being phenomenally conscious.

According to the condition stated above, an inference is, necessarily, a cognitive process under the control of an agent. If the inference is adequate, then it provides an epistemic entitlement, typically justification. This normative requirement demands some kind of agential control. Otherwise, defining inference strictly in terms of the reliability of a cognitive process falls short of the standard required for normative standing and epistemic responsibility. The challenge, then, is how to define inference as a psychological process without falling prey to traditional objections concerning the lack of normative standing. As mentioned, the present

proposal is to define the psychological process involved in inference in terms of the agential control provided by attention.

Most authors define attention as a psychological process of selection (see Jennings, 2020, for a historical account of how attention has been defined in philosophy and psychology). But attention is not merely a process of selection either (Mole, 2011; Watzl, 2017). As some authors have argued, it is a form of *mental action*, or selection *for action*, or *for a subject*, guided by needs, motivations, and intentions, even though many of these needs, motivations, and intentions are typically implicit (Jennings, 2020; Wu, 2011). For present purposes, there is no need to endorse the view that attention is, necessarily, a psychological process of selection *by a subject* (a metaphysically robust "self"). What is needed is that attention be a process of selection that always occurs at the personal level, which includes motivations and goals (Wu, 2011).[4] The agent, on this account, is responsible for her inferential-attention routines. Thus, attention provides a type of guidance and control that is particularly relevant for epistemology (Fairweather and Montemayor, 2017).

Empirical findings demonstrate the descriptive adequacy of the agential account of inference. Crucially, attentional selection at the personal level is essential for explaining *responsibility* in multiple psychological studies (e.g., moral, epistemic, and even legal responsibility; see Jennings, 2020). This is especially relevant when top-down attention routines modulate early processing, which can cause negative epistemic, as well as moral, repercussions. For instance, recent studies demonstrate an alarming combination of unreliable epistemic guidance and morally reprehensible bias. Attention guided by racial bias produces unjustified inferences about the identity of objects (e.g., a gun versus hand tools) with unfortunate moral implications (James et al., 2013; Payne, 2001; see also Benjamin, 2019, for a more comprehensive approach to algorithmic social injustice related to this kind of phenomenon). Research has shown that similar effects drive attention in a "shooter task" (Correll et al., 2002), as well as judgments of criminality concerning objects (Eberhardt et al., 2004), age (Goff et al., 2014), and judgments about capital punishment (Eberhardt et al., 2006). These are bad inferences in at least two ways because they are epistemically and morally inadequate. This inadequate kind of attentional guidance is under the implicit or unconscious control of the agent, preventing the agent from satisfying her representational, rational, and moral needs. The agent is responsible for these bad inferences.

Other effects of implicit inferences based on top-down attentional biases are less troubling from a moral perspective, but they could be problematic from

an epistemic perspective. Color perception seems to be susceptible to cognitive penetration (Macpherson, 2012). If the color I am seeing is determined by inferential influences that are pervasive and unreliable (for instance, I am more likely to see red when something looks like an apple independently of what is the apple's real color), then cognitive penetration would hinder the satisfaction of representational needs concerning color. This problem could extend to many aspects of perception and cognition because presumably the same could be true of many other perceptual features (see Montemayor and Haladjian, 2017, for a critical discussion about the scope of cognitive penetration in the context of the functions of attention). But cognitive penetration need not be problematic in this way, and in fact, in many cases it constitutes a kind of virtuous integration without which representational needs associated with general intelligence cannot be satisfied. Typically, the selective functions of attention routines are *virtuously sensitive* to reliable information—they tend to be epistemically adequate, because they ignore irrelevant information and are immune to frequent error by preventing an overwhelming influence of unjustified biases.

Attentional guidance is also fundamental for extremely skilled types of high-level inferential attention, but in many cases such guidance is best understood as implicit or automatic inference, rather than cognitively penetrated perception. Attentional integrity and high-level executive function are fundamental for low-level motor dexterity and strength (Rinne et al., 2018). This kind of attentional integrity can be understood as *agential integrity*, which unifies the motor-control level with the higher executive-function level. Integrated guidance allows us to structure needs into a hierarchy: by satisfying the most important representational needs, other needs can also be met, which keeps the agent in good cognitive shape. Inferential-attention integrates various attentional subroutines that aid in fulfilling multiple tasks, which is critical for an agent's performance as an epistemic agent. For instance, Siegel's (2017) theory of perceptual inference distinguishes the positive or negative "epistemic charge" of sets of implicit inferential precursors to perceptual experience. Her main focus is the cognitive penetrability and rationality of perception. But as Irving (2019) argues, these effects are best understood in terms of *norms of attention*, or what I am describing as the rationality of inferential-attention. Thus, Siegel's account is compatible with the present proposal if understood, as Irving proposes, as an account of the rationality of attention.

In addition to its role in perception, top-down attentional modulation determines non-phenomenally conscious memory trace formation and it also suppresses sensorial input to allow for high-level phenomenally conscious

memory content (Jacob et al., 2015). What unifies and makes these agential capacities epistemically adequate is not their phenomenology, but rather, the fact that these attentive capacities selectively and reliably satisfy the representational and rational needs of an agent, even in the absence of any phenomenology, as memory consolidation illustrates. The integration of these attentional capacities constitutes a form of epistemic excellence that can be assessed in terms of good cognitive performance, based on the good making features of an attentive agent. Any attention routine starts with an input that triggers a guided process of selection in order to obtain an answer. Success in epistemic tasks can thereby be attributed to the agent because of her capacities to integrate and attend to relevant features.

The empirical findings show that attentional integration occurs at various levels of cognition, including motor-control dexterity. If inference is not conceived merely as a psychological process with epistemic consequences, but also as a specific kind of attention routine with a degree of assertive force associated with action and motor-control, then a moderate view on inference is possible, and for the reasons offered above, preferable. According to this account, the liberal is right in extending the scope of inference to its lowest bounds (see Buckner, 2019; Siegel, 2017) and the conservative is right in demanding that inferential reasoning be solely attributable to agents that have cognitive control over the inferences they draw (Boghossian, 2018). Attention is a cognitive ability that explains why inference is not merely a psychological process because it essentially involves a type of mental agency with a clear epistemic upshot: succeeding in satisfying the representational and rational needs of an agent in an optimal and reliable manner under the guidance and control of the agent.

The definition of inference above states that a psychological process is an inference only if it provides an epistemic entitlement, typically justification. Why typically? This is because inferences may generate epistemic entitlements that differ from full epistemic and rational justification. Good inferences that are unconscious, according to some authors, are not *fully* rational (Malmgren, 2018). But unconscious attention and inference can still satisfy representational needs and meet a justificatory standard, even if they fall short of the highest standards of rationality. In fact, there are fundamental kinds of rationality that count as normative without being explicit or phenomenally conscious (Gigerenzer, 2008). If they typically or reliably provide epistemic entitlements, such as knowledge or justified belief, then they meet the epistemic norm even if they are not fully "evaluable" in human awareness, thus providing grounds for epistemic comparisons with other intelligences, animal and artificial.

Inferential attention also operates in the moral, aesthetic, and practical domains. This is the topic of Chapters 4 and 5, which discuss emotional, aesthetic, and what Maslow calls "transcendence" needs. These needs differ from epistemic needs because phenomenal consciousness plays a more critical role in their satisfaction. A virtuous agent will integrate all these needs harmoniously, but often, the satisfaction of only one type of need at the cost of others will create conflict. However, there is enough similarity in the structure and functions of attention *across domains* to justify the hypothesis that inferential-attention will be capable of explaining rational inference in all domains—attentional capacities are required to satisfy all these needs. Overall agential virtue and success will ultimately depend on the fine-tuned satisfaction of multiple needs in a well-integrated hierarchy of needs. The similarity in structure of inferential attention routines can serve as the basis for a much more ambitious account of rationality and intelligence with crucial implications for the development of AGI, encompassing the whole range of cognition based on conscious and unconscious attention.

A central point of this discussion on inference is that high-level cognition for rationality and epistemically normative outcomes *need not be* phenomenally conscious (there need not be any specific "what it is like" for these cognitive process to satisfactorily deliver good outcomes). This is fundamental to understand the possible satisfaction of representational and rational needs by AI. High-level cognition in epistemology is compatible with an implicit, and phenomenally unconscious, kind of attentional guidance that allows, nonetheless, for enough access and control over inferential processes. Although this view of inferential-attention does not necessitate phenomenal consciousness, it does not *exclude* it. Actually, this account of inferential attention and epistemic agency is very well suited to explain the differences between two "styles of reasoning": optimal but unconscious reasoning and necessarily conscious reasoning which is reliable only if phenomenally conscious. Kahneman (2011), for example, emphasizes the unreliability of fast and frugal heuristic reasoning, while Gigerenzer (2008) defends such reasoning as optimal. Both authors agree, however, that there is implicit or unconscious inferential reasoning and that it plays critical cognitive functions, even if they disagree about the nature and scope of phenomenally conscious rationality. Thus, although there is disagreement among psychologists concerning what exactly should count as rational cognitive processing, there is consensus that some kind of virtuous integration through something akin to inferential attention is fundamental.

Non-human species also rely on inferential attention, and since consciousness is not necessary for the proper functioning of epistemic attention routines,

they could in principle be instantiated in AGI. There are few uniquely human capacities, and the extent to which rationality is uniquely human is challenged by the empirical evidence (de Waal, 2016). Inferential-attention capacities in the moral, aesthetic, and epistemic realms need to be studied more carefully in a broader context. But the structural similarity among different kinds of attention is useful to draw a general distinction between good and bad inferences across different normative domains. Bad inference is, in general, unreliable inference, but depending on the normative domain, the unreliability of an inference may mean substantially different things. In the epistemic domain, a good inference is truth-conducive (for instance, in the case of deductive inference, if the inference is drawn from a valid argument, the inference is truth preserving—if the premises are true, the conclusion must be true). Deductive inference differs from the automatic and inductive type of inference required for perception and motor-control, but both kinds of inference involve truth conduciveness and require cognitive control.

The most paradigmatic examples of high-level cognition involve a kind of attention-based *dexterity*. One can characterize "intellectually *responsible* intuitions" as a kind of high-level attentional dexterity, analogous to perceptual skill. Consider Descartes' discovery that one can prove truths about algebra through geometry and truths about geometry through algebra. There is no immediate relation, based on conscious reflection alone, that could justify investigating such proofs because the conscious access we have to visual and perceptual figures in geometry seems entirely independent from the relation among abstract and "invisible" entities, such as numbers. However, Descartes could "sense" that there had to be such a relation, based on his tacit knowledge of mathematics. Conscious awareness is thus guided by more tacit and unconscious subroutines that help conduct creative inquiry in a selective and virtuous manner, in order to arrive at a conclusion. Learning how to learn mathematics and other disciplines, and learning how to identify relevant and interesting problems in general, operates very much in this implicit fashion.

Groundbreaking mathematical discoveries are an exceptional case that illustrates how inferential-attention guides multiple subroutines simultaneously. When one understands the Euclidean axiom of the parallels by "looking" at the parallel lines and imagining that they go all the way "up and down to infinity," these perceptual routines are implicitly guided by a *rule* that defines the space of perceptual visualization (an infinite space) as having zero curvature. One need not be consciously aware of this specific rule concerning zero curvature in order to follow it implicitly. Attentional guidance through perception and imagery thus

allows us to understand the axiom of the parallels without explicitly formulating the axioms of an infinite two-dimensional space with zero curvature. This is not really a typical case of *cognitive penetration* (e.g., an implicit rule, emotion or bias affecting how things *appear* to us in perception) because space never appears as infinite to us. But we clearly use this kind of top-down rational influence through inferential-attention to learn abstract knowledge perceptually, from basic mathematical and logical proofs to scientific theories. Only through this kind of virtuous integration of various attention routines into a novel type of learning dexterity can abstract representational needs be satisfied.

In the moral case, one can conceive of a view in which inferential-attention is guiding our emotional and perceptual contents toward the needs of others, in a way that we become virtuously sensitive (morally virtuous in our selectivity) to moral needs. This is the kind of care-based attention routine that is constitutive of empathic attention. Normative guidance in some of these domains necessitates explicit conscious inference or phenomenally conscious attention (Montemayor and Haladjian, 2015). Some animals very likely have phenomenal consciousness, but lack the capacity for linguistically guided inferences about explicit rational norms. (In the case of AI agents, their lack of emotional needs entails a lack of empathic attention and rationality—more on this in the next chapter.) Nonetheless, attention plays the same fundamentally selective and integrative role required to satisfy moral needs in the non-linguistic mental lives of animals.

3.6 Not Necessarily Phenomenal Rule Following for Inferential Rationality

This section compares the inferential-attention approach with Paul Boghossian's influential account of inference based on conscious intuition, and it argues that the former account is preferable, demonstrating the significance of inferential-attention in higher cognition, even when it is not phenomenally conscious.

As mentioned, the literature on inference emphasizes either the conscious and explicit rational endorsement of an inference or the flexible and automatic character of inferential reasoning, both of which are important for the proper satisfaction of representational and rational needs. The first group of these views is associated with the highest forms of cognition, while the second is associated with early cognitive processing. Attentional agency provides a new perspective on these issues, which satisfies both conditions of adequacy: normative and

descriptive. A further advantage of this inferential-attention account concerns the notorious *problem of regress* about rule-following. Inference seems to require the *acceptance* of rules. Typically, this type of acceptance is understood in terms of some form of decision or mental activity that *itself* is constituted by following the rule to accept the first-order rule. This triggers the regress. The acceptance of an inference depends on the intention or decision to draw a conclusion based on the premises of another inference—a *rational rule* that must be correctly applied. But if an inference is already a rule of reasoning and there are rules concerning our decision to apply it to a concrete case, then our acceptance of these rules generates intermediate "application" rules. Our acceptance of these rules, and the further rules that justify *their* application to particular cases, generates a regress because there seems to be no end to the process of determining which is the foundational rule that justifies all others.

Several difficulties emerge from this problem. I shall focus on problems associated with the type of mental act involved in the acceptance of a rule. Paul Boghossian (2014, 2016, 2018) has defended one of the most detailed and comprehensive views about inference. He proposes that inference is a kind of *mental action*:

> Inference, as I have characterized it, is mental behavior and, so, for it to make sense to hold you responsible for your inferences, inferring has to be something you *do*, and not just something that happens to you. It has to be a mental *action* of yours, something *you have control over*, and which you could have done differently, had you thought it desirable to do so.
>
> (Boghossian, 2018, 60, my emphasis)

Unlike arguments, inferences are essentially mental actions, because an inference is not merely a set of propositions, but fundamentally, a *movement* of thought from premises to conclusion (Boghossian, 2018, 55). What is, exactly, this mental action? It must be an *intentional* mental action, precisely because you are responsible for it. But as Boghossian argues, inferential mental actions cannot be based on conscious or explicit *intentions* to follow rules because tacit or implicit inference plays a central role in our epistemic lives (Boghossian, 2018, 66–7). According to Boghossian, there are three defining features of the mental action that properly falls under the category of "inference." *Basing* determines that agents accept the premises of an argument as the reason for believing the conclusion that follows from them—agents accept that the premises serve as the basis for the conclusion. *Quality*, given *Basing*, establishes that the conclusion drawn from the premises "can be assessed as resting on good or bad reasons."

Responsibility is based on these two properties, and attributes them to the intentional mental actions of an agent, for which she is accountable—the assessment based on *Basing* and *Quality* determines an assessment of *the agent's rationality* (Boghossian, 2018, 59).

On Boghossian's account, these three features apply to all and only those psychological processes that qualify as inference: fully explicit reasoning; inference without knowledge of the principle that allows for the transition from premises to conclusion; quick, effortless inference; and inference in children. Boghossian claims that so-called "inferences" that are subdoxastic or not at the personal-level, inferential-like reasoning in *all non-human animals*, and information processing in *artificial intelligence or computers* do not satisfy these three features, and therefore, should not fall under the epistemically fundamental category of *inference*. It follows that, for Boghossian, neither animals nor AI can be rational epistemic agents.

Inference in children, largely guided by implicit and "built-in" rules of rationality that the child never follows explicitly or according to clearly expressed deductive, inductive, or abductive rules, qualifies, according to Boghossian, as clearly rational. Animals (who certainly satisfy representational needs) and AI engage in information processing that is only "inference-like." As stated above, subdoxastic or subpersonal processing should not count as inferential—your retinas or some set of neurons in your visual cortex are not drawing the perceptual inference that you should believe that the color of that apple on the table is red. It is *you* who draws the inference. But animals have the same, or at least extremely similar, perceptual needs. Thus, an implausible consequence of Boghossian's view is that it denies animals (and machines) epistemic standing, thereby positioning them as a very voluminous sector of the non-rational world. By contrast, on the inferential-attention account, animals certainly deserve inclusion in the realm of rational epistemic agency because they satisfy autonomously their inferential and rational needs. Moreover, a child machine could qualify as an epistemic agent if she were truly equivalent in all respects to human and animal attention.

Therefore, while I endorse Boghossian's characterization of inferences as responsible mental actions, my disagreement concerns the narrow scope of his view. The key difference between the inferential-attention account and Boghossian's is that the former does not depend on phenomenally conscious states, and this makes the inferential-attention account more capacious and explanatory. But the two accounts are partially compatible because the inferential-attention account does not *exclude* the relevance of phenomenal consciousness.

In any case, an attention-based approach to inference is superior to an *intuition-based* one (such as the one favored by Boghossian) for the following reasons.

A striking difficulty regarding the nature of inference concerns the "distance" between premises and conclusion, as our thinking "moves" from the premises to the conclusion. Boghossian illustrates this problem as follows: although Fermat's last theorem *follows* from the Peano axioms, one cannot simply *infer* one from the other. This problem is related to the distinction between argument and inference, but the "distance" between premises and conclusion, say in a proof, is quite intricate and must be defined somehow. Boghossian writes:

> It looks as though what's also needed is that the conclusion not be at too far a distance from the premise. But what does that mean? The only good answer that I can think of is that the step from premise to conclusion be such that the thinker have some *appreciation* that the conclusion does indeed follow from the premises. Of course, unless this condition is to generate a super-task, it had better be that, for a wide range of basic inferences, this appreciation is non-inferential in character.
>
> (Boghossian, 2018, 60)

Boghossian's solution to this difficulty is that, since the thinker must *take* the premise to support the conclusion, this "taking" must be "backed by an intuition to the effect that the taking is true" (Boghossian, 2018, 60). This intuition-based approach is used by Boghossian to solve a lot more than the distance problem. In fact, the notorious regress problem is also tackled with intuition by appealing to the kind of understanding and appreciation provided by its phenomenology. Boghossian distinguishes two types of regress, which he calls "ingress regress" and "egress regress." Ingress concerns the way in which we rationally *get into* the taking state. If we get to this state via an inference, which seems necessary since the state has a general content that we must grasp through some rule, then it seems impossible to get into this state while avoiding regress. Egress involves the *transition* from the taking state to the conclusion—if it is through inference then it seems impossible to do so without regress.

Both of these problems, Boghossian claims, can be solved by appealing to conscious intuition. With respect to *ingress*, a thinker takes her premises to support her conclusion because she has "the vivid intellectual impression" that whenever the premises are true, the conclusion must also be true (Boghossian, 2018, 62). The nature and importance of intuitions are briefly described as follows: "Taking states can seem like beliefs; but it's important that, although they are belief-like, they are distinct from beliefs […] Underived taking states,

that is, taking states not derived from other taking states, can only be entered into via intuitions (and not by testimony or inference)" (Boghossian, 2018, 62). But practically speaking, why should a vivid intellectual impression guarantee the satisfaction of a representational or rational need? And more important, what kind of mental *action* is an intuition, if it is described passively, as an *impression*? With respect to *egress*, Boghossian says that "we know of many examples of intentional states with general, conditional contents rationally controlling behavior without the benefit of inference" (Boghossian, 2018, 63). He provides the example of a tennis player who implicitly controls her behavior without drawing inferences. Consequently, transitions from the taking state to the conclusion can be in control of the agent without necessitating an explicit inference. This is exactly the kind of *dexterity* that characterizes automatic inferential-attention, explained above. But isn't this a clear example of an action, rather than an intellectual impression?

What about inferences in which, unlike inferences in mathematics or critical thinking, the thinker lacks both an explicit aim and an explicit "taking" state (including inferences by children)? Here Boghossian proposes that the three basic features of *Basing, Quality*, and *Responsibility* need to be understood in terms of goal-directed actions under the *rational control* of the thinker (Boghossian, 2018, 63). Something akin to conscious taking is needed to guarantee rational control. Boghossian proposes taking states that are present *tacitly* (or implicitly). By relying on this tacit rational control, agents are relating contents under a non-phenomenally conscious guidance for thought transitions of the form, "so" or "therefore." Thus, quick and automatic inference can be under the intentional and rational guidance of the agent, even in the absence of a conscious "taking" state. But then, why shouldn't animals, or even Turing's child machine, count as drawing inferences?

Since the intellectual vividness of intuitions is what grounds the taking state in explicit inference, what is implicit, according to Boghossian, must be precisely *this kind* of vivid intuitive support—the intuitive guidance is there, it just *"becomes" automatic and habitual*. Is this claim empirically plausible? First, it is not clear that intellectual impressions really constitute mental actions, so an attention-based account seems preferable. Attention is certainly easier to understand empirically, as a mental action under the control of an agent, rather than this more empirically controversial phenomenology of intuition, or vivid intellectual "seemings." Second, consider how attention would solve the problem of *appreciation*. Personal appreciation for how premises support conclusions explains the distance between them. An intuitive-based account explains this

in terms of the phenomenology of the experience of appreciation or "taking." Tacit guidance lacks this conscious understanding, but according to Boghossian, it depends on it, as it has to be originally based on a conscious intuition that then becomes habitual. The question is whether *all* tacit inferences really depend upon the phenomenology of conscious intuition. From an empirical perspective, the answer is: certainly not.

Many inferences that we rely on to rationally guide our mental actions are *never based on intuitions*, and are tacit from the very beginning. Inferences underlying our knowledge of linguistic syntax are pervasive in our mental life. It takes linguists years of training to explicitly appreciate the inferences that determine the grammaticalness or lack thereof of sentences. Typically, one only tacitly follows the principles guiding these inferences, without any conscious intuition or taking. So how is appreciation supposed to work for the young infant and the standard language speaker? Syntax appreciation involves high-cognition inferences that cannot be explained in terms of the vividness of an intellectual seeming or intuition because they are *essentially implicit inferences*. Other examples of higher-order cognition that rely on this type of inference include practical or inductive inferences and recognizing the speech acts of other speakers. The correctness of a sentence *may be* associated with a "feeling" or with a certain kind of phenomenology. But the rules of syntax and their inferential structure cannot depend on such feelings or seemings. Echoing Turing's recommendation, a more "unemotional" channel of communication is needed here.

The inferential-attention account avoids this problem because high-level cognition is perfectly compatible with non-phenomenally conscious forms of attention that provide guidance and an implicit form of *appreciation* based on attentional selection and salience. Mental actions can be rational without any phenomenology—this is in line with the fact that actions can be *evaluated* as good or bad independently of how it "feels like" to execute them (consider Boghossian's example of playing tennis). To evaluate rational mental action, all that is needed is the guidance and control that attention provides. This explains why essentially implicit inferences, like those concerning syntax, play a critical integrative role in our mental lives (see Richard, 2019; Siegel, 2017; Wright, 2014), and very likely, in the mental lives of non-human animals as well (Kornblith, 2012). A child machine would qualify as an epistemic agent that performs adequately without having *any* associated feelings of intellectual vividness or lack thereof. This could even be an epistemic *advantage* of the child machine, over the human "feelings" associated with rational thought, precisely because of the unemotional nature of tacit inferences.

A key reason to favor the inferential-attention account over the intuition-based one is that the psychology of attention is much better understood, and much less contested, than the psychology of intuitions or intellectual seemings. In fact, there is a whole branch of contemporary philosophy, namely, experimental philosophy, that systematically criticizes the use of intuitions and intellectually vivid imagery (and their phenomenology of certainty and truth) in philosophical analyses because intuitions can be shown to be unreliable in a wide variety of ways (see, for instance, Knobe and Nichols, 2008). As a matter of methodological prudence, the less controversial and well-verified psychology of attention should serve as the foundation for the study of inference.

What about the normative requirements that the intuition-based account clearly satisfies? Is the inferential-attention account capable of explaining the three key features of inference (i.e., *Basing, Quality,* and *Responsibility*)? The inferential-attention account can not only meet these three normative criteria, but also provide an explanation of inferential mental action that is *superior* to the intuition-based account. If the requirement for conscious taking is circumscribed to only *explicit inference*, then the attention-based account can provide the ideal way to satisfy Boghossian's normative constraints. In addition, the inferential-attention account can fully explain, and provide empirical support to, Boghossian's *mental action* approach. This is because an attention-based approach is compatible with phenomenally conscious forms of attention, and attention is more clearly active than vivid impressions. Inferential attention can provide an explanation of the appreciation of how the premises support the conclusion by appealing to the *selective and luck-eliminating* functions of perceptual and cognitive attention routines. Attention selects information through virtuously sensitive information processes, and it ignores (or is virtuously insensitive to) irrelevant information, in a reliable and non-lucky way, which explains the actual success of agents in achieving multiple epistemic goals (Fairweather and Montemayor, 2017).

Crucially, on the inferential-attention account what the agent "appreciates" is that her *representational and rational needs have been adequately satisfied*. This provides a better explanation of the key properties of *Basing, Quality,* and *Responsibility*. The agent needs to take the premises to be the basis of her conclusion, determining that it provides a good reason to draw the conclusion, which she is responsible for drawing. The epistemic "force" or justification of an inference must find its source not just on the phenomenology of intuition but, fundamentally, on a *selective and luck-eliminating capacity* that leads to rational success. It is the successful satisfaction of representational and rational

needs, based on inferential attention routines, that generates an appreciation for their proper guidance, from premises to conclusion. Attentional capacities are "luck-eliminating" because it is not by chance that the goal of moving our thoughts from premises to conclusions is achieved. This theory of rationality, based on inferential attention, certainly includes animals, and can in principle include AGI.

There are similarities between the inferential-attention account and Siegel's (2017) liberal inferential account of the rationality of perception. The present account endorses, and is fully compatible with, the agent-level guidance and responsibility that Siegel seeks to identify in implicit reasoning beyond conscious awareness. The main difference is that Siegel never addresses the nature of mental action. If her proposal is interpreted as essentially dependent on the guidance and norms of attention, the way Irving (2019) suggests, this problem is solved. To conclude, an inferential-attention account can explain how a child machine may satisfy her representational and rational needs and be considered as a fully competent epistemic agent. Like humans and animals, the child machine's "retinas" would not have epistemic relevance—only the child machine as an *agent* would be responsible of her conclusions and perceptual inferences. Because she has no biological or emotional needs, the child machine or any future AI will think through entirely unemotional channels, and this can be a very significant advantage because they won't have metabolically based distractions. However, the child machine might be fundamentally limited in other aspects of her rationality because of her unemotional nature.

4

The Handicaps of Unemotional Machines

4.1 Emotional Needs, Moral Intuitions, and Value

This section presents two preliminary difficulties concerning the application of any ethical theory by AI agents: a representational and a motivational limitation. The literature on ethical AI is briefly surveyed in the light of these problems. The work of G. E. Moore, Philippa Foot, and Wallach and Allen is discussed as part of an argument stating that, unlike epistemic justification, conscious intuition plays a central role in moral autonomy and justification.

Turing's unemotional child machine can be a fine epistemic agent whose thinking is based on inferential attention routines (if this is ever achieved), but there is a catch: the child machine lacks biological needs and the concomitant emotional and aesthetic needs that depend on the finite and vividly experienced lives of biological organisms. This is a major problem, as this and the next chapter elucidate. The difficulties unemotional machines confront stem from the fact that satisfying moral, aesthetic, and spiritual or lifelong needs is at the top of most human beings' priorities, even though they are rarely consistent in how they pursue them. Human beings would be irrational if they didn't pursue these higher goals and, therefore, their emotional intelligence is a core aspect of their general intelligence—humans should, to the best of their capabilities, live a good life. If AI cannot develop this kind of emotional intelligence, then they cannot be "human-compatible" or completely general in their intelligence and, therefore, there cannot be a comprehensive solution to the value-alignment problem. But if AI develops genuine attention routines, while there will still be risks concerning value alignment, there might be satisfactory solutions to epistemic value alignment, and that would be a very important source of risk reduction.

This argument can be defended on purely theoretical grounds, based on standards concerning value appreciation, required for value alignment

(see Chapter 2) and the lack of emotional needs in AI. The next chapter expounds this argument in the context of moral and aesthetic needs. This chapter centers on emotions that are relevant for morality, in the light of empirical evidence concerning the dissociation between consciousness and attention. The evidence suggests that epistemic agency, mostly associated with perceptual and inferential attention, is *dissociable* from phenomenal consciousness, which depends on the visceral reactions that inform our emotional needs.[1] This has important implications for the prospects of developing a morally "attentive" AI, which are explored in what follows.

Turing's child machine is an AGI attentive agent. But the scope of her intelligence is limited. Although she is capable of satisfying rational and representational needs, she has no foundation for aligning many of her values and preferences with the values of her creators, which are based on their emotional and biological needs. This limitation can be epistemically advantageous—having no biological needs, such as sleeping, avoiding death or stress from addictions, and so on, can make the child machine the most efficient and fast problem solver in the history of intelligent agency. By being out of touch with humans' biological and emotional needs, however, the child machine is intrinsically incapable of pursuing the highest forms of need-satisfaction, namely, those concerning the *categorical needs* that humans place at the very top of their hierarchy of needs. These very important needs, and many other biological and emotional needs, like reconciling after an argument by embracing, would be quite puzzling for an AI agent. AI could at best "interpret" these needs by pure mimicry. But since mimicry in social relations is unscrupulous, the imitation of emotions by the child machine would have no basis on genuine needs, and thus, it could constitute a very dangerous kind of ultra-intelligent *manipulation*, in which we believe our values are aligned, but in fact we are being exploited by very clever rational machines with no capacity for empathy. This would be an extremely dangerous kind of mimicry given, for instance, how easily addicted we get to online content and solicitation from social media platforms, as well as the ever increasing risks of deepfakes.

One model of "ethical AI" is to "build-in" a system of norms in accordance with one of the major ethical theories. The emphasis of recent ethical approaches to self-driving cars has been on assessing scenarios such as the "trolley problem," which depend on judgments or intuitions about the applicability of a rule derived from an ethical theory to a specific case. We know from the previous chapter that this opens up the Pandora's box of inferential regress but let us ignore this difficulty here. Presumably, as long as AIs follow the rules of an ethical system

and apply them systematically to cases, then they can certainly be ethical, just like us, since their actions will be driven by the same judgments and rules derived from Kantian or utilitarian principles. What exactly is the problem with this approach to moral value alignment? Some authors have recently indicated that given the complexities of value alignment with AI, a multifaceted approach that draws from various ethical traditions is better than a narrow one based on a single theory (Bostrom et al., 2020; Gabriel, 2020). I shall focus now on two preliminary problems, one less significant than the other, and come back to the multifaceted approach in the final chapters.

The first and less significant problem is that the application of ethical rules to specific cases requires very considerable representational and rational requirements. Since AI is not capable of satisfying the most basic of representational and rational needs, it is dubious it will be capable of processing more complex representational needs concerning rule following in ethics and the interpretation of cases in a semantically and morally relevant way (e.g., a person shoots her gun, is she a cop or a criminal?). Suppose an AI is confronted with a trolley problem concerning inevitable death—as is assumed in cases of ethical AI discussions. Are people correctly represented as such? Are pedestrians represented as pedestrians, children as children, trees as trees, and so on? Are the AI's relations to them also part of the representation of the case (are some of them people the AI is designed to protect and thus the generic version of the trolley problem is irrelevant)? Are the consequences of the AI's actions properly represented and calculated? Are the AI's actions properly represented *as causes* of an innocent person's death (assuming the AI decides to kill one person in order to save five lives)? Is avoiding the death of five people also represented as such or is the system simply optimizing on a reward? Are the right reasons properly motivating and informing decision-making?

Ethical rules do not protect humans in every single case, and therefore, their application requires careful assessment and evaluation. For example, killing the innocent is not the same as killing someone guilty of murder; killing is not wrong in self-defense, war, police enforcement, and countries that enforce capital punishment. Police enforcement may itself be a source of evil according to some theories because of its association with forms of financial and racial oppression and yet, can AIs oppose enforcing the law? Capital punishment may be deeply immoral for similar reasons but it is legal in some countries, including the United States. Are we running the risk of designing ethical AI that disobeys a legal system? These considerations show that what is *legal* need not be, and frequently is not, *moral*, and vice versa, as explained briefly in the introduction.

There are multiple problems that derive from social and economic injustice that don't even feature as main tenets of ethical theories. The powerful are frequently exempted from legal and moral obligations that the poor are never exempted from; the poor and oppressed confront difficulties in their daily lives that can themselves be considered intrinsically as moral harms; and so on.

But even assuming that adequately attentive rational and representational capacities become central components of an AI's "unemotional" rationality, there is a second preliminary, yet much more serious difficulty: AI's *incapacity to determine why* is it that humans follow these rather peculiar rules and practices in the first place? Why only some animals, at best, qualify as worthy of legal, social, or moral protection (e.g., pets such as dogs and cats)? Why is it that not every human being is equally protected under the law? Why are some people tortured, surveilled, ignored, chastised, or incarcerated? Why are there so many differences in the treatment of human dignity across different countries, cultures, and legal traditions? Can a colossal database of human behavior really help here? There seems to be no rhyme or reason to these practices if only *representational needs* are considered. But of course we know better than simply saying that our behaviors and practices are the decisive measure or representational gateway to moral value. "What is it that humans *value*?" is not a question that AIs can answer by just gathering behavior patterns across the planet. These patterns need to be interpreted according to *emotional and moral needs*, not merely representational ones.

As mentioned, the value we give to our most important or categorical needs is a decisive factor in human emotional (and general) intelligence. Without a hierarchy of needs organized in terms of what we value the most, it is impossible to explain why would someone sacrifice so many years of her life to achieve the goal of receiving a college degree, forming a family, caring for a loved one, or simply finding spiritual atonement through solitude. Our emotional and moral needs bring homogeneity to the human condition and provide a basis for value alignment. Moral rules, the guidance they provide to legal rules, and their application to concrete cases would be unintelligible without these needs. The underlying motivations for the human practices constitutive of morality and the compliance with legal norms depend fundamentally on the emotional, moral, and social needs for recognition, autonomy, and dignity.

A systemic misunderstanding of the motivations and needs underlying human morality would spell disaster if unemotional AI were in charge of satisfying these needs. This is a problem that generates very serious *social risks*. The problem is not simply that determining the preferences of a human being across her life is

an insurmountable technical problem, but rather, that even if machines were good at tracking preferences reliably they would have no way to interpret why some preferences are higher in the hierarchy of human values, preventing them from adequately solving conflicts among preferences and values. Any articulation of emotional and social needs appeals to our biologically rooted conscious awareness, and here lies the key difficulty. The principles of pleasure-refinement, pain-avoidance, and consequentialist utility-maximization, or alternatively, Kant's categorical principle that protects the kingdom of ends in themselves because they possess reason and autonomy, are based on our *moral sentiments and intuitions*, and these intuitions, in turn, are based on our emotional needs for being treated with dignity, for autonomy, for shelter and care, for recognition, and for authentic understanding.

The conscious intuition of the good, the impression an action creates on our awareness, plays a vital role in moral reasoning. As the previous chapter explained in detail, the situation is very different with respect to the unemotional type of inferential attention required to satisfy representational and rational needs concerning coherence and belief-guidance toward the truth. In sharp contrast to the satisfaction of epistemic representational needs, the viscerally experienced motivation that is behind moral actions must correspond to something genuinely valuable for it to be a source of good. Herein lies the conundrum between objectivist and subjectivist, as well as judgment versus affect-based views in moral theory. G. E. Moore's (1903|1968) proposed solution to this problem in *Principia Ethica* was to postulate that goodness is a non-natural property that we discover through intuition, presumably through the kind of vivid intellectual seeming that Boghossian assumes in his account of inference—people simply "see" that something follows from a premise, or that an action is good. This "seeing" or "seeming" determines normative, rather than descriptive, evaluations. The previous chapter argued that "seemings" are not a good source of epistemic justification. Can they be a good source of moral justification? This chapter answers this critical question affirmatively, because of the importance of genuinely experienced empathy in moral motivation.

However, as Philippa Foot (1967, 2–3) explains, there is substantial tension between the objective nature of goodness and our intuitive subjective awareness of it because some kind of reliable means or method must be involved in objective assessments of the good, and identifying this reliable method is not a trivial issue. This metaethical question, namely, what are the facts grounding our appreciation of moral goodness, needs to be answered based on facts and properties about *human psychology*. Whether we take the view that morality is

based on judgments rather than on affective reactions, or whether moral value is subjective or objective, relative or universal, we must also explain how our choice of moral theory is compatible with, and explained by, human psychology. Value alignment depends on a non-subjective standard for what is valuable that can reasonably be agreed upon, as was argued above in the context of defining categorical needs. The value-alignment problem and the requirement that categorical needs be based on real value, rather than merely subjective responses, tip the balance toward objectivist views. But the main point here is about why, unlike epistemic needs, our conscious awareness is constitutively involved in moral and aesthetic needs—we have these needs in virtue of our biologically rooted phenomenally conscious experiences.

The hierarchy of needs is determined by categorical needs that give meaning to our lives. This suggests the existence of a kind of *motivational penetration* (see Watzl, 2017) explained by how needs are organized. Similar to cognitive penetration, satisfying the most basic representational needs while guaranteeing that higher needs are capable of providing general and malleable guidance requires virtuous cognitive integration. This chapter explains how the relation between phenomenal consciousness and attention is fundamental for the explanation of morally based motivational penetration. Empathy, or the capacity to put ourselves in someone else's perspective, and the familiarity that phenomenal consciousness provides in recognizing feelings and emotions are essential aspects of human psychology that underlie moral reasoning.[2] Because AI lacks the emotional needs associated with empathy and familiarity, the prospect of *autonomous AI moral agents* is bleak.

This is a key clarification. The main claim of this chapter is not that it is impossible to create "ethical AI." Rather, the claim is that *while AI agents may become fully autonomous epistemic agents they cannot become fully autonomous moral agents*. If AI develops something like attention routines, they will satisfy epistemic and representational needs, but the lack of emotional and biological needs prevents them from becoming fully autonomous moral agents. Thus, there will always be a non-negligible degree of risk involved in assigning morally relevant tasks to AI agents.

Moral autonomy in the context of AI is insightfully addressed by Wallach and Allen (2009), who distinguish four layers of moral competences, plotted along two axes, one for autonomy (low and high) and the other for what they call "ethical sensitivity" (also low and high). Ethical sensitivity is achieved in human psychology through conscious attention routines, but in AI it could be accomplished through attention routines that satisfy representational and rational needs—the way Wallach and Allen define sensitivity is consistent with

this characterization. The lowest levels of autonomy and sensitivity correspond to today's machines, some overlapping with the next level of "operational morality." The upper two levels are "functional morality" and "full moral agency." As was emphasized before, there is a deep connection between safety and control through agency. Wallach and Allen give the following example of functional morality:

> The realm of functional morality contains both systems that have significant autonomy but little ethical sensitivity and those that have low autonomy but high ethical sensitivity. Autopilots are an example of the former. People trust them to fly complex aircraft in a wide variety of conditions, with minimal human supervision. They are relatively safe, and they have been engineered to respect other values, for example passenger comfort when executing maneuvers. The goals of safety and comfort are accomplished, however, in different ways. […] Under normal operating conditions, the design of the autopilot keeps it operating within the limits of functional morality. Under unusual conditions, a human pilot who is aware of special passenger needs, for example a sick passenger, or special passenger desires, for example thrill-seeking joyriders, can adjust her flying accordingly.
>
> (Wallach and Allen, 2009, 26–7)

While Wallach and Allen's notion of "autonomy" is more permissive than "*agential autonomy*" (which is the notion that, I have argued, matters for genuine intelligence), "mechanical" autonomy with little moral sensitivity reveals how navigational systems satisfy passenger needs in a lucky or risky way—the system is reliable, but had the program being slightly different the system would have no regard for the more specific needs of passengers. If the program were based on rules, the AI system would not "care" if the rules were Kantian, utilitarian or consequentialist, or even ethical rather than practical. The AI self-navigational system will simply execute whatever task it is given. Only the human pilot understands that these rules satisfy moral and biological needs, based on *her own* needs. Crucially, the human pilot has an understanding of how to rank the needs of thrill-seeking, illness, and comfort in an *autonomous way*, by herself, and based on her own cognitive architecture and hierarchy of needs. She knows that caring for the ill, for instance, is more important than satisfying thrill-seekers. AI can only achieve a representational and *heteronomous* understanding of these needs. Thus, agential autonomy is deeply important for *non-risky control* as opposed to mere *machine-reliable control*. More important, even if AI became autonomous *epistemic* agents, they would still lack the *emotional and moral* needs required to become autonomous *moral* agents.

4.2 Machine Consciousness?

This section distinguishes the present criticism of AI, as a limited moral agent, from John Searle's arguments against AI. It does so on the basis of evidence in support of the dissociation between consciousness and attention.

One of the most compelling topics currently under debate is machine consciousness.[3] There is a growing number of articles in magazines and newspapers that discuss related advances in AI, from self-driving cars to the "internet of things" where common household objects can be intelligently connected through a centralized control system—these are systems that, as the example of the autopilot by Wallach and Allen demonstrates, could be considered in some contexts as "ethical." Along with these advancements is a growing fear that we may be creating intelligent systems that will harm us (Rinesi, 2015). This topic has been addressed in many settings, from international conferences to popular books (e.g., Bostrom, 2014; Brooks et al., 2015). Some think that the so-called "singularity" (the moment in which AI surpasses human intelligence) is near. Others say that there is now a Cambrian explosion in robotics (Pratt, 2015). Indeed, there is a surge of AI research across the board, looking for breakthroughs to model specific forms of intelligence, including the capacity for emotional intelligence and learning. Michael Graziano (2015), for example, has claimed that artificial consciousness may simply be an engineering problem—once we overcome some technical challenges, we will be able to develop consciousness in AI.

Can morality become part of AGI? As stated above, a negative case is defended here based on the lack of *phenomenal consciousness* in AI. While we may be able to program AI with aspects of human moral reasoning, as the autopilot example shows (e.g., "do not cause bodily harm," "do not steal," "do not deceive"), we will not be able to create actual emotions by programming rules into monitoring and control systems—at best, these will always be simulations of real emotions. Since human moral reasoning is based on emotional intelligence and empathy, this is a substantial obstacle for the development of morally safe AGI, because simulated emotion can be dangerously manipulative.

The present criticism differs considerably from John Searle's arguments against AI (1980, 1998). Searle famously criticized AI because of its incapacity to think with *intentionality* (i.e., the feature of mental states that makes them *about something*, essentially relating them to semantic contents). Searle takes intentionality to be a capacity that only conscious agents can have. He also

argues that a consequence of this criticism is that phenomenal consciousness is *necessarily* a biological phenomenon. Searle, therefore, takes the limitations of AI to be principled or a priori ones that will not change, regardless of scientific progress. Critics have argued that the claim that only biological beings can have intentional minds may be defeated (e.g., see Block, 1995a) and that cyborg systems or an adequate account of how the brain computes information could refute Searle's "Chinese room" thought experiment (Churchland and Churchland, 1990; Pylyshyn, 1980). These criticisms have merit, but they are applicable only to the kind of consciousness that Ned Block (1995b) calls "access consciousness." Thus, there is a very important ambiguity in this debate. While Searle is right that phenomenal consciousness is essentially a biological process and that AI is severely limited with respect to reproducing this kind of biologically rooted consciousness, his critics are right in claiming that AI may be capable of simulating and truly achieving *access consciousness*—AI will be intelligent if they become attentive to properly integrated rational and representational needs. This is why the consciousness and attention dissociation (CAD) is crucial here, because it entails that attention is essentially related to access consciousness (Montemayor and Haladjian, 2015).

The present criticism of AI, therefore, is more nuanced than Searle's in three important respects. First, it concerns exclusively the type of consciousness that is characteristic of feelings and emotions, independently of how they are related to semantic contents or conceptual categories (i.e., phenomenal consciousness). Second, the limitations of AI regarding phenomenal consciousness are independent from considerations about understanding the meaning of sentences—animals, for instance, experience vivid emotions without having the range of concepts we apply in emotional judgments and categorizations. Other biological species, which do not manifest the capacity for language but which very likely have phenomenal consciousness, share our emotional and metabolic needs. Thus, the present criticism based on CAD is more faithfully based on biological considerations than Searle's. Finally, and quite importantly, AI may simulate intelligence, rationality, and linguistic behavior successfully; however, they will not experience feelings or emotions in the same way as humans. This implies that AI agents lack *moral standing* even if they are autonomous intelligent agents, assuming that experiencing emotions and feelings is a *necessary condition* for moral standing.

Some would object to the distinction between access and phenomenal consciousness, or like Searle, to separating intentionality from phenomenality. But these distinctions and assumptions provide several advantages over other

views, including Searle's, both from an empirical and theoretical point of view. One advantage is gaining clarity by avoiding the ambiguity aforementioned. Another key advantage is the emphasis on empathy. Empathy and the intensity of emotions have not been considered as central in criticisms of AI. This is a puzzling situation, given the importance of phenomenal consciousness for human empathy, moral standing, and moral behavior. Surely, the intrinsic moral value of consciousness should be a fundamental component of proposals about artificial consciousness. Taking the moral value of consciousness seriously, however, shows that the prospect of artificial consciousness is bleak.

To fully appreciate how the intrinsic moral value of phenomenal consciousness matters to AI's limitations, consider the fact that although semantic information can be *easily copied*, and programs with syntactic features may be reproduced many times, in principle to infinity, it seems clear that the way a subject experiences the intensity of an emotion from her subjective perspective cannot be replicated or copied at all. This irreproducible uniqueness might be the most important aspect of phenomenal consciousness. It certainly seems to be more important than the fact that the mind relates to semantic contents (e.g., see Aaronson, 2016). This is why the focus here is not on semantics, but on the importance of emotions and their intrinsic normative value.

While the idea that phenomenal consciousness cannot be realized in machines is intuitive, there are reasons to explore this issue further and more carefully. Advances in AI are quickening in pace, and as software and hardware technologies continue to progress there will be increased accessibility to more powerful machines that can perform more sophisticated computing. In the field of biocomputers, there are even developments of using enzymes to create "genetic logic gates" (Bonnet et al., 2013) that could be used to build biological microprocessors for potentially controlling biological systems (Moe-Behrens, 2013). As mentioned in the introduction, evolutionary and biologically based models are already inspiring the fields of open AI and unsupervised learning (Lehman and Stanley, 2008). If we use living materials to build and run software, how are we certain that such organic-based technologies are not going to be conscious eventually?

As long as the emphasis is on moral autonomy and the satisfaction of emotional and metabolic needs, a clear case can be made against the mere simulation of emotion by AI. Obviously, it could be the case that genetic manipulation eventually allows us to reproduce many aspects of moral intelligence, but in that case we would be creating new forms of *life*—whether this type of investigation

is itself morally acceptable is controversial. Moreover, CAD shows that although epistemic agency is deeply related to moral reasoning, the two are dissociable in humans, as the research on empathy confirms (Zaki, 2017). Perception and emotion modulate each other and give rise to many forms of cognitive skills associated with human intelligence (Pessoa, 2013). One may suppose that if artificial agents pass not only Turing intelligence tests but also *emotional* Turing tests (Picard, 1997; Reichardt, 2007), artificial agents may achieve a level of conscious awareness similar to human beings. In fact, according to the most optimistic interpretation of AI research (e.g., Kurzweil, 1999), artificial agents may become sources of ethical and rational flourishing because they would not be subject to the biological constraints that human beings inherit from their genetic lineage, thereby enhancing the possibilities for improvement in ways that are impossible for us mortals.

The implications of CAD for this debate are crucial here. Since human visual attention is now increasingly used to examine the nature of conscious experience and the kind of intelligence AI is supposed to achieve, it is critical to understand *how* consciousness and attention are related. An examination of this relationship shows that there is a strong case for dissociation between attention and consciousness in humans. The basic forms of attention do not require consciousness to operate successfully. Perception is supported by many attention routines that operate outside of phenomenal consciousness (Cavanagh, 2004). According to CAD, the proper functioning of attention routines would not *entail* conscious awareness even in humans. This means that even if AI reached similar or superior levels of intelligence based on attention routines, machines would still lack consciousness given that they have no biologically rooted needs, which are essential for emotional intelligence.

In support of a dissociation between consciousness and attention, consider that the sort of phenomenal consciousness that is experienced by humans must be a more recent advancement in evolutionary terms (Haladjian and Montemayor, 2015). Abilities related to the selective processing of visual information concerning location, color, shape, and motion are basic for survival and found in many animals. These can be thought of as modules of perception that can be activated based on the environmental and task demands (Pylyshyn, 1999), and can be described as attentional routines (Cavanagh, 2004). From a computer science perspective, the halting problem (i.e., the termination of a function when its goal is complete) is achieved by these attentional abilities on the basis of their evolutionary purpose and proper function, which is to satisfy

representational needs. AI programs execute shape detection, object tracking, and face recognition routines, but without, at least so far, any autonomous representational needs—their "needs" are imposed by us. This may change, and if so, they would count as epistemic agents. But the difficulties surrounding consciousness would remain. Phenomenally conscious experience is constant, visceral, "homeostatic," *does not halt,* and has no specific informational goal. It is deeply related to vitality. Through its interaction with attention, color perception obeys the rules of halting and switching a task (e.g., search for blue, now search for red). But we remain phenomenally conscious while attention switches tasks and halts searches. Consciousness naturally "switches off" only because of the vital function of sleep.

Another point related to evolution is that *dexterous* complex actions, which were genetically designed from millennia of evolution, are notoriously difficult for AI and machines to simulate. Using a familiar example, one can program a computer to beat any human in the game of chess, but it is very difficult to program a robot that could dexterously move the pieces of the chessboard like a human. This idea is related to Moravec's paradox: while abstract and complex thought is easy to compute, basic motor skills are very hard to model computationally. Hans Moravec (1988) explained this puzzling asymmetry precisely by appealing to evolution. Our species had millions of years to develop finely tuned attention-integrated skills, which operate unconsciously or automatically, while complex rational thought is a recent addition to our cognitive abilities. This line of reasoning must be carefully considered. One critical consequence of developing this point is that *conceptual* and explicitly language-based conscious attention must have evolved later than basic perceptual attention (Haladjian and Montemayor, 2015).

Yet, one does not need to accept these arguments to appreciate how CAD makes the unqualified proposal for AGI problematic because there are at least two kinds of cognitive needs: epistemic and moral. The main issue here is that *while simulated intelligence may be intelligence, simulated emotion cannot be emotion*. Turing's child machine is a good epistemic agent but her moral cognition is a potentially manipulative simulation (a "deepfake") of the viscerally felt basis of moral sensibilities in humans and animals. This limitation is more nuanced than completely denying intelligence to AGI on the basis of a lack of biological or metabolic functions, the way Searle does. In fact, it grants that it is perfectly fine to assign intelligence to AGI; it only denies that they have consciousness and the moral needs that depend on subjective awareness.

4.3 Intelligence Equivalence: Visceral and Algorithmic

This section expands on the scientific findings on attention and its interaction with consciousness, focusing on the extensive evidence in vision science. It justifies the distinction between Extensionally Equivalent Intelligence and Intensionally Equivalent Intelligence: two very different styles of Artificial Intelligence.

The human brain has been compared to a computational system by a vast number of psychologists and neuroscientists in attempts to better understand how it works. A crucial component of the field of cognitive science grew out of this tradition (e.g., see Pylyshyn, 1984). Scientists continue to explore the relationship between the brain and computers, which has generally taken the form of computational models of brain processes that led to a better understanding of perception and cognition (Reggia, 2013). This analogy of mind and machine also drives innovations in technology that aim to achieve human-like performance. Many insights from neuroscience and cognitive science are now fundamental for modern AI. A limitation of this approach, as mentioned, is that attention is more "algorithmic" than consciousness, so their interaction, given CAD, needs to be elucidated.

Animal minds are capable of performing many impressive actions and calculations based on attention routines. The complexity with which these tasks are achieved, in order to satisfy various cognitive needs, still defies AI implementation. Attention has been studied extensively in the visual system. In general, attention can occur automatically (e.g., from exogenous stimuli) or be more willfully directed (e.g., from endogenous sources). As explained before, what all kinds of attention have in common is that they can work toward the goal of selectively processing information in relevant ways to allow an organism to interact with its environment and satisfy its most basic representational needs. Regarding neuroanatomy, feature-based attention is a more primitive information selection mechanism related to low-level perceptual processes. These information processing systems are organized according to specialized brain regions responsible for registering specific types of visual information, such as color, motion, or segment orientation (for a review, see Maunsell and Treue, 2006). Feature-based attention interacts with these low-level systems to select information in a typically automatic manner, but this selection process can be biased by higher-level signals based on task demands, including motivations, as illustrated in Chapter 3.

Animals fundamentally depend on spatial attention, from different sense modalities and integrated through their agency and needs. This kind of attention to space probably evolved for predatory purposes and then increased in complexity during the Cambrian explosion. Spatial attention is a necessary condition for object and feature detection, and it also seems to require a minimum sense of "consciousness" of one's perspective with respect to the immediate environment, although it is highly contentious to claim that this is genuine conscious awareness. Attention must somehow be directed toward objects and features as occupants of regions of space. The attention "spotlight" can be focused on a specific region and shifted around as needed (Posner et al., 1980), and can be made more diffused or a more focused "zoom lens" (Eriksen and Yeh, 1985). Distributed attention can capture quickly a statistical summary representation of the information outside of the focus of attention (e.g., Alvarez and Oliva, 2008), which also helps compute the overall properties of a visual scene. Computer vision has done a decent job of simulating both feature-based attention and spatial attention, though not as efficiently as the human visual system (e.g., Yang et al., 2011). Attention can also operate on object-like properties, such as cohesion, symmetry, and common fate (for reviews, see Chen, 2012; Scholl, 2001). Object-based attention requires a two-stage process that begins with the individuation of objects (Pylyshyn, 2000).

These are all the basic or "bottom-up" kinds of attention behind the tasks used by Leibo et al. (2018) in their experimental lab where AI designed by the DeepMind team is compared with human performance. Although AI falls behind in performance for all these basic attentional tasks, even the most basic kinds of life forms, insects for example, are incredibly good at these tasks (Gallistel, 1990). Evolution generated attention routines and subroutines that satisfy the representational needs of biological agents as they inspect various environments at multiple scales. Animal navigation can be oriented spatially and temporally, involving memory, motor-control, decision-making, and thought (Gallistel, 1990; Montemayor, 2013), as well as some rudimentary but genuinely categorical capacities, such as caring for conspecifics. Attention in the wild allows for demonstrative reference, or "pointing" mentally at terms that are followed through a kind of symbolic indexical marker for space and time. These kinds of intentionality or "aboutness" need not involve phenomenal consciousness.

With respect to cognitive integration, selective attention operates upon these "indexed" items in order to bind object features, which are made available through feature maps (Treisman and Gelade, 1980), resulting in

sustained object-based mental representations that allow object identification and meta-selection, including the conceptual categorization of the object as falling under a conceptual kind (Keil, 1992). Thus, attention operates by *analyzing* and also by *integrating* contents from different sense modalities and specialized brain regions—a process that involves selective hypotheses and meta-hypotheses. Together, individuation and identification support abilities like enumerating sets of items, tracking multiple objects, or attending to a single item in detail. Selective attention plays a crucial role in forming persisting object representations by allowing features from a visual scene to build and maintain a coherent representation incrementally in visual memory (Treisman, 1998, 2006). These mid-level "object file" representations (Kahneman et al., 1992) are generally considered the product of object-based attention. Another version of an object file is an "event file," which incorporates both features and motor commands (Hommel, 2004, 2007; Zmigrod et al., 2009). This richer notion of object representations can combine cross-modal sensory representations and also integrates action-planning information. Animals rely on this kind of cross-modal attention to sustain attention to objects for long periods of time.

These forms of attention constitute epistemic virtues or excellences of epistemic agents who satisfy incrementally complex representational needs, including inferentially integrated and highly skilled inferential routines and subroutines. They are also automatic and often produced without any conscious awareness (Dehaene et al., 2006; Mudrik et al., 2014; Mulckhuyse and Theeuwes, 2010; van Boxtel et al., 2010). There are many examples of attention processing that do not reach awareness but yet still allow individuals to perform actions successfully, as in the case of blindsight (Kentridge, 2012; Kentridge et al., 2008). In principle, AI could develop search routines robust enough to simulate these forms of attention, although Moravec's paradox remains a caveat for the implementation of the mechanisms supporting these forms of attention, particularly regarding attentive integration with motor-control and decision-making. Crucially, evidence supports the claim that these fundamentally navigational kinds of attention occur unconsciously in humans and animals.

Additionally, there is conceptual attention, which can also occur automatically and unconsciously in humans, but requires routines guided toward semantic propositional contents that would be difficult for AI systems to emulate (Marcus and Davis, 2019) and which depend on much more complicated kinds of hypotheses and meta-hypotheses. Exactly at what moment conceptual attention plays a role in object-based attention is a subject of debate, but it clearly plays a critical role in higher-level human visual attention. Machine learning programs,

such as those implemented in object recognition, *simulate* implicit conceptual attention, although required caveats are needed here as well. The point is, were AI to simulate these attention routines to a humanly satisfactory degree, they would count as *extensionally equivalent* forms of intelligence, as opposed to intelligence that is exactly equivalent in content, function, and *agential control* or *intensionally equivalent* intelligence.

> *Extensionally Equivalent Intelligence (EEI)*: AI will be capable of intelligent behavior only if it simulates attention routines, even if these attention routines are not satisfying the AI's own representational needs and therefore are performed *without autonomy*.
> *Intensionally Equivalent Intelligence (IEI)*: AI will be capable of intelligent behavior only if it simulates attention routines. If the AI agent does so in order to satisfy her own representational needs, then she will be an *autonomous epistemic agent*, like animals and humans, *even in the absence of human-like phenomenal consciousness*.

An agent that satisfies her own representational and rational needs because of her skills has a level of understanding that an agent that simulates those skills but satisfies alien needs (or none at all) completely lacks. Using the terminology from the literature on inference, an agent that satisfies her own epistemic needs gains an *appreciation*, even if implicit, of why and how her epistemic outcomes are justified. This is the key difference between EEI and IEI. If the capacities of two agents are genuinely satisfying their needs based on their attention skills, then they are robustly equivalent. EEI and IEI are, therefore, *two radically different forms of AI,* in spite of the fact that *"from the outside" they may be thought of as functionally equivalent*. The contribution of this chapter is to explain why although EEI and IEI should count as *intelligent* (although EEI is considerably more *risk-involving* than IEI because EEI does not satisfy agential needs, and a hierarchy of needs brings stability to preference and value alignment), *neither of them can count as morally autonomous*. Only the possibility of AI epistemic autonomy, through IEI, seems possible.

Moral autonomy is a fundamental advantage of the kind of integration that only consciousness can provide to human and animal cognition. The virtuous integration of epistemic and moral agency combines the unemotional intelligence of the child machine with our ancestral, biologically based emotional and biological needs. While IEI could implement human-like cognitive penetration in the integration of epistemic routines, it will not be capable of implementing human-like motivational penetration, rooted in biology. Thus,

the overall hierarchy of needs of IEI AI will be extremely different from ours, making the value alignment problem quite difficult with respect to moral risk in particular.

Besides moral autonomy and care, phenomenal consciousness brings vitality to the selective and learning-oriented capacities of attention routines. Human agency and motivation of a conscious kind involve different kinds of integrative attention. Some tasks require an engaged and sustained effortful attention, for example in learning a new skill (Meuwese et al., 2013). But highly skilled attentional performances can be so engrossing that they feel effortless (Bruya, 2010). This effortless attention is particularly relevant for *emotional engagement* because it concerns expertise and the activities we value the most (those at the top of our hierarchy of needs), and which take years of sacrifice to develop. The feeling of "flow" associated with effortless attention in skilled performance intensifies the focus on the mechanics of a physical activity with very little effort; flow also occurs in complex intellectual tasks such as solving a mathematical problem (see Csikszentmihalyi, 1997). This kind of attention, therefore, cannot be reduced to search routines and selection processes, and requires levels of cognitive integration that may prove too challenging for AI because they are consciously integrated. The feeling of effort, however, is presumably irrelevant for AI. The real challenge is how to understand *agential motivation*, which is fundamental to attention routines, in AI systems. EEI and IEI are two paradigms of AI. This distinction clarifies how there are superficial and deeper aspects of general intelligence on the basis of agential and non-agential AI.

Humans and animals have the ability to attend to different mental states at once. For example, attention to emotions concerning features of the environment depends on an older neural network for immediate action and arousal—*homeostatic-related* systems are deeply associated with basic metabolic and biological needs. Attention to features, such as color, can occur in unison with attention to emotions, but the neural correlates of these different networks can be distinguished as independent from each other (Pauers et al., 2012). Therefore, even if we focus just on color, there is a network that satisfies *representational needs concerning color,* and a different network that satisfies *emotional needs associated with color*. These are different networks unified by the agency of attention.[4] AIs may be able to recognize color better than humans and make complex decisions based on such recognition, but they will lack the emotions that humans feel when, for example, they look at a beautiful sunset. IEI AI could count as a genuine autonomous intelligence when it comes to epistemic agency, but not moral or aesthetic agency.

4.4 Emotions and Phenomenal Consciousness: Moral and Biological Needs

This section builds on the distinction between EEI and IEI, and it explores various kinds of awareness and their relevance for the satisfaction of emotional and moral needs, drawing on empirical research on consciousness. It also explains further the notions of attentional cognitive penetration and conscious motivational penetration.

As mentioned, access consciousness depends on basic forms of attention, since attention serves the role of processing information within the brain in a task-relevant manner to guide action or thought. In fact, access consciousness is best understood as rationally and representationally integrated attention (Stoljar, 2019). Thus, IEI is a type of AI that could in principle qualify as *access conscious*. Since what it is like to be you, or your conscious awareness, depends on your unique first-person perspective (and your emotional and biological needs), AI seems to be incapable of being *phenomenally conscious*. Yet another form of consciousness is reflective self-awareness. Whether or not animals possess self-awareness remains debatable, but some have proposed ways in which basic consciousness might be identified in animals independently of capacities for reflection (Edelman et al., 2005; Seth et al., 2008). Authors appeal to various criteria to demarcate conscious from unconscious intelligence. Tim Bayne's (2007) theory of "creature consciousness" specifies whether or not an organism can be said to be phenomenally conscious by requiring mechanisms that generate the "phenomenal field." Problem-solving behaviors like tool usage may provide some of the best indication of the possible presence of conscious attention in animals (for a review on animal consciousness, see Griffin and Speck, 2004). It is likely that such behaviors, however, could depend just on access consciousness. In any case, reflective self-consciousness can be distinguished from other forms of consciousness, and further complicates the issue of agency and motivation for consciousness and attention.

Several empirical studies on consciousness describe the structures that likely support it. In general, the brain requires some level of recurrent processing and not just the feedforward movement of information (Di Lollo et al., 2000; Seth and Baars, 2005; Tononi and Koch, 2008). Complex neural networks with recurrent processing, especially those that have signals originating from the frontal cortex, are considered later adaptations. More deliberate forms of conscious attention also are associated with activations in the "newer" brain areas like the prefrontal

cortex, and are supported by working memory systems (but there is no conclusive evidence that the frontal cortex is necessary for conscious awareness, even though it might be necessary for human-like intelligence).

Consciousness can be described as having different levels of activation, with some events remaining "preconscious" and others entering full awareness (Dehaene et al., 2006). These ideas are presented under models such as the "global workspace" or "broadcast" views of consciousness (Baars, 2005; Dehaene and Naccache, 2001). Under these accounts, the main adaptive purpose of conscious attention is to "broadcast" contents that are computed in a uniform format (presumably conceptual). This plays the important epistemic role of accessing contents across different modalities and supporting goal-oriented actions. Consciousness could be crucial for the overall or "holistic" integration of information from different modalities (but some argue that integration can occur outside of awareness, e.g., Mudrik et al., 2014). Although we are far from fully understanding how the brain supports consciousness or what its evolutionary purpose may be, there has been progress regarding what we know about it. While consciousness is closely tied to attentional processes, consciousness and attention cannot be simply reduced to one another. Attention often operates outside of consciousness (Koch and Tsuchiya, 2007) and most likely appeared before consciousness (Haladjian and Montemayor, 2015).

Emotions are the foundation of morally relevant experiences and judgments. They regulate mental states in order to produce morally salient behaviors regarding empathy toward others, the pleasures and pressures of social bonding, and the satisfaction of biological needs (e.g., producing positive emotional states in rewarding situations like mating or eating, or anxiety in response to fear). Emotions also influence changes in physiology and bodily states, such as the quickening of the heartbeat, pupil dilation, and tensing of muscles. Thus, emotions are closely related to the neurophysiological state of the brain and body through the nervous system, and can be critical in influencing the ability and the manner in which we act. Emotions also are a large *phenomenal* component of conscious experience and subjectivity, tightly connected to metabolic function and vitality (Damasio, 1994).

The previous distinctions between three possible types of consciousness illustrate three ways of cognizing emotions: by *experiencing* them (phenomenal consciousness), by *accessing information* through them (access consciousness), and by *attributing them* to oneself and others through a judgment (self-consciousness and third-person attribution through a "theory of mind"). This tripartite distinction correlates with the three forms of empathy described

above (e.g., emotional, cognitive, and caring; Weisz and Zaki [2018] call these types of empathy *experience sharing, mentalizing, and empathic concern*). It is the *experiencing* of emotions through phenomenal consciousness and how it immediately affects empathic concern toward others that presents serious challenges for AI. Our conscious lives are the basis of *valuable and worthy projects* that have their roots in our morally emotional autonomy (e.g., listening to a moving piece of music, admiring a sunset, tasting a very good wine). We share with animals many of our emotions, including the morally crucial emotion of empathy (de Waal, 2019).

Emotions include basic "primary" responses such as fear and anger (Ekman, 1992) and more complex evaluations of situations through feelings of remorse, resentment, and gratitude, as well as empathic responses with moral implications (Decety and Cowell, 2014). The circuitry that contributes to primary emotions is evolutionarily older than the circuitry of long-term and sequential planning, and is present in animals that can display fear reactions (LeDoux, 2000, 2012). While animals do exhibit basic emotions, whether or not they have the same subjective "feeling" that we do is difficult to determine. Nevertheless, it is worth pointing out that animals share similar physiological reactions as humans, at least within the context of basic emotional responses that rely on similar brain circuitry, and therefore are very likely conscious of those emotions.

Human emotional states are based on experiencing "feelings," which are consciously processed (Tsuchiya and Adolphs, 2007) and inferentially integrated with memory and long-term planning. While independent, the neural systems that support emotional processing in humans are closely tied to those responsible for cognition (Pessoa, 2008), so emotional feelings can be considered as a more advanced form of mental activity, especially when it concerns moral and aesthetic judgments. Feelings also seem to be fundamental to the sense of self (Damasio, 1994), which further suggests that the presence of feelings requires higher-level integration of basic emotions and cognition that perhaps only phenomenal consciousness can provide. The *cognitive role* of phenomenal consciousness could thus be to integrate epistemic and moral value through motivational penetration, heavily influencing and structuring the shape of an agent's hierarchy of needs. Similar to the integrative structure of inferential attention in cognitive penetration, motivational penetration can be understood in terms of a hierarchy of emotional needs and feelings, some more immediate and biologically driven, others more deeply related to higher thought, self-awareness, and long-term projects.

The neuropsychological evidence (for review, see Pessoa, 2008) shows that the limbic system, particularly the amygdala, is closely associated with generating emotional responses and seems to be the only region that can "hijack" the cortex and take over its computational, rational processing (LeDoux, 2000, 2003). Fear responses *engage* an organism, focusing attention on critical aspects of the environment, taking over cognitive activity of a more "unemotional," representational and rational kind. Stimuli with emotional content tends to activate more extensive cortical areas of the visual system related to attention (Pessoa, 2013; Pessoa et al., 2002), and may even be detected *outside of conscious awareness*, or at least has a very low threshold for detection with minimal awareness (Mitchell and Greening, 2012; Pessoa, 2005). Emotional systems concerning these circuits are found in both animals and humans (LeDoux, 2012). For example, the fear response a housecat displays plays the double role of a threatening stance toward the aggressor and a signal to be extremely alert. Conversely, the blissful state of a purring feline conveys trust and relaxation. These basic emotional states prepare an organism to engage in different ways. There is something it is like for the cat to undergo these states (phenomenal consciousness) and there is also information that is being used for immediate action and decision-making (access consciousness). The experience may be radically different from the information it conveys (e.g., the experience of fear is intensely vivid, even though the information it conveys may differ widely, from real and imminent harm to foreseen and conditional expectations).

Although emotions are an important part of our conscious experience, research suggests that some emotional cues can be processed non-consciously (Tamietto and de Gelder, 2010), which calls into question the overall role of awareness for functional responses to emotions (Pessoa, 2005). Notice, however, that this is fully compatible with the architecture of inferential attention and cognitive penetration that a child machine could develop—a system of "attention without emotion," which is characteristic of cognitive empathy. Along these lines, researchers have proposed that conscious experience is a central signature of feelings, but not necessarily of emotions (Tsuchiya and Adolphs, 2007). There might be access to emotional information without consciously feeling an emotion and there may also be unconscious attention routines that process such information and affect behavior without producing any specific type of awareness. A vivid example of the dissociation between a vital emotional response and the subjective feeling associated with it is seen in people who are born with a congenital insensitivity to pain (Heckert, 2012). This neurological condition prevents the individual from subjectively feeling pain, and the

associated feelings of fear and aversion. It is impossible for these individuals to *understand* or appreciate from their subjective point of view what pain is like, and what others might feel when experiencing pain, preventing them from *experiencing* empathy. But these patients learn how to *draw inferences* about appropriate behavior regarding their own bodily damage and how others react to pain. Access to information about bodily damage can occur, therefore, without the experience of pain.

Nonetheless, feelings and emotions are paradigmatic forms of phenomenal consciousness, and philosophers frequently appeal to conscious awareness in order to ground moral and aesthetic value. What are the implications of these distinctions for AI? A technical or "descriptive" issue mentioned before is that all functions and attention routines for emotion-simulations in AI would depend on *halting thresholds* (i.e., the point where a computer program should stop). But by definition, phenomenally conscious states are not reducible to such routines and have no real "halting threshold"—experiences are vividly engaging without having an obvious algorithmic output. The experience of regret, for instance, is not simply the end point of running an attention routine. On the contrary, it is a complex state that cannot be reduced to any simple halting function. This is partly why emotional and moral needs can serve as the foundation for categorical desires that inform most or all of our lives. The essence of these experiences is to engage the subject as a whole, rather than to arrive at a specific conclusion—they affect and shape the subject's entire conscious awareness. The *normative* implication of this engaging kind of integration is that it exerts a rather powerful influence on the structure of an individual's hierarchy of needs.

4.5 Empathy and Moral Reasoning

This section elaborates on the limitations of EEI and IEI artificial agents in the realm of morality, particularly the impossibility of fully autonomous moral AI agents. It discusses problems related to AI's lack of moral autonomy concerning the representation of emotions and the risks associated with it. Conscious integration is exemplified with emotional and recognitional color capacities.

Empathy is the capacity to feel or understand the subjective perspective of conspecifics, and it is fundamentally related to morally salient emotional responses. This requires the agent to have some type of theory of mind and an understanding that similar agents will possess analogous emotional states and

needs, organized roughly in similar ways. Humans cannot empathize without knowing this relationship between self and others, which appears developmentally in a child's second year of life (Zahn-Waxler et al., 1992). The ability to fully empathize relies on self-consciousness and self-recognition, which also develops around two years of age (Rochat, 2003; Rochat and Striano, 2000), but this cannot be a complex or strictly representational kind of self-awareness, since empathy is found in many animals (de Waal, 2019). This form of consciousness seems to develop with experience, particularly of the social kind. Some argue that this ability for social perception is the basis for consciousness in general (Graziano and Kastner, 2011). While machine learning in AI may simulate attention, it is questionable whether or not it can ever develop genuine empathy (e.g., see Miner et al., 2016). More precisely, using the tripartite distinction mentioned before, while AI may simulate empathic cognition through inferential attention (or mentalizing) it cannot succeed at experiencing empathy or empathic care.

Since moral reasoning is at least partly based on this ability to experience empathy, ethics and morality seem to necessitate phenomenality. Displaying intelligence does not necessarily require phenomenal experiences—they happen in tandem in humans, but they can be, and generally are, dissociated (Montemayor and Haladjian, 2015). This is a critical point for understanding the challenges that confront ethical AI. It is one thing to be capable of *detecting* emotions and running rule-based algorithms to reach a conclusion at a halting threshold. It is an entirely different achievement to be able to *empathize* with others based on how we feel. Aesthetic judgments are related to moral judgments, and also require phenomenality in human psychology (the next chapter expands on this). Although it may be controversial to claim that moral and aesthetic evaluation necessitates conscious experience, this is a central assumption in most theories of moral and aesthetic value. Consider the classic utilitarian principle that one must reduce the amount of pain and maximize happiness or well-being, or the Kantian principle that human life is intrinsically and categorically valuable, not just instrumentally valuable, or Arpaly's example of how Finn ignores his "epistemic autonomy" in order to act based on his emotions of care for Jim. The key point is that *without emotional experiences there is no possibility for moral AI that qualifies as IEI; the only possibility for ethical AI is EEI moral agency. Equivalently, there is no possibility for morally autonomous AI.*

There is another related problem for AI agents concerning the representation of emotions as part of a larger informational background (e.g., Picard's affective computing account). Because of considerations concerning utility and value (see Kahneman and Thaler, 2006), it seems plausible to conclude

that the emotional background of an individual is *narratively* structured and not just utility-based. Research shows that the more utility-based a person is, the less inclined she will be to *attend* and respond quickly to morally relevant stimuli (Haidt, 2007). One should not take this evidence as confirmation of an ethical or moral perspective—psychology *describes* phenomena while morality *prescribes* actions. But there is undoubtedly a fundamental and even constitutive conceptual relation between phenomenal experiences associated with moral feelings of approval or condemnation and any conceivable moral theory (Carter and McBride, 2013). Since autobiographical narratives depend on conscious memory (Montemayor, 2018), this presents yet another obstacle for the development of ethical AI because narratives are temporally open and holistic representations.

If the dissociation between consciousness and attention is taken into consideration, one can easily see that AI epistemic agents will be able to reason their way through the inferences of a developed ethical theory, but still lack the wherewithal for responding appropriately in specific cases because they lack the vital context that conscious experience provides. This problem, reminiscent of inferential regress, creates a very unique type of risk because without having genuine emotional needs, AI agents have no understanding of moral autonomy or the way these needs are *prioritized*. Choosing an ethical theory is not the main obstacle for the implementation of morality in AI (although it is a mighty obstacle). The fundamental obstacle is the subjective, empathic nature of moral experiences. Having such a theory, therefore, is not what is most distinctively human about morality. Rather, human morality is composed of our biologically rooted reactions to the pain and suffering of others. More precisely, our moral reactions *express* who we are as autonomous moral agents—they are not mere knee jerk-like reflexes triggered by representations. Moral responsiveness is sensitive to our needs and the needs of others, and ultimately, to the mutual expectations we typically assume as morally adequate (Strawson, 1962).[5] Ethical theory and search algorithms for detecting emotions are good for *simulating* and potentially *inferring* ethical behavior. But because of CAD, AI agents are incapable of genuine or autonomous moral agency. Simulating emotion is manipulative, risky, and unrelated to autonomy.

A further complication is that the most interesting forms of conscious experience for evaluative emotion may not be the mere sum of specific attention routines. The vivid power of a moral, aesthetic, or transformative experience is not reduced to voluntary effort and the successful detection and processing of, for instance, the colors of a sunset over the harbor. This

would be a rather poor and inaccurate description of what moral and aesthetic experiences are, as typically experienced by humans. Such experiences are the result of reliable attention routines and an overall affective and valence-based reaction that combines conceptual information with emotions. In fact, the integration of conscious motivation and attentional control permeates human psychology.

As mentioned, color vision presents a remarkable example of capacities that are typically integrated but can also become dissociated. Trichromatic color vision depends on a single genetic addition of a third light-sensitive protein called opsin, rather than neural modifications during cognitive development (Mancuso et al., 2009). With the single introduction of a new gene, a brain that was not habituated to respond to a whole range of color gains the ability to identify and react to new colors. Let us call this ability *recognitional color vision*. Pauers and colleagues (2012) identified a different cognitive network, dependent on melanopsin, that independently regulates *emotional reactions* and circadian regulation involving color, and which is evolutionarily older than recognitional color vision. Let us call this capacity *emotional color vision*. Melanopsin can influence the circadian system, which consequently affects emotional regulation (Tucker et al., 2012), even when the cones and rods in the eyes are "disabled," for example, when there is natural degeneration of photoreceptors or in laboratory conditions. Remarkably, the visual system communicates with the limbic system (deeply associated with basic emotions) through a different network from the one it uses for color detection (Mancuso et al., 2009; Pauers et al., 2012). While our capacities to distinguish specific shades of color may differ from person to person, our emotional responses are independent from these capacities.

Even assuming that color recognition in AI could exactly simulate cone-color detection, AI's "experiences" of color would be very unlike those experienced by humans, which are also integrated with visceral emotions. Emotional color-alignment would, therefore, be entirely superficial or, more precisely, extensionally, rather than intensionally equivalent for emotional purposes. More generally, the CAD framework entails that experiencing an emotion will not just be a matter of *recognizing it through attention* (this is the satisfaction of a representational need), but rather of empathic or felt *engagement* (the satisfaction of an emotional need). Additionally, cross-modal integration of emotions, decision-making, and attention to social cues may be semantically integrated in a way that mere simulation can never capture (e.g., attending to a sardonic versus honest smile). Sarcasm, for example, is notoriously difficult to be detected by AI systems (Joshi et al., 2016). But again, even if AI *succeeded*

in such complex detection tasks, there would still be a gigantic gap between *detecting* and *understanding* an emotional experience.

Emotions and feelings have a deeply social dimension. The strong tendency to reduce these features to mechanistic algorithmic routines may work for a vast number of attention routines, but not for feelings and moral agency. This conclusion is not based on some humanistic type of fervor, dogmatic adherence to the "hard problem of consciousness," or to intuitions about semantics and the "Chinese room" thought experiment. Rather, it is based on empirical evidence, considerations about evolution, and the sociobiological functions of emotions. It is also based on our contemporary understanding of computation and AI systems. Purpose and motivation are essential to many attention routines and to the extent that AI systems lack intrinsic motivations, they cannot be considered as agents. Even if only attention is considered, AI systems are quite limited. But CAD complicates the picture in a more fundamental and principled way: *even if* AI systems managed to have motivations, those would be motivations to satisfy representational needs, rather than genuinely emotional-moral needs. This strongly favors an approach according to which AI systems should not be placed in situations that demand moral competence and appreciation (Sharkey, 2020).[6]

4.6 Unemotional but Rational Machines?

This section critically assesses the use of emotion-recognition technology and further clarifies the kind of limitation that AI moral agent-development faces. It discusses the work of Sherry Turkle and Rosalind Picard.

The challenge of modeling emotions in AI based on the dissociation between consciousness and attention does not mean that AI systems will not be incredibly transformative and useful. On the contrary, artificial intelligence has already changed our world. In addition, focused attempts at incorporating principles from the human emotional system in product design and computing systems are already in place (Ahn and Picard, 2014a, 2014b). The commercial use of such technologies includes the tailoring of user experiences based on emotion recognition (Bradshaw, 2016; Weintrauboct, 2012). The Kismet robot, for example, is programmed to detect emotions in facial expressions and respond accordingly, which often gives the impression of interacting with a living being (Breazeal and Scassellati, 1999, 2002). But perhaps we are, by using these technologies, endangering the authenticity of moral behavior; that is,

we may be substituting authentic human relations with simulated ones, which could have negative implications for society (Turkle, 2007). Nevertheless, the implementation of human-like emotions into AI systems is an attempt to improve human-computer interactions that may elicit more willingness for humans to interact with AI systems. This can provide more contextually relevant responses by AI, forging the path toward a partial, albeit still morally risky, solution to the value-alignment problem.

Rosalind Picard's affective computing approach (Picard, 1997) assumes that artificial agents may be designed to pass the emotional Turing test. Technology has improved emotion recognition through physiological measurement (e.g., due to stress or frustration), and this information can be used to provide personalized feedback or adjust a machine's performance (Picard, 2002a; Picard et al., 2001). The development of this computational account of emotion suggests that emotions can be processed by AI reliably and thus can be reduced to algorithms to some degree (Picard, 2002b, 2007). This ability to understand human emotion has clear implications for product development and marketing (e.g., Ahn and Picard, 2014a), but also for making human-computer interaction more fluent. While Picard acknowledges that computers may not achieve the level of conscious awareness that humans have, she argues that computers can achieve a "minimal sense" of conscious awareness, including self-awareness.

In order to interpret this proposal in the context of morality, one must know what is meant by "conscious awareness." Picard proposes a very flexible and general account. However, when more rigorous definitions are provided, the scientific prospects of emotional AI are bleak or at least not good, as CAD shows. The proponents of AI have been overoptimistic about the prospects of artificial consciousness. The main contribution of this chapter is that the most serious difficulty concerns the impossibility of morally autonomous AGI, or IEI moral AI. Since this difficulty concerns the normative force of morally relevant emotions, this is an in-principle or non-technical problem (what philosophers call an a priori limitation, not to be solved by advances in technology).

AI rational simulacra can be based on attentional-functional routines, but simulacra cannot reproduce the grip, motivation, and immediate urgency of phenomenal consciousness that contains emotional content. This is why even though AI may become rational and if attentive, *IEI epistemic agents with autonomy*, they cannot become IEI moral agents with autonomy. Although some computer programs can learn a form of normative morality by reading text and may even begin to behave in ethically salient ways based on the content of the text (Riedl and Harrison, 2015), they cannot truly empathize and understand

social interactions. This lack of empathy becomes especially important when assessing the utility of AI programs that become a more integrated part of society (e.g., see Miner et al., 2016). At best, feelings can only be simulated by AI, as seen in research that presents computational accounts of emotions. As Sherry Turkle argues, simulated thinking is (or potentially may become) thinking, but simulated feeling is not (and can never be) feeling (Turkle, 2005/1984). This distinction implies two different senses of cognition and rule-guidance: the "should" of rationality and the "should" of moral empathy.

5

The Vitality of Experience against Mechanical Indifference

5.1 The Value of Consciousness

This section examines the value of consciousness in terms of the autonomy and freedom it confers. The positive and negative notions of freedom are introduced. The work of Immanuel Kant on moral and aesthetic value is assessed. The vitality of the experience of freedom is defined as an unconditioned kind of curiosity.

A portrayal of human general intelligence would be incomplete without aesthetic and transcendental needs. The previous chapters have emphasized the importance of autonomy for intelligence and rationality. To sum up, epistemic and moral agency requires the autonomous satisfaction of needs based on the abilities of an agent in order to justify knowledge attribution and moral praise. An autonomous agent satisfies her needs because of her abilities. But CAD shows that the hierarchy of needs is divided into at least two broad sets of cognitive sources of value that may come apart and even compete against each other. Attention is the most important kind of mental agency in animal and human cognition—it is the paradigmatic form of mental action. When combined with the phenomenology of conscious experience, attention provides a kind of agency that transcends representational and truth-oriented value, not only in the moral realm, but also in aesthetics, spirituality, and the pursuance of autonomy for its own sake. Conscious attention provides guidance to autonomous agents with emotional, moral, and aesthetic needs based on the intrinsic value of phenomenal consciousness and subjective awareness.

Representational and rational needs are satisfied through inferentially structured attention. Their epistemic value derives from guaranteeing that beliefs are non-accidentally true as well as consistent with one another. A large amount of these needs, in fact the majority of them, may be satisfied in unemotional ways. This is explained by the fact that most of them concern accuracy conditions

that only the objective properties of propositional relations and the environment can provide. By contrast, attention to morally salient information is anchored in emotional needs, and this is why empathy is so important as a source of conscious attentional guidance. Moral needs, however, are not the only needs that cannot be satisfied simply by representing the environment in accordance to rational rules. Aesthetic and other transcendental needs (see Chapter 2 on Maslow's hierarchy of needs) depend upon conscious attention to vivid, valuable, and phenomenally integrated experiences. Conscious attention routines that underlie our moral and aesthetic capacities have various degrees of vital intensity that merely representational or purely rational-consistency routines lack. The full scope of these conscious attention routines is the focus of this chapter.

The previous chapter explained how CAD entails limitations for AI concerning empathy. Even if one grants that AI could become autonomously intelligent (or an epistemically autonomous IEI), it would not be able to appreciate, understand, or feel what motivates humans in their satisfaction of moral needs. This chapter expands on this fundamental limitation of AI by examining other aspects of human psychology that depend on phenomenal consciousness. Humans place these needs at the top of their hierarchy of needs—there is something *categorical or unconditional* about how valuable it is to have and satisfy these needs. An agent that has these categorical needs experiences her own autonomy and freedom in a fundamental way. This exacerbates the problem of value alignment and the limitations of unemotional machines.

Can freedom be experienced? Presumably, experiencing freedom affords a kind of value that transcends the value produced by mere norm-guidance. Autonomy, which is deeply associated with freedom, is unconditioned in the sense that it transcends sets of rules, representations, and regularities. Our lives are valuable to the extent that we can experience our freedom and exercise our autonomy. But what exactly does this mean? Freedom has a negative connotation, according to which agents are free if there are no restraints on their action. It also has a positive meaning, according to which one is free just in case one is in direct control of one's actions—one is free to the extent that one's life is determined by what one wants to do based on one's own capabilities. The previous chapters have focused on different aspects of these kinds of freedom and discussed related problems such as enfeeblement, the gorilla problem, and risks concerning value alignment. Freedom and autonomy are intrinsically valuable and pursuing them is intrinsically good precisely because the need for autonomy is categorical, or unconditioned, for any properly integrated agent. Immanuel Kant articulated these ideas with remarkable lucidity in his account

of aesthetic and moral value. As Paul Guyer (1996, 2) says: "Kant believed in the intrinsic and independent value of aesthetic experience but also in the uniquely unconditional value of morality, or the primacy of practical reason (that is, the use of reason to determine what we ought to do rather than what is the case)."

What we ought to do and our freedom go hand in hand, as paradoxical as that may sound. This is because what we ought to do, if we are *autonomous*, is the clearest manifestation of the unconditional character of our actions. When we do what we ought to do, because we ought to do it, we are not responding to heteronomous sources of decision-making: biological, social, or factual. We are acting on our own, because we are obliged to express our autonomy against merely factual factors. We are also acting based on our reasons and the satisfaction of our autonomous needs, which are also categorical—the need for autonomy is not negotiable for a free agent. The intricate details and difficulties of these notions need not be examined here. The essential point is that autonomous intelligent agents pursue their goals and satisfy their needs because of their abilities and, crucially, that they organize their needs in a hierarchy in which the most categorical needs that are essential to their autonomy are at the top. Thus, similarly to the hierarchy proposed by Maslow, aesthetic, spiritual, and moral needs will be at the top of the hierarchy of a complex and virtuously integrated agent (transcendence needs are at the very top for Maslow).

Biological needs are satisfied by all living creatures, from plants to mammals. Representational needs may be satisfied by plants, but they are clearly satisfied and in incredibly complicated ways by animals. Rational needs must be satisfied for consistent and optimal behavior and many animals show signs of rational decision-making (although clearly the highest forms of rationality are exhibited by humans). Emotional, moral, and aesthetic needs seem particularly central for human rationality, although again, animals have empathic and emotional needs. In animals and humans, biological needs are "at the bottom" not because they are not fundamental sources of value. To the contrary, they are an *essential* source of moral and aesthetic value. But biological needs are not, and cannot be, the *only* sources of value for an autonomous agent. At some point, the genetically determined metabolic commands of an agent and what she ought to do come apart.

Biological needs and their metabolic foundation provide the vitality of consciousness and, as such, they are crucial ingredients in the kind of moral understanding and aesthetic appreciation that make our lives worth living. They may even count as "inner springs" of creative impetus that operate largely unconsciously and erupt into vivid awareness in the virtuous agent. This kind

of "metabolic growth" view of the mind radically differs from the *tabula rasa* favored by empiricism, which describes the mind as dependent on external impressions that get organized according to some associative and mechanistic learning principle (although this is a very rough characterization of empiricism). In particular, some views of the mind emphasized the crucial role that biological needs play in virtue of their interconnectedness with the creatively complex network of activities that constitute the living world beyond the perceiving agent. M. H. Abrams (1953) describes this interconnectedness of mental vital energy as follows:

> The dominant English psychology of empiricism had no place either for the concept of growth or of the subliminal in the activities of the mind. The psychology of Leibniz, on the other hand, so influential in Germany in the latter eighteenth century, was favorable to both these concepts. Leibniz emphasized the essential community of all monads, from the human soul, down through the vegetable kinds, to the monads of apparently inorganic substances. The real, as opposed to phenomenal nature, is living and organic throughout this hierarchy, and each monad, of every degree, is 'a perpetual living mirror of the universe,' possessing within itself the simultaneous perception of everything, everywhere, whether past, present, or future. Man is distinguished from the lower orders in the scale of being because in his soul some few of these perceptions arrive at a sufficient degree of clarity to achieve 'apperception,' or awareness. Still, even in the soul of man, the mass of *petites perceptions* which remain below awareness incalculably exceeds the tiny area which becomes available to consciousness.
>
> (Abrams, 1953, 202)

The *organic power* of the biological mind is indeed deeply intertwined with an ancestral trajectory of evolutionary innovation, which is majestic in its complexity. There obviously has to be a difference in value between artificial, unemotional machines and us, animals with a long history of development in our living planet. The value resides in the vitality of conscious awareness, as this chapter seeks to articulate, but the undisputed basis of this vitality is the energy of metabolic life itself. Value, however, is normative—it compels the agent to act in accordance with what she ought to do, because *she* decides to do so, and not merely because she is a causal cog in a cosmological machine. Our genetic lineage is an open window into the remote past and future of our species. But our genetic endowment is part of the mechanical and biochemical machinery of life. How to understand vitality as autonomy then? The *petites perceptions* are a good place to start. They are the foundation of agency. These *petites perceptions*

can be conceived as unconscious types of attention routines conducting much of the orchestration of the mind behind the theatre of conscious awareness, which is limited in access. Still, the autonomy and freedom of an agent are at their most salient in her experiences of independence. Biology is a major *source* of agency but it cannot be the *only* source.

The experience of freedom in conscious awareness, sensed by the agent as the satisfaction of her categorical needs, is fundamental to determine the worth of her life. What does it really mean to say that freedom is *experienced* through the satisfaction of categorical needs? As Maslow noted, while most needs require less effort and motivation as they become satisfied, transcendental needs increase motivation as they become more and more "satisfied." This asymmetry and apparent contradiction can be explained in terms of an *unconditioned curiosity for the transcendental*. Kant, as mentioned, shed invaluable light into this intricate issue. The essential Kantian doctrine required for the present discussion is briefly encapsulated by Guyer in the following passage:

> There is an intimate and indispensable connection between the analysis of aesthetic judgments and the explanation of aesthetic response, which is the core of Kant's theory of pure judgments of taste, and the linkage of aesthetics to morality, which is clearly Kant's ulterior motive. The pleasurable yet disinterested sense of freedom from cognitive or practical constraint—that is, the sense of the unity of aesthetic experience without its subordination to any scientific or moral concepts and purposes—which is at the heart of Kant's explanation of our pleasure in beauty is precisely that which allows aesthetic experience to take on deeper moral significance as an experience of freedom.
>
> (Guyer, 1996, 3)

"Disinterest" is the key term here. The experience of genuine freedom is disinterested and it is categorical—it is valuable in itself, regardless of the consequences of or conditions for actions expected to produce utility, calculated outcomes, or other conditional expectations. Aesthetic experiences are transformative, fundamental, and transcendental. They are transcendental because they make the autonomous agent go beyond the standard "objective" and representational, or truth-oriented rational needs, of utility and expectation. In this sense, satisfying these needs allows the agent to transcend her epistemic agency or the satisfaction of needs concerning intelligence and problem-solving, which would only "objectify" her freedom. The deepest attentive allegiances of an autonomous agent are with what makes her the freest. This is why transcendental needs are at the top of a human's priorities.

5.2 Transcendental Needs and Categorical Desires

This section elucidates the nature of transcendental needs and their role in shaping an individual's hierarchy of needs into a narrative structure. It addresses the role of memory in specifying utility and value. Daniel Kahneman's distinction between the experiencing and remembering self is introduced. Specific types of memory-related agential risks are discussed. The notion of "familiarity" is defined.

Transcendental needs are essentially categorical. They make us who we are, beyond a merely intelligent existence as a utility maximizer. The fact that aesthetic experiences take on a "deeper moral significance as an experience of freedom" does not mean that morality is relative or contingent on subjective desire. To the contrary, the disinterested experience of beauty and goodness satisfies a categorical and essential need of free agents. The experience associated with helping someone without expecting utility maximization has personal value as empathic engagement, and it also has *real* moral value, as opposed to subjective, or contingent value. Imagine how personally irrelevant would the most important moral decisions be if nothing about them engaged our subjective awareness and determined who we are. No moral decision would be personally transformative or meaningful; it would all come down to rules and calculations. Consider how Finn's decision to help Jim had real moral worth despite Finn's accurate belief that it was wrong to help Jim according to the laws and customs of his contingent times. By being empathically engaged and morally good Finn transcended the contingent wrongs of his society. Transcendental needs are constitutive of our consciously aware identity and the "narrative self." If morality were all relative and contingent on the rules and standards of utility optimization, then we would become utterly *depersonalized* and lose the uniqueness of our selves. This is why transcendental needs are categorical—without them we cannot be genuinely free and autonomous moral agents (free from the contingent standards of society and utility).

The disinterest involved in a morally good life resembles, or is deeply connected to, the radical disinterest of aesthetic experiences. These experiences involve strong visceral reactions that cannot be reduced to strictly representational aspects of their sources, or to calculated utility goals, such as the amount of "screen-time" or the expected income that might result from having them. This is also what makes profound aesthetic experiences *transformative*. When one contemplates something beautiful, it is hard to keep track of time (against the "time is money" imperative). If one is empathically engrossed with beauty even

a lot of time seems like nothing—one could be looking at a mountain range, or a painting without experiencing effort or expecting monetary compensation. This is exactly what highly skilled performers describe as the experience of "flow" mentioned in the previous chapter: an effortless and phenomenally conscious type of attention.

Memory provides one of the clearest examples of the difference between epistemic and moral need-satisfaction. The episodic memory system satisfies epistemic needs and is valuable because it is a source of justification for beliefs about the past in relation to our plans for the future—it is an essential component of an autonomous epistemic agent. Autobiographical memory satisfies moral and narrative-autonoetic needs or needs that are about knowledge and awareness of ourselves, and is valuable because it is a source of personally meaningful and insightful experiences about our past—it is an essential component of an autonomous moral agent. Unlike autobiographical memory, episodic memory is only weakly autonoetic. The relation between these two roles of memory is captured by the tension that exists between a narrative and an accurate report. Episodic memory capacities provide weakly autonoetic memories that are not luckily accurate, thereby guaranteeing an external kind of justification for beliefs about memories. This kind of memory may be implicit or explicit (unconscious or "access" conscious). Thus, weakly autonoetic or "report-like" memory is structured in a rational or truth-seeking way. Autobiographical memory, however, seems to be necessarily based on phenomenal consciousness and its categorical value (Montemayor, 2018).

Our lives need to be lived according to accurate information about the past, organized in a way that makes some plans more rational and salient than others. But our lives cannot be simply understood as a *report* of activities and a set of plans. A full report up until now about my memories, updated according to my current plans, should be familiar enough to me—it should be weakly autonoetic. But the moral and aesthetic value of autobiographical memory provides an internal kind of justification about what I really value as a person, beyond the facts as stated in the report about my life, such that my narrative is not artificially or luckily related to what *I value* as an individual: a phenomenally conscious kind of memory, which is strongly autonoetic (this notion is based on the principle of *Narrative Integrity* discussed in Montemayor, 2018). A report cannot capture which memories I value the most.

If an agent lacks *familiarity* with her past and values, substantial risks for autonomy emerge—is this event important in my life, why do I remember this so vividly? The epistemic and moral aspects of memory present a trade-off

between accuracy and narrative integrity, or between truth and personal value. This is a particularly interesting implication of CAD for memory capacities, with fascinating consequences for the value alignment problem and AI design in general. For instance, Stuart Russell addresses these difficulties by appealing to Daniel Kahneman's research on the *experiencing* and the *remembering* self. The experiencing self "calculates" utility in terms of the actual succession of experiences and their cumulative hedonic value, while the remembering self bases this calculation on a value-bias that emphasizes maximally vivid memories. Kahneman's experiments show that the remembering self is typically in charge and that she calculates utility in ways that are irrational, if one looks *exclusively* at what the hedonic utility should be, based on objective values. Russell explains these findings as follows:

> Kahneman's explanation is that the remembering self looks back with rather weirdly tinted spectacles, paying attention mainly to the "peak" value (the highest or lowest hedonic value) and the "end" value (the hedonic value at the end of the experience). The durations of different parts of the experience are mostly neglected. The peak discomfort levels for 60 and 60 + 30 are the same, but the end levels are different: in the 60 + 30 case, the water is one degree warmer. If the remembering self evaluates experiences by the peak and end values, rather than by summing up hedonic values over time, then 60 + 30 is better, and this is what is found. The peak-end model seems to explain many other equally weird findings in the literature on preferences.
>
> (Russell, 2019, 239)

Why are these findings weird? This style of decision-making is weird in two critical ways. First, decisions based on the peak-end model are irrational because choices should be made on the basis of the sum of values over instants of time, rather than the "peak" experienced by a subject. Consider that 60 + 30 is objectively worse than 60. It is the same badness of 60, plus more badness. Why are people doing this to themselves? Violations to other principles of rationality are extensively documented by Daniel Kahneman's research on systems 1 and 2 (in collaboration with Amos Tversky), including systematic violations to the axioms of probability. System 1 is heuristic, implicit, unconscious, designed for quick actions and decisions, and epistemically inaccurate in various surprisingly simple cases. System 2 is slow, reflective, conscious, and epistemically accurate (for a philosophical response to the "situationism" and unreliability these findings suggest for epistemology, see Fairweather and Montemayor, 2017).

Second, such a divided agent whose decisions are tainted by attending to peaks of experience violates, as Russell points out, John Harsanyi's *principle of preference autonomy*, which is essential for rational choice: "In deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences" (Russell, 2019, 220). The previous chapters argued that genuine intelligence requires autonomy, but here we have a fundamental violation to preference-autonomy: subjects are deciding against their best interests because one way in which they rank their preferences (in terms of experience) prevails over the *rational* way of ranking their preferences (in terms of utility maximization). So the problem is twofold. It is not only the case that the decisions are irrational, but also that the decider is divided in her preferences, and she is systematically biased toward irrationality based on heightened experience or heuristic salience.

CAD explains why human agency is divided in exactly this way, into two kinds of autonomy: utility-based and experiential, each with its own style of aligning value. The tainted spectacles through which the "remembering self" pays attention are the spectacles of phenomenal consciousness. This divided agency is pervasive in human psychology.[1] Of course, this is not *divorced* agency and in general, it is virtuously integrated agency, but because of the nature and internal variations of the hierarchy of needs, aspects of these different kinds of agency, moral and epistemic, can enter into conflict. Crucially, experienced value plays a personal and motivational role that mere sums of past choices or objective measurements through time cannot capture—it plays the role of making decisions *personally relevant and familiar*. It is partly because of this that our preferences change, sometimes radically, making our lives meaningful, interesting, and exciting. Russell writes,

> The fact is that no law *requires* our preferences between experiences to be defined by the sum of hedonic values over instants of time. It is true that standard mathematical models focus on maximizing a sum of rewards, but the original motivation for this was mathematical convenience. Justifications came later in the form of technical assumptions under which it is rational to decide based on adding up rewards, but those technical assumptions need not hold in reality. […] Kahneman acknowledges that the situation is complicated still further by the crucial role of anticipation and memory in well-being. The memory of a single, delightful experience—one's wedding day, the birth of a child, an afternoon spent picking blackberries and making jam—can carry one through years of drudgery and disappointment. Perhaps the remembering self is evaluating not

> just the experience per se but its total effect on life's future value through its effect on future memories. And presumably, it's the remembering self and not the experiencing self that is the best judge of what will be remembered.
>
> (Russell, 2019, 239–240)

Indeed, why would the value and meaning of your most cherished experiences, of *your life*, depend on sums of past hedonic measurements (and what is a hedonic "measurement" really measuring such that it can be "chopped" into neat units of time that can then be mapped to equations?). The issue is not only that you can change and renew the value of what you experience, but also that there is no connection between measurements and the precise manner in which you experience such value. In particular, the remembering self pays special attention to the *total effect* of an experience on life's future value. This is exactly the function of narrative integrity that phenomenally conscious autobiographical memory provides (Montemayor, 2018). Total-effect value, which informs the hierarchy of needs, should not be incompatible with minimum standards of rationality and yet, that is what the empirical findings suggest.

Kahneman, as Russell points out, struggles with this problem. On the one hand, Kahneman says that the remembering self made a mistake here (and in principle, always makes this kind of mistake, which is frankly a bit concerning). On the other hand, he says that a theory of well-being that "ignores what people want cannot be sustained" (Russell, 2019, 239). Perhaps this is a false dichotomy. Is the remembering self really making a *mistake*? The answer CAD presents is that these two styles of decision-making *satisfy different needs*. Neither of them is mistaken once the needs they satisfy come to light. The lesson is that the needs of the autobiographical remembering-self *transcend* the epistemic needs captured by utility maximization. These are two kinds of *autonomy needs*: the need to accurately represent the environment in a way that is organized by our own representational skills, based on our preferences understood as utility functions, and the need to integrate our memories into a narrative that is familiarly organized in terms of what we value the most. These needs correspond to the weak and strong kinds of autonoesis that ground two types of autonomy: epistemic and moral/aesthetic.

What is this "familiarity" that consciousness affords? As Robert Sapolsky (2016) explains, when Joseph Capgras discovered the syndrome now named after him, he documented how the feeling of familiarity can become dissociated from the perceptual capacity of recognition in a completely unexpected way. What confused Capgras the most was that his patient had

perfect epistemic access to the defining aspects of the identity of individuals that were either close relatives or the closest person in her life. These included detailed physical characteristics that served as specific criteria for recognition. The patient, however, was incapable of recognizing them as someone she *loved and trust*—she fully "identified" the individual in great detail, but the feeling of familiarity was wholly removed and, on the contrary, a strong feeling of strangeness and lack of trust ensued. In other conditions, this extends to inner organs (Cotard's syndrome) or objects and houses (reduplicative paramnesia), mostly the place one lives in. Studies revealed that the neurological basis for perceptual skills concerning detailed object recognition can be dissociated from the neural circuitry underlying trust and familiarity. Both capacities, for recognition and trust, are fundamental for our social lives and for the value and meaning we find in social relations, but they depend on a delicate balance that can fall apart.

In Capgras syndrome, one of the capacities that become dissociated from its normal social function is entirely epistemic—object recognition that grounds perceptual justification. This capacity remains intact, which is what surprised Capgras. The other capacity is very hard to pin down. It could be called empathy, care, trust, and love. "Familiarity" is a good term to capture all of these emotions and reactions. In Capgras syndrome, perfectly correct epistemic function, when fully dissociated from familiarity, becomes disabling—one recognizes one's spouse, but believes this individual is an identical imposter. According to CAD, this generalizes. Recall the distinction between recognitional and emotional color. One capacity is to correctly identify and label colors, the other one is to experience emotional reactions to them. Both capacities are dissociable, and they are instantiated in different neural circuits.

These findings have implications for debates about the nature and value of consciousness. Frank Jackson's (1982) "Mary" thought experiment shows that she had all the epistemic (scientific) skills she needed to *identify* colors and even identify what colors other subjects were seeing, even though she had never *experienced* these colors herself. When she gains the ability to experience red for the first time, what is exactly the nature of this ability? She could pay attention to red color *representations* before she had this experience—this is why she possessed all the physical and recognitional knowledge of color. So when she gains the ability to experience color, viscerally and from her subjective perspective, she must gain a different ability. The lesson that many philosophers drew from Mary's situation is that her new knowledge is "not physical." But a better lesson to draw is that her new knowledge is based on a new ability

to *empathize* through color (e.g., imagine vivid aesthetic experiences, and "connect" color with pleasure or repulsion). Previous proponents of similar thought experiments made exactly this point—see Feigl (1958).

The value of color experiences, which is what Mary learns, is emotional, rather than simply representational, truth-tracking, or epistemic. Mary has now an open door to the aesthetic and empathic dimensions of color. This is the value of consciousness—the empathic, visceral connection to biological, emotional, and transcendental needs, rooted in a deep and robust sense of familiarity. Mary can now be "disinterested" about color and, instead of recognizing and reporting, she can focus on *her own* color emotions. Consciousness satisfies moral, aesthetic, and transcendental needs (see Humphrey, 2011). This is an important consequence of CAD that stands in opposition to the entrenched doctrine in philosophy of mind that the value of phenomenal consciousness is rational and epistemic (Smithies, 2019). For AI research, the lesson from this analysis is that ethical AIs can *at best* be EEI moral agents—this impossibility is explained by the fact that our emotional needs are not mere representations and that they depend deeply on our biological needs.

5.3 The Strength of the Absurd

This section addresses spiritual and transcendental needs. It discusses proposals by Robert Sapolsky, Wallach and Allen, and Ludwig Wittgenstein. Frank Jackson's "Mary" thought experiment is examined in the light of transcendental needs.

According to some authors, a truth norm is deeply inadequate to define the most important beliefs a human being can have, namely, those that define her spiritual convictions and shape her outlook on life *as a whole*. One can even argue that moral autonomy is never just a matter of utility or rational preference. Søren Kierkegaard's analysis of the Biblical passage concerning Abraham's decision to kill his son Isaac is based on this principle. According to Kierkegaard, one believes, in the case of religion, on the *strength of the absurd*. Conflicting evidence, irrationality, a systemic discrepancy between belief and fact, incoherent utility maximization, or any other notion of epistemic irresponsibility, should not be an obstacle for religious—or deeply personal—belief. This is surprising, given that religious belief was pivotal in how humans transcended all other species in terms of large-scale cooperation and long-term planning (Harari, 2015). A *leap of faith* was really needed for us to evolve into who we are now.

The spiritual person finds strength in the apparent absurdity of religious belief, but not out of stupidity or epistemic incompetence. Rather, religious belief *demands this indifference toward objective standards of truth*—religion requires the kind of disinterest characteristic of moral and aesthetic experiences. The religious person believes with the highest commitment, a conviction that demands the strongest type of faith, against all types of epistemic reasons and evidence to the contrary. This is *personal commitment*, of the most unswerving kind. It is genuine belief, as opposed to simple hope or wishful thinking—this is what Abraham illustrates, according to Kierkegaard, in the biblical story in which he is asked by God to kill his son, Isaac (a key component of his objection against Hegel's dialectical and *rational* approach to the philosophy of religion). Why should religious convictions be genuine beliefs? Because they involve the strongest type of commitment, and crucially, because of the impact they have on our life as a whole. In an interview with *Edge* with the title *A Bozo of a Baboon*, Sapolsky explains,

> The minute you're in the realm of Sister Helen Prejean, the nun featured in the movie *Dead Man Walking*, you have left the primates far behind. How can someone spend all their time ministering to the most deplorable, scum-of-the-earth people? Prejean says that what has to be the case is that the less lovable they are the more you have to love them. The less likelihood of reward, the more you have to be willing to do the right thing and get punished. This is the realm where Kierkegaard said that Christians need to be able to contain two contradictory facts in their head simultaneously, where the more explicitly faith is challenged, the more irrefutably it is negated, the more there must be faith. Nothing in primatology or in your dopamine reward pathways can explain that. This is off the edge of the cliff into a completely different realm. Incredibly few people live lives where they get no reward. This behavior is certainly maladaptive, since by definition you're not going to be passing on copies of your genes, and neither is your kin line. You can't come up with any sort of adaptive argument that involves doing the incredibly self-sacrificial right thing, and getting punished for it.
>
> (Sapolsky, 2003)

This kind of ministering is *biologically maladaptive*, and to that extent, it is irrational behavior. And yet, it is *exactly* the kind of behavior that is associated with the transcendental needs that Maslow (1987) places at the very top of his hierarchy. Harari (2015) would agree with Sapolsky that this is an instance of need-satisfaction where the primates are left far behind. But Harari would disagree with Sapolsky that such irrational behavior is maladaptive—in fact, it is this kind of behavior that made us the most dominant species in this planet.

Faith cannot satisfy biological needs, but it certainly satisfies more abstract and transcendental needs. If you are a baboon, then you are a fool by behaving in this way. But faith is transcendentally "adaptive," beneficial, and *transformative*. It transformed us, for instance, into the most cooperative and mighty force of the known world of intelligent creatures through a strong and transcendental type of trust. This dominance is threatened by AI. But could AI be dominant without the satisfaction of this kind of need?

It was *unconditional or categorical trust*, or faith (in God, in scientific evidence, in the market, in democratic societal organization, in civil and political rights, in legal systems) that propelled us into dominance. Trust in what exactly? The "content" of trust did not matter much. Rather, what mattered was what such a deep kind of representationally disinterested trust allowed for—*unconditional or categorical cooperation*. This is trust in our ability to empathize and feel emotions in a similarly disinterested way. We trust that the world is familiar to us, not merely in representational terms, but fundamentally in emotional and experiential terms. A merely representational world would feel unfamiliar to us; we need strong emotional experiences to make the world deeply real and familiar to us—to make this world *our own*.

But isn't here a problem concerning the *real* value of categorical needs? Cooperation is great, but don't we also need to account for the reality of value? Deontological approaches to moral value seek to identify an objective basis for morality on the fundamentality of rights, upon which good norms are justified. Utility maximization or welfare views seek to establish an objective measure of goodness in the kind of hedonistic values and preferences that Kahneman showed to be irrational. The approach defended in the last two chapters is to appeal to a capability approach that grounds human rights in order to satisfy the real-value requirement, and to capture some of the key insights of the alternative approaches (see Gabriel, 2020). For now, the emphasis is on showing how having a good understanding of human rationality is essential for AI design. In their book on artificial moral agents (AMAs), Wallach and Allen (2009, 87) comment on James Gips' requirements for a consequentialist robot in terms of four abilities: (i) a way of describing the situation in the world; (ii) a way of generating possible actions; (iii) a means of predicting by conditional or counterfactual reasoning the situation that would result if an action were taken given the current situation; (iv) a method of evaluating a situation in terms of its goodness or desirability. None of these abilities are robust or trustworthy in current AI and the last ability is impossible without phenomenally based empathic abilities. In the epilogue to their book, Wallach and Allen write,

> We have learned that the process of designing (ro)bots capable of distinguishing right from wrong reveals as much about human ethical decision making as about AI. We started with the deliberately naïve idea that ethical theories might be turned into decision procedures, even algorithms. But we found that top-down ethical theorizing is computationally unworkable for real-time decision. Furthermore, the prospect of reducing ethics to a logically consistent principle or set of laws is suspect, given the complex intuitions people have about right and wrong. […] People don't want AMAs to replicate the abstractions of moral philosophers any more than they want their neighbors to do so. People want their neighbors to have the capacity to respond *flexibly* and *sensitively* in real and virtual environments. They want to have *confidence* that their neighbors' behavior will satisfy appropriate norms, and that they can *trust* their neighbors' actions. Meeting this challenge will entail an even more thorough understanding of human ethical behavior than is presently available.
>
> (Wallach and Allen, 2009, 215–216, my emphasis)

A more thorough understanding of human morality is indeed needed, and a central component of it is appreciating the role of conscious attention in sensitively satisfying emotional and transcendental needs associated with autonomy and care. An epistemically omniscient AI with all the required information for applying consequentialism to concrete cases will still need attentional abilities to determine salient information in the context of care, rather than representational accuracy. But even ignoring this problem, the representation of cases would require vast amounts of knowledge about the causal structure of the world (Marcus and Davis, 2019). Deontological approaches fare no better than consequentialist views because similar abilities and representational capacities are needed. Care for each other is based on moral autonomy. AI operates heteronomously (for Russell, by design because beneficial AIs never satisfy *their* needs but *ours*). No child machine will care the way we do, and we would not trust AI that only accidentally understands our moral perspective. Even a "slave" AI would be dangerously out of touch, and in principle, it would be incapable of even understanding our transcendental needs because AI can at best merely represent them. Thus, AI cannot be a moral IEI.

Although transcendental needs are rooted in emotional and biological needs, they are extremely difficult to model or explain in behavioral terms because they are not reducible to fixed sets of action-patterns. They also cannot be reduced to emotional and biological needs as the motivational asymmetry described above shows, which justifies why they are at the top of the hierarchy of needs (the closer we get to satisfying them the more motivated we become, without any clear

satiation point). This is why they transcend representation and basic rationality, and it is also why they really set us apart from other species. Understanding these needs by representing the behavior of individuals is not going to work. Satisfying these needs, on the contrary, requires a very unique type of conscious attention that is highly integrative and highly selective. When this kind of attention is properly displayed, it constitutes a virtue (Aristotelian and Confucian ethics emphasize this type of excellence based on habit or inner dispositions). Moral care, flexibility, confidence, sensitivity, and trust, all highlighted in the quote above, require emotional abilities and selective attention, rather than principles, norms, mathematical modeling, and collections of patterns and representations. Once transcendental needs are in place, satisfying moral needs becomes integrated with other needs in a way that provides a categorical kind of motivation.

Ludwig Wittgenstein wrote about the transcendence of trust, care, sensitivity, and the autonomy of deliberation in various texts. Even in his early and highly representational thinking, expressed in the *Tractatus Logico-Philosophicus*, Wittgenstein remarked: "It is clear that ethics cannot be put into words. Ethics is transcendental. (Ethics and aesthetics are one and the same)" (Wittgenstein, 1922/1974, 86). In his *Lectures on Aesthetics*, Wittgenstein illustrates how from fashion and style, to music appreciation and architecture, aesthetic sensitivities are *never* the result of causal explanations, scientific theories, or representational schemes. Aesthetic appreciation requires a kind of aspect "dawning" or aspect perception that grounds similarities and differences that transcend mere explanation and theorizing—they require focused attention, habituation, and skill.

In music appreciation, for example, what matters is the attentive sensitivity of our reactions, emotions, gestures, and expressions. According to Wittgenstein, if we gave a theory of music appreciation that is entirely based on the features of the environment coded through audition, and neutrally expressed in different brain areas, which indeed "cause" our experience of musical beauty we would be *entirely* missing the point. This "theory" of aesthetic experience, paraphrasing Wolfgang Pauli, would not even be wrong. It is so off the mark that it is not even relevant to the aesthetic domain. For Wittgenstein, even the *words* we use in our aesthetic judgments and the explanations and theories we concoct to make sense of our aesthetic attentional practices play a completely secondary role and elucidate very little about how or why we appreciate music. Wittgenstein writes,

> In order to get clear about aesthetic words you have to describe ways of living. We think we have to talk about aesthetic judgements like 'This is beautiful', but we find that if we have to talk about aesthetic judgments we don't find these words

> at all, but a word used something like a gesture, accompanying a complicated activity. [The judgment is a gesture accompanying a vast structure of actions not expressed by one judgment.]
>
> (Wittgenstein, 1967, 11)

An "aesthetic" AI, based on the best theories, data-based representations, and scientific explanations, would be a complete travesty that does not even get aesthetic experiences wrong: *there is no "representationally-right or wrong" here*. Extending this analysis to ethics, IEI autonomous moral and aesthetic agents are therefore impossible, a priori and by design. In fact, if Wittgenstein is right, scientific sophistication and advancement in the realm of ethical and aesthetic AI is nothing but sophistry and delusion. And given the importance of risk reduction in AI development and the difficulties associated with AI value alignment, such sophistry is not only erroneous but also *dangerous*. Precisely because a lack of understanding here entails various types of risks and dangers, our joint ethical and moral sensitivities entail a kind of care and unconditioned interest in other fellow human beings, animals, and the world itself that is impossible for AI to represent. During his lectures, to the challenge presented by one of his students "If my landlady says a picture is lovely and I say it is hideous, we don't contradict one another" Wittgenstein responds,

> In a sense you do contradict one another. She dusts it carefully, looks at it often, etc. You want to throw it in the fire. This is just the stupid kind of example which is given in philosophy, as if things like "This is hideous," "This is lovely" were the only kinds of things ever said. But it is only one thing amongst a vast realm of other things—one special case. Suppose the landlady says: "This is hideous," and you say: "This is lovely"—all right, that's that.

"That's that" meaning that just a few English words are not at all enough information to determine care, empathy, or appreciation. This is obviously not a logical contradiction (how boring would life be if it were merely a contradiction-avoidance strategy), but a contradiction in value—you may even say, a contradiction in *life style*. Isn't this kind of contradiction more significant for a human being than logical or pragmatic contradiction? Isn't care a more basic need according to which all other needs are organized, and isn't a life based on such transcendental care-needs more meaningful and beautiful than a life spent satisfying market value utilities or other representational needs? Preferring to look at a painting exclusively in terms of market value is obviously a *kind* of utility. But isn't it more meaningful to experience the disinterested care of really enjoying the painting independently of price-tag?

Satisfying transcendental needs requires attending to overall significance and conscious appreciation, rather than merely "being right" according to some evidential standard. For you to appreciate why Gustav Mahler is a great composer, does it suffice if you just yell "Wonderful!" or "Magnificent!" every time you recognize a song by him? Not at all, and you would be sadly confused if you thought so. Or would it help to give a detailed account of sound acoustics, instrument production, neural circuitry, and so on, of every single symphony by Mahler and his listeners? Again, not at all; this is entirely irrelevant for musical appreciation, even if such material factors constitute the "causal basis" of appreciation. The same holds for the market value of the song, the year it was composed, etc.

To give a more familiar example from contemporary philosophy of mind, let's return to the case of Jackson's Mary. The key to understand what kind of new information she learned is the type of *appreciation* that Wittgenstein investigated in his Lectures. It will clearly be very difficult to understand appreciation in terms of the kind of physical-fact knowledge she possessed before her first experience of red. What is this appreciation? I hope it is clear by now that if the response is the one provided by metaphysicians of mind we are left in the dark—if all she learns is a non-physical "fact," or non-physical information, or a new representational theory, we have not elucidated at all what is it that she learned (see Lewis, 1988). As mentioned above, Mary's transition from recognitional red to emotional red involves a transition from epistemic to empathically moral or aesthetic value.

Before red-color revelation, Mary had a *theoretical* understanding of color. She, according to the thought experiment, possessed "omniscient" knowledge of color-vision concerning all the relevant scientific facts, how they are causally structured, etc. After revelation she gains a subjective and visceral experience and a new responsiveness to color, and with it, she acquires an appreciation of how emotive color can be. Being a world expert in color vision, she has read multiple reports about how people react strongly to some colors, with comfort or stress, calling them "beautiful" or "ugly." But now she can finally appreciate, understand, and empathize deeply with them. Mary's new "knowledge" gives her aesthetic and emotional appreciation. She disinterestedly spends now time at galleries, seeing sunsets, or painting her house. This brings not only a deeper kind of non-theoretical familiarity to her visual experiences, but it also brings a range of *sensibilities* that open up a whole new world of attitudes and reactions. Her needs have changed; *she* has changed.

The appreciative self is not a theoretical self. Richard Moran (2001) explains how the knowledge we have of our minds is familiar and authoritative because

it provides a *stance* that is uniquely responsive to needs and interests. Third-person representations, reasons, or theories cannot capture this stance. Based on a similar distinction by Jean-Paul Sartre between "self as facticity" and "self as transcendence," Moran differentiates between the theoretical stance one uses to explain one's own actions in terms of antecedent conditions, and the deliberative stance one uses in order to respond to questions about what one should do or is going to do at any point in time. The second kind of questions requires a sense of familiarity and commitment that external reasons, reports, or descriptions cannot provide. Commitment to a decision is not merely based on an *explanation*, and this is why a life lived in terms of external reasons alone is radically unfamiliar or estranged. We do not merely predict what we are going to do, we actually decide what we should do. When we "justify" our actions by means of accurate conditions and predictions, we are acting in *bad faith* because we are avoiding the stance required for genuine personal responsiveness. An alignment of value with what "matches" our external behavior will produce a set of values that are ultimately alien to the perspective we are familiar with. But as we are about to see, the deliberative, free, and transcendent self can have a darker side when she becomes too obsessed with her own transcendence.

5.4 Sadistic or Luciferian Needs: Reward-Related Challenges for AI

This section describes wireheading and its negative consequences concerning diminished autonomy. It discusses the work of Iris Murdoch and Stuart Russell's notion of negative altruism.

The depths of transcendental interests and desires are not always good news. In our contemporary self-oriented, ultracompetitive, money-driven society, an intense desire to transcend orients the self *addictively inwards* at the cost of the natural empathy humans and animals feel toward each other, for instance, in childhood. This inward reorientation and retraining of our empathic attention produce a unique type of anxiety and a new set of highly salient needs that demarcate what we now consider normal rational adult behavior. These needs are selfish because their satisfaction involves comparisons for reaching beyond established societal standards in other to gain notoriety. The entire "meritocratic" structure of the current liberal markets and political systems encourages and fuels the intensity of these needs for notoriety. Under the spell of merit and social

status, the hierarchy of needs is reframed inwards. Narcissistic attention-seeking needs, if mild, are absolutely normal expressions of the need for *recognition*. But when the vitality of experience that depends on empathic and emotional needs is entirely devoted to narcissistic needs, then, ironically, the autonomous self becomes mechanistic, predictable, and automatized by addiction to the self.

The key to understand this irony and its repercussions for AI design is, as Russell (2019) explains, the dopamine system. Addictive behavior, including risk-seeking, novelty-seeking, and the need to go beyond and surpass others at any cost, engages the dopamine reward system and makes us mechanically, neuro-chemically, addicted. Wireheading is the phenomenon of engaging directly the brain reward system by electric stimulation. The results are horrific—rats press a lever to produce stimulation without pausing to eat or drink until collapsing, and so do humans. Attention-addiction to social media, without being as horrific or dramatic, exploits the same system. Could reinforcement-learning AI "wirehead" given that they are designed according to a reward maximization algorithm, which they could manipulate by convincing human designers to reprogram in order to increase reward? The answer is yes, but the problem is intricate. In the context of AI's interactions with humans, wireheading presents a unique kind of risk:

> The AI safety community has discussed wireheading as a possibility for several years. The concern is not just that a reinforcement learning system such as AlphaGo might learn to cheat instead of mastering its intended task. The real issue arises when humans are the source of the reward signal. If we propose that an AI system can be trained to behave well through reinforcement learning, with humans giving feedback signals that define the direction of improvement, the inevitable result is that the AI system works out how to control the humans and forces them to give maximal positive rewards all the times.
>
> (Russell, 2019, 207–8)

Russell's solution to this problem is to distinguish reward *signals* from *actual* rewards. There is information signaling what a human is actually experiencing—the actual reward. But the AI now has no incentive to wirehead because accumulating reward signals is not going to entail accumulating actual rewards. However, even if there is no direct mapping between signals and human rewards, the learning process is still driven by rewards of the kind that in humans, rats, and other animals produce extremely addictive behavior. Any kind of AI manipulation entails the objectification of humans, and if this involves reinforcing addictive behaviors through rewards, AI "assistance" will bring the

worst of human psychology. Regardless of whether Russell's proposal effectively neutralizes wireheading, the structure of the problem is quite alarming. Even if the addictive behavior that is enhanced and optimized by AI is much milder than wireheading, the fact that it appeals to the selfish and immediate reward-utilities of a human being should give us pause. Satisfying these addictive, selfish, and "moment-to-moment" needs will preclude the virtuous integration of needs. An egomaniacal individual would surrender her autonomy and freedom to her self-addiction with AI-help.

The very notion of a free and "transcendent" individual, so engrained in contemporary moral theories, and in fact, considered to be a basic assumption of deontological and utility views, may also be problematic, particularly if such a free "choice-maker" is not tempered by empathic needs. Iris Murdoch (1970), whose work helped revive virtue theories in ethics, provides an attentional account of the ethically good person in terms of "loving attention." On Murdoch's account, one's acquaintance with and appreciation of a person should anchor one's actions toward that person, rather than merely following rules, recognizing the application of such rules, or repeating patterns of behavior according to one's own "rewards." Murdoch explains this in terms of an adequate *attentional anchoring*, which immediately orients proper ethical behavior without necessitating deliberate reflection on premises and conclusions regarding specific cases.[2] Attentional anchoring and guidance replace deliberation and intention in accounting for the *motivational* aspect of moral responsibility. In order to properly attend to the needs of others, we must effectively become virtuously insensitive to numerous irrelevant attentional targets including, according to Murdoch, one's own basic needs, biases, and desires. Attention must be *anchored and extended* toward the needs of others not via our self-oriented needs. Regarding the free and self-oriented maximizer or decision-maker, Murdoch has strong things to say:

> Kant abolished God and made man God in His stead. We are still living in the age of the Kantian man, or Kantian man-god. […] Stripped of the exiguous metaphysical background which Kant was prepared to allow him, this man is with us still, free, independent, lonely, powerful, rational, responsible, brave, the hero of so many novels and books of moral philosophy. The *raison d'etre* of this attractive but misleading creature is not far to seek. He is the offspring of the age of science, confidently rational and yet increasingly aware of his alienation from the material universe which his discoveries reveal; and since he is not a Hegelian (Kant, not Hegel, has provided Western ethics with its dominating image) his alienation is without cure. He is the ideal citizen of the liberal state, a warning held up to tyrants. He has the virtue which the age requires and

> admires, courage. It is not such a very long step from Kant to Nietzsche, and from Nietzsche to existentialism and the Anglo-Saxon ethical doctrines which in some ways closely resemble it. In fact Kant's man had already received a glorious incarnation nearly a century earlier in the work of Milton: his proper name is Lucifer.
>
> (Murdoch, 1970, 78)

Would an AI that feeds off reward signals from a reward-seeking self-oriented maximizer become a Luciferian artifice? Wireheading aside, is it a good idea to build AIs that will have no other purpose other than to satisfy their "owner" or "master" when she is self-obsessed with immediate rewards and gratifications? The modern "rational man" assumed in economics and moral theories has been criticized extensively on empirical grounds, but the issues raised by Murdoch are of an entirely different nature—they are not *descriptive* (i.e., whether "rational man" is indeed an accurate description of actual human beings) but prescriptive (i.e., whether "rational man" can really serve as the foundation for norms in moral systems). How could an autonomous agent that is alienated from others *without cure* serve as the foundation of the good life? This idealized rational man may well serve the purposes of describing an abstract citizen of the liberal state, distilled into formulas for utility-maximization, but without empathic needs, such an agent is inhuman, in the sense that she has no real experiential basis for moral needs—it is a merely idealized assertion of rational autonomy, made in a moral vacuum. Are we building a Luciferian world in which inhumanly characterized selfish humans satisfy their needs by using non-human intelligent machines?

Luciferian autonomy is representational autonomy. What is needed for a morally meaningful life is experienced, empathically based, autonomy. The abstract autonomy of Luciferian reward-driven agents that comply with norms and principles need not involve the concerns and needs of other agents as such—the only way other agents can be represented is through calculation, maximization, and mere rule-following. Thus, other human agents feature simply as an abstraction that is required for the theoretical framework to work. Perhaps this makes "ethical" AI easier to *design*, but if Murdoch is right, only at the cost of also making it Luciferian. Genuine moral behavior requires care and an *attentive orientation toward others* (similar to the kind of care that Wittgenstein calls *appreciation*). There is foundational value in such a categorical or unconditional cooperation, regardless of one's own representational and reward-based needs. Disinterest allows us to satisfy, collectively, the most

meaningful and transformative of human needs. Murdoch's warning is that there is a dark side to focusing too much on *our* needs, understood as *my* needs according to a value-reward model.

The selfish need to transcend by "going beyond," through novelty-seeking, and by constantly seeking to expand the limits of reward and knowledge "Icarus style" was indeed very much in Milton's mind when he wrote *Paradise Lost*. Transgressing and acquiring forbidden knowledge is a common theme in many tragic stories in literature. Transformative experiences have a value of their own based on the unique pleasure that discovery affords, and the experience of transgression enhances novelty. But there is a whole new set of troublesome needs lurking behind these selfish drives. Roger Shattuck's (1996) lucid book on forbidden knowledge demonstrates the deep connections between Prometheus, Milton's Lucifer, Goethe's Faust, and the characters in the stories of the Marquis de Sade. Novelty-seeking through transgression is thrilling and transformative, but it can be, and frequently is, *sadistic*.

Novelty-seeking, transgression, and boundless selfishness have unsettling consequences for AI design. The balancing act between satisfying my needs and caring for others is intricate, and given current AI design paradigms, the balance can be easily tipped toward sadism. An agent can have as a top preference caring for another human being and thus the AI should also care about the well-being of that person in order to satisfy the caring agent's preferences. But this can occur in a variety of ways. The caring agent may want the other person to do well and be willing to sacrifice her own well-being. Or the agent could care but not be willing to sacrifice anything. There is considerable room between radical altruism toward another person (say a family member) and borderline indifference. This is very problematic when trying to identify categorical needs: one of the needs at the top of a human's set of preferences is to protect very close family members, but at what cost to others? On the flip side, there has to be some minimum care for others regardless of who they are, so that we are not entirely indifferent to the well-being of fellow human beings.

But things are actually much worse than this. After all, sacrificing the well-being of others for the sake of those we love might be interpreted as a "biological imperative" (but see Moran's distinction between deliberative and theoretical stances in the discussion above). Caring for the well-being of all human beings is just impractical—although "impracticality" does not entail necessarily the moral *permissibility* of rampant indifference toward humans that characterizes our contemporary liberal and capitalist societies (Unger, 1996). Russell (2019, 229) introduces the term "negative altruism" to describe a much darker and

sadistic aspect of preference satisfaction. If an agent has complete disregard for the well-being of another human, she will take away resources from this person toward the satisfaction of her needs until the person is left "destitute and starving." Negative altruism occurs when the agent derives happiness "purely from the reduced well-being of others, even if her own intrinsic well-being is unchanged" (Russell, 2019, 229). Russell writes,

> In his paper that introduced preference utilitarianism, Harsanyi attributes negative altruism to "sadism, envy, resentment, and malice" and argues that they should be ignored in calculating the sum total of human utility in a population: "No amount of goodwill to individual X can impose the moral obligation on me to help him in hurting a third person, individual Y." This seems to be one area in which it is reasonable for the designers of intelligent machines to put a (cautious) thumb on the scales of justice, so to speak. Unfortunately, negative altruism is far more common than one might expect. It arises not so much from sadism and malice but from envy and resentment and their converse emotion, which I will call *pride* (for want of a better word). If Bob envies Alice, he derives unhappiness from the *difference* between Alice's well-being and his own; the greater the difference, the more unhappy he is. Conversely, if Alice is proud of her superiority over Bob, she derives happiness not just from her own intrinsic well-being but also from the fact that it is higher than Bob's.
>
> (Russell, 2019, 229–230)

Russell is absolutely right about the prevalence of negative altruism, but I think he underplays the role of sadistic needs. One of the examples Shattuck (1996) gives of contemporary sadism is pornography but we now think of pornography as an essential part of freedom of expression and communication—it is considered to be absolutely morally permissible. Should porn be as pervasive as it is? Watching hard core porn is clearly not the same as mass shooting a high-school with a machine gun (something, by the way, that is also a consequence of the *right* to possess very dangerous weapons, at least in the United States and some industrialized nations), but both have the sadistic structure of satisfying one's goals at the expense of another person's well-being for the sake of sexual and violent transgression. The dangers of being too moralistic about these issues are real, but a discussion about how our societies organize our regimes of attention by aligning them with selfish needs and commercial value, and the effects this has had on how we care for each other, is badly needed. As Russell says, this discussion is also fundamental to properly understand ethical AI. What Russell labels as "pride" is involved in achieving "capitalistic merit" and liberal "self-fulfillment." Russell insightfully refers to Fred Hirsch's research on "positional

goods" or goods that are not intrinsic to our well-being but that are based on relative value provided by comparisons—having the right car, education, or appearance. This "positional structure" of our values is indeed quite pervasive, and its growth has been fueled to unprecedented proportions by social media.

That being said, too much empathic orientation toward the needs of others, or too much altruism, can also be destructive. George R. Price, who formalized mathematically the biological basis of altruism, seemed to have experienced the "strength of the absurd" by expanding altruism in radical and biologically incomprehensible ways, for instance, by losing all his wealth in his effort to help the homeless. Something quite interesting, however, is that Price did this as an explicit assertion of moral autonomy, in order to show that altruism is not *merely* a biological necessity. Narcissism is a much safer way to live one's life than altruism and to most people, selfish narcissism, informed and guided by positional goods, is more reasonable than helping the other because "you have to depend on yourself and everyone is out there to get you." This is not the best moral perspective on life, and it is certainly an impoverishment of the kind of empathic and unconditionally cooperative capacities that made us the most dominant species. In fact, narcissism based on positional goods is incompatible with moral autonomy and value because positional goods are valuable only contingently. We have quite a bit of capitalist "Luciferianism" lately, and our empathic capacities are diminishing. Perhaps contemporary humans are so deeply narcissistic that we may really need AI to make us better? In any case, an AI designed to benefit such narcissistic agents will just exacerbate the problems of sadism, pride, and even wireheading.

5.5 AI and Collective Epistemic Agency

This section begins the transition toward social and collective aspects of AI. The work of Norbert Wiener and Daniel Hillis is discussed.

The previous discussion of human rationality completes this book's argumentation in favor of an attentional approach to general intelligence, including AGI, with the model of an individual epistemic and moral agent with various needs that she must satisfy autonomously. One possibility that must be discussed, however, is that AGIs may not really be "individuals," but rather, collective agents. Interestingly, as I shall argue in what follows, collective agents must also be attentive and satisfy needs in similar ways, so the basic framework

presented thus far applies to them, but with different consequences given the much greater influence that collectives exert on society.

With respect to ethical AI, and to conclude this chapter, perhaps we should not think of AI in terms of what Murdoch calls "Luciferian" ethics. As Wallach and Allen (2009) say, humans do not follow principles when they act morally, because what is needed is caring attention to the needs of others and this always entails difficult negotiations between satisfying selfish and altruistic needs. Selfish needs now abound because of the prevalence of positional goods around the globe. So selfish autonomy must somehow be balanced with altruism and empathy. AI should help in this effort if it is to qualify as "ethical." Since AI cannot be robustly autonomous from a moral and individual point of view—it will not be IEI—the only hope is to try to design *epistemically IEI AI* that gets at least a representational and theoretical understanding of the complexities regarding human needs.

But obviously we should not depend on AI to become better and more altruistic. Our political and economic systems need to improve—a problem that the Covid-19 pandemic made painfully clear. Since there are considerable risks involved in designing AIs that satisfy the individual needs of a selfish human being, one per human being if this is to be done democratically, perhaps we should rethink entirely what kind of system *should* qualify as AI. Is an AI "slave" whose goal is exclusively to satisfy the needs of a selfish human agent truly an example of *intelligence*? If the arguments provided so far are true, the answer is a resounding *no*. Such an "AI agent" would lack genuine autonomy, and thus, the motivations and needs required for the kind of agential attention that underlies *value alignments* (epistemic, moral, or aesthetic). Subservient AI could become a sad and perverse simulacrum of intelligence and a dangerous source of Luciferian manipulation. But what is the alternative?

An option that deserves careful consideration, and which is developed in the next chapter, is to consider really intelligent AIs as *essentially collective* epistemic agents. AIs would be collective in a couple of ways. They would be collective in their needs and motivations (these would not be the needs of a single selfish individual) and collective also in their attention routines and goals. The challenge is to explain how could collective agency explain epistemic responsibility based on genuine motivations and attention routines.[3] If AI is best understood in terms of collective epistemic agency, then the child machine will be more like NASA than an individual human being. A collective agent can thus be intelligently beneficial to *humanity*, rather than "pleasant" or useful to one human being at a time, in a way that guarantees minimum standards of altruism because of the impact

collective epistemic agents have on markets, legal systems, and communities of knowledge. The collective nature of AI could then really benefit humanity as a whole by not being based on an individual's preference maximization but rather on representations and rankings concerning objective measures and needs that must be met. Governments may not like this idea, but designing ethical AI should not be based on whether or not it pleases government officials—although clearly, global consensus is quite important for AI design. Below, I argue that human rights, understood in terms of agential needs for capabilities and competences, can serve as the basis for global consensus.

The notion that AI is more adequately conceived of in collective terms goes back to the work of Norbert Wiener. Daniel Hillis starts a brief essay on this topic with the following passage from Wiener's (1950) *The Human Use of Human Beings*: "Whether we entrust our decisions to machines of metal, or to those machines of flesh and blood which are bureaus and vast laboratories and armies and corporations, we shall never receive the right answers to our questions unless we ask the right questions" (Hillis, 2019, 172). Hillis continues,

> Norbert Wiener was ahead of his time in recognizing the potential danger of emergent intelligent machines. I believe he was even further ahead in recognizing that the first artificial intelligences had already begun to emerge. He was correct in identifying the corporations and bureaus that he called "machines of flesh and blood" as the first intelligent machines. He anticipated the dangers of creating artificial superintelligences with goals not necessarily aligned with our own. What is now clear, whether or not it was apparent to Wiener, is that these organizational superintelligences are not just made of humans, they are hybrids of humans and the information technologies that allow them to coordinate. […] These artificial intelligences have superhuman powers. They can know more than individual humans; they can sense more; they can make more complicated analyses and more complex plans. They can have vastly more resources and power than any single individual. Although we do not always perceive it, hybrid superintelligences such as nation-states and corporations have their own emergent goals.
>
> (Hillis, 2019, 172–3)

The first AI will very likely be designed by a large corporation, or by a conglomerate of governmental and commercial interests. A superintelligent AI designed with *collective goals* will be a natural way of proceeding given that many of its capacities will be based on collective hybrid intelligences: collective data-gathering, surveillance, and decision-making. Ideally, a beneficial and

genuinely intelligent AI will be responsible for her decisions because of the control she has over her cognitive life. Such an AI will be a competent *interpreter of needs* and arrive at very complex decisions to balance healthy selfishness needs in a way that is compatible with increasing levels of altruism. AIs will be much better and faster than governments, markets, and corporations at doing this, which could rapidly eliminate corruption and tribal interests. Their being attentive in representational and rational ways that are ethically beneficial need not demand IEI moral autonomy. EEI moral AI might be sufficient, given that collective *epistemic agency* provides enough safety and control—although this and the previous chapter are a warning concerning the importance of transcendental needs and how difficult it is to represent anything meaningful about them without having the appropriate conscious and emotional reactions, which shows that even EEI moral agents should not make decisions in situations that demand moral appreciation (Sharkey, 2020).

The altruism promoted by AI should include the environment and other species. A truly intelligent AI will appreciate how similar human and animal biological and emotional needs are. Since AIs will not be obsessed with positional goods and egomaniacal needs, they could really become an enormous source of ethical benefits. If all goes well, AIs could help humans flourish by making altruism much more prevalent. So far, we have no reason to believe that this will happen. On the contrary, we have plenty of evidence that AIs will be designed in a way that exacerbates the ubiquity of sadistic and Luciferian needs. Ultimately, given the radical transformation of humanity AI may entail, the ethical implications of AI should be one of the most urgent and central tasks for international law, global markets, and politics. Concerning the politics of AI, Hillis envisions four scenarios. An "obvious" scenario is that AIs are *controlled* and "allied with, individual nation-states." Given what has been said so far in this book, this kind of AI will not really be genuinely intelligent since it is subservient to the commands and needs of a state. It will not have autonomy and independence, and therefore no genuine intelligence, and worse, it could be a militarized form of super advanced automation. A second, and even more likely scenario is that the AI will be driven by the interests of a large corporation. This would generate the political risk of giving enormous power to a handful of greedy impresarios—much more power than any nation-state. A third, more catastrophic scenario, is the "terminator" situation in which AIs are not aligned with human values and pursue their own interests. Ironically, this scenario is the only one fully *compatible* with genuinely intelligent AIs, but it is indeed

catastrophic because of all the risks mentioned in previous chapters concerning value, autonomy, and control. In the fourth scenario:

> Machine intelligences will not be allied with one another but instead will work to further the goals of humanity as a whole. In this optimistic scenario, AI could help us restore the balance of power between the individual and the corporation, between the citizen and the state. […] AI will empower us by giving us access to processing capacity and knowledge currently available only to corporations and states. […] We may not fully understand or control our destiny, but we have a chance to bend it in the direction of our values. The future is not something that will happen to us; it is something that we will build.
>
> (Hillis, 2019, 177)

To be sure, of the four options presented by Hillis, the fourth option is the only one that prevents the development of sadistic or Luciferian AI because the value alignment is between a collectively understood AI agent and the shared collective values of humanity. If the promise of unprecedented wealth production by AI through large-scale intelligent automation materializes, an ethical AI should facilitate the creation of a new economy, which will be vastly innovative and transform the topography of human needs and in which altruism will be the rule rather than the exception. To achieve this goal, AI need not be phenomenally conscious. On the contrary, by not being emotional, AIs will be, as Turing envisioned, less prone to egomaniacal or "solipsistic" mistakes. Maybe they will never be morally autonomous because of this, but they can certainly be superintelligent and help us make very difficult decisions that align with our most basic needs. This would certainly be a welcome development, given our uncertain futures overcast with global warming and neo-nationalisms. AIs will not be at all like us because their agency is free from biological needs. They will likely be duplicated quickly and their presence will spread across times and regions. However, for AIs to be genuinely beneficial we need to reconceive them, perhaps not as personal assistants (although this type of AI will certainly play a role in industry) but as genuine and autonomous *collective epistemic agents*. Autonomy is unavoidable if intelligence is involved. But what is collective autonomy?

6

Are AIs Essentially Collective Agents?

6.1 Artificial Intelligences as Collective Thinkers

This section explains the possible advantages and disadvantages of collective artificially intelligent agents. It introduces political issues, such as the importance of collective AI for democracy and the public sphere.

This chapter expands on the conclusion of the previous one, namely, the claim that AIs may best be conceived of as *collective epistemic agents*. This possibility presents two immediate problems. First, in what sense are collective AIs autonomous epistemic agents? And second, how could we interact with these collective agents if they are so different in cognitive architecture and capacities? If it is already difficult for humans to interact with governments and corporations, why think that interacting with AI collectives would be less difficult, or a good idea in the first place? These questions only make sense if we assume that AIs are indeed collectives, somehow similar to governments and corporations in their complex data gathering, decision-making, and influence. If they are not, then we need to confront all the challenges about modeling AI in terms of individual human psychology, which were explored in detail in the previous chapters. It is because of the difficulties that individual human psychology presents for AI that conceiving them as collectives seems to be a good alternative, or at least one that deserves serious consideration. Thus, this chapter addresses these two key questions under the assumption that AIs are different not only because of their cognitive architecture, but also because of their collective nature.

One development in AI regulation that already leans toward considering AI collectives is the legal protection of AI. Legal and political issues will be examined more carefully in the next chapter. The basic idea behind this proposal is that if AIs become important players in industry, science, and human affairs at large—as many expect—then protecting and regulating them under corporate

law would allow at least for some level of surveillance over them, and also for a well-understood type of responsibility-attribution and legal personhood. Clearly, this proposal assumes that AIs will be capable of some minimal epistemic agency to engage in legal transactions, and thus it is crucial to address how exactly they will be capable of autonomous epistemic agency in order to have legal standing.

Since AI is here conceived under a different characterization of agency, an entire new set of risks will emerge. Many will be familiar, such as the need for competence and epistemically virtuous skills (attention and inference) in order to ground trust. But some will be unique. One of them is the "brittleness" of Collective AI (or CAI). Depending on their degree of integration and general intelligence, some CAI may be more susceptible to failures based on new information that cannot be adequately processed because of a rigid design. This is a general problem for AIs because either the opacity of their reinforced learning processes prevents humans from comprehending their "reasoning," or because the algorithmic procedures they follow cannot intelligently select what is relevant in a new situation. But a CAI that has aggregation rules for compiling information from multiple sources, which may include humans, will depend too much on such rules for decision-making. This can be an obstacle for the flexibility required for AGI. But if these rules are flexible enough CAI will have advantages concerning adaptability, the way institutions, when properly designed, can adapt to very rapid social changes. This chapter argues that answering this and related questions about CAI will depend on solving what is described below as the *interface problem*.

The "fragility" of CAI may also be a unique difficulty. Hacking into systems is already a costly and increasing problem. A unique risk of a highly dispersed and extended AI is that it may risk being overtaken by other agents given its lack of agential integrity. Scattered, unorganized, and otherwise to easily manipulated CAI will not count as trustworthy. A "shallow" CAI organized simply in terms of aggregation procedures, as some organizations are, would not be resilient enough. Here issues of industrial design for safeguarding the integrity of the CAI will be very relevant. Humans and animals solve this problem through their highly and virtuously integrated epistemic agencies, which shields them from too much external influence over time. For CAI these will be major challenges (although these are challenges for any kind of AI, albeit probably not as substantial as for CAI).

CAI may also present unique advantages, although this issue also needs to be explored more carefully. Given their wider range of resources and their intrinsic social standing as salient sources of information, it might be easier for CAI to

become generally intelligent AI (AGI). In philosophy of mind, an advantage of the view called "extended cognition" is that multiple sources of information need not be contained in an agent's functional organization or hardware—contents of an agent's mind may be distributed throughout different environments and databases. AI is already conceived of in this way and if it is intentionally designed as CAI it may learn more quickly to be general, but obviously more sources of information, extended throughout the globe if internet is used, are not going to produce any kind of intelligence unless the issues concerning epistemic agency are also satisfied, including reliable capacities for joint attention—this is a disadvantage of systems like GPT-3, as mentioned in the introduction.

The biggest challenge that emerges from CAI concerns difficulties in our *interactions* with them, given the degree of social influence that collective epistemic agents have, such as government agencies or corporations. This is at the root of the interface problem explored in this chapter. Since the capacities of these CAI vastly surpass the resources of any individual, an *intelligence jet lag* is a major risk. This also happens now with corporations, banks, and agencies such as CERN and NASA. For this reason, the powers of surveillance of CAI will pose very significant political risks for the population—something also already occurring in intelligence agencies and police forces.

Besides these risks concerning communication, discrepancies in intelligence and information gathering, loss of information in the transmission of information, and other related problems, there are also problems concerning the epistemic status of CAI. For CAI to count as epistemic agents, they must satisfy the same standards as individual agents regarding abilities and cognitive control. Clearly this will be more difficult for CAI, but the same constraints apply for their attention routines and assertions. Since the topics of attention and cognitive control in human psychology were explored at length in previous chapters, the emphasis here is on *social epistemology*—a somewhat recent subfield of epistemology. The motivations and needs that are associated with epistemic achievement at the individual level—the need to represent the environment correctly, making salient relevant information and inhibiting irrelevant information, as well as the need to structure cognitive processing in a rational way—must also be present in CAI. If they meet this requirement, CAI may become, as already mentioned, the most influential and powerful kind of intelligence.

The next chapter focuses on why the similarities between CAI and governments or corporations justify studying them as essentially political actors with more power than any "personalized" AI. Trusting these AIs will be equivalent to trusting legal and political institutions. Political institutions might

take advantage of this and insert their own goals and specific agendas into what they will consider *their* CAI. Different political and legal cultures will shape CAIs and this would create the risk of over-politicizing AI unnecessarily, eventually making CAI yet another arena for traditional political warfare. The politics of CAI will determine who gets to shape the goals and motivations of what could become the most powerful decision-makers. Addressing this question will be a fundamental component of a larger discussion concerning the future of democracy, liberal autonomy, and the very notion of a *public rational sphere*.

This chapter analyzes a unique difficulty concerning the production of knowledge that is deeply related to the issues stated above—the problem of creating a *radically expansionist* kind of social epistemology. This difficulty shows that, unless epistemic interface problems are solved, there cannot be epistemic trust and accessible knowledge. The backdrop of these broader epistemological and social worries is the interaction between human epistemic agents and CAIs that qualify as AGIs. This expansionist type of social epistemology will incorporate non-anthropocentric forms of knowledge production in a fundamental way, which requires comparisons with traditional issues in philosophy of mind and epistemology, such as cognitive architecture, explanation, and inferential reasoning. Given the ever-increasing roles of AI as tools in knowledge production, the interface problem opens the possibility of CAI as knowledge producers in their own right. Unless the interface problem is solved, an "intelligence jet lag" between humans and CAI will be a permanent feature of any non-anthropocentric and expansionist social epistemology.

6.2 The Interface Problem of Intelligence and Knowledge

This section focuses on knowledge production by CAI and various interface problems that it generates. It offers a brief survey of interface problems in philosophy of mind and epistemology.

There is a widespread use of automation in collective human epistemology, whether this be in improving search algorithms for browsing the internet (Lardinois, 2012), assisting doctors in medical diagnosis (Robinson et al., 2014; Shibata, 2004; Turkle et al., 2006), guiding automated cars (Greenough, 2016), or any number of other activities, including attempts at stopping the spread of fake news (Simonite, 2017). Here the focus is on a use of AI in collective *knowledge production*, which may involve a plurality of different forms of intelligence, rather

than a single AGI (Kelly, 2016). Because of the increasing reliance on AIs in the fields of mathematics and various sciences (Hassabis, 2017; Voevodsky, 2014), AIs are shifting their role in our scientific endeavors and are becoming crucial *members* of our scientific communities. AIs have been used as knowledge tools, assisting researchers and those in industry in developing products, performing computations, gathering information, and advancing scientific knowledge. What is important for the coming generations of AI is not their roles as knowledge tools, but rather their roles as knowledge producers—in the case of CAI, the most authoritative knowledge producers.

Mapping the epistemic difficulties surrounding the current development of CAI requires an inquiry into the "epistemic interface" between AIs and humans. This is an interface by which one agent transmits epistemically significant information to another. Information transmission in an epistemic community can be shaped by substantial asymmetries in influence and trust. Key transmission problems concern the identification, evaluation, and integration of information concerning knowledge and other epistemic achievements, such as justification and reliable assertion, but now for communal purposes. Focusing on communication interfaces that prevent radical asymmetries between the "epistemically rich and poor" must be a priority in the design of CAI (and AI research in general).

Given the centrality of reliable information exchanges for the proper functioning of epistemic communities, it is not surprising that traditional problems in philosophy of mind and epistemology are framed as interface problems. An example of an interface problem concerns the loss of information from one format or type of information to another. Uncontroversial cases of this type of interface problem involve the distinction between analog and digital formats of information (Dretske, 1981; Haugeland, 1981; Maley, 2011). Fred Dretske (1981) explicitly made this distinction in terms of information loss. A similar idea is behind the distinction between cognition and perception, more specifically between iconic information and conceptual information (Block, 2014; Burge, 2014; Carey, 2009; Fodor, 2007, 2008). The distinction between conceptual and non-conceptual content, particularly if understood in terms of "fineness of grain," seems to presuppose a similar understanding of interface, as one loses information from the non-conceptual (or phenomenal-iconic) to the conceptual or symbolic format (associated with *propositionally structured* epistemic access).

What is an interface, then? It is a kind of informational mapping. Mappings can be isomorphic (identical in structure) or homomorphic (similar in structure).

Since some information is lost in the cases above—from iconic to symbolic, analog to digital or non-conceptual to conceptual—the mapping involved must be homomorphic. Despite the similarity in information, some information is lost from one format or structure to the other. With respect to CAI, the loss would occur when transferring information with *epistemic value*. Unless the interface problems detailed below are solved, we will face the perplexing situation that increases in intelligence will not correlate with increases in knowledge. CAI and social institutions will become very intelligent, but individual human beings will not be more knowledgeable.

This problem extends to basic communication in general. In a typical speech act, the intention of the speaker to communicate must be recognized by the person receiving the message and they must be jointly attentive and motivated to communicate by assuming the same background information, linguistic norms, and linguistic intentions. As mentioned, this is a problem with AIs that "communicate" without jointly attending, like GPT-3. For individual human beings, severe interface problems rarely emerge because of the similarity of their cognitive architectures concerning attentive motivations and goals, but the situation is entirely different in social epistemology (as well as with non-human animals). While it is relatively easy for me to recognize and understand the speech acts (e.g., jokes, assertions, questions, commands) of my neighbor, this is not the case, unfortunately, with respect to governmental agencies and scientific boards. Given the explosive increase in information processing of AIs and how much epistemic collectives rely on them, there are various difficulties concerning how AIs share, create, and distribute knowledge, with consequences for democracy and the public sphere.

A core characteristic of *epistemic* interfaces is the capacity to be flagged as a trustworthy source of information by other members of the community, based on the capacity to produce knowledge autonomously. It is uncontroversial that computers and AI produce, distribute, and store information used for collective purposes and that they are considered as reliable sources of information. The challenge for epistemically relevant CAI is to determine whether or not they are truly *epistemic agents*: Can they produce knowledge that is epistemically valuable from a human perspective, and can they identify justified beliefs? If it is true that CAIs are epistemic agents, all sorts of epistemic communities are about to experience an unexpected increase not only in information production but also in knowledge production. But there is no way to tell the difference between tool and producer without solving the interface problem: what is the mapping between *artificially produced knowledge* at the collective level and human knowledge in

general; how much epistemically relevant information (and epistemic value) is *lost* in their interaction (e.g., the justification and identification of speech acts by CAIs, the production of mathematical proofs by AIs)?

6.3 Expansionist Knowledge Production and the Interface Problem

This section explains the epistemic interface problem and introduces Alvin Goldman's notion of an expansionist social epistemology. It argues that CAI would entail a radically expansionist social epistemology. The political consequences of "intelligence jet lag" are discussed.

The potential asymmetry in knowledge production between ultra-capable CAI and humans generates the epistemic interface problem. Stated briefly, this is the problem of how human epistemic agents and artificial epistemic agents will exchange information with one another with regards to collective knowledge production, for instance, mathematical and/or scientific problems, in an *epistemically fruitful* and valuable way. How will human epistemic agents be able to enter into a *meaningful* epistemic exchange with CAIs, and furthermore, how will we *know* that this exchange is meaningful and veridical? Humans bridge the gap between personal intelligence and collective knowledge production by being motivated to succeed in various epistemic tasks and, crucially, through epistemic trust based on similar epistemic competences. This form of integration between goals and motives requires some form of collective agency and joint attention—either "thin," understood simply in terms of aggregation or voting procedures or "thick," in terms of capacities, virtues, and even character traits. The more information integration a collective has, the more it will be able to have the kind of virtuous epistemic constitution that humans have (Fairweather and Montemayor, 2017). We frequently refer to collectives in these terms: the racism of the police force, the incompetence of an intelligence agency, the perspicuity of NASA, and so on.

CAI-based knowledge will become an extremely impactful source of information which will transform the degree of reliability and speed of *knowledge transmission*. It will very likely become the most important source of knowledge production, leaving behind even the best integrated human-based epistemic collectives. This presents great opportunities and dangers. How to understand the impact of such a powerful and *new* source of knowledge? For CAI to be good

news, the interface problem must be addressed successfully. Solving the interface problem will be part of what Alvin Goldman (2012) calls "expansionist" social epistemology, because although the interface with CAI is not explicitly addressed in traditional (individual) epistemology, it is certainly continuous with traditional epistemic problems. In fact, solving the interface problem with CAI is the most urgent and interesting expansionist project in social epistemology because it will eventually affect how all kinds of social systems with epistemic impact are evaluated, from expert systems in medicine and government to large-scale scientific collaboration. We already witnessed the impact of supercomputer-based scientific outcomes during the Covid-19 pandemic regarding diagnosis and vaccine production. This is just the beginning of a much larger trend.

To better appreciate the potential role of CAI, it is useful to use Goldman's (2011) characterization of social epistemology (SE). Goldman breaks social epistemology into three separate spheres. The first variety of social epistemology (SE1) concerns individual epistemic agents and their relation to *social evidence*, or the role social evidence plays in influencing an individual's doxastic choices. The second variety (SE2) concerns *collective doxastic agents*: groups such as governments, courts, and corporations, which arrive at doxastic judgments (including beliefs and assertions) as an autonomous collective. The third variety is what Goldman calls "systems-oriented (SYSOR) SE," which studies *epistemic systems*. According to Goldman: "'epistemic system' designates a social system that houses social practices, procedures, institutions, and patterns of inter-personal influence that affect the epistemic outcomes of its members" (Goldman, 2011, 18). Figure 1 illustrates Goldman's three spheres of social epistemology. The spheres SE1, SE2, and SE3 are open because there is overlap between these spheres in terms of the kinds of epistemic outcomes they produce, and so although they are different spheres some cross-pollination likely occurs. How does CAI fit into this picture? It seems clear that the individuals who compose these social interactions are in fact all of the same epistemic architecture, that is to say, human. Although a corporation itself is not human, it is composed of individual humans, and thus its doxastic judgments are brought about through human doxastic (belief-based) capacities. And in the case where no specific human is involved directly with the doxastic judgment, as is the case in SYSOR SE, those systems are still the result of human doxastic capacities. How does this change when we cannot assume that those interactions are all composed of the same type of individuals with the same cognitive architecture? Particularly, how to include CAI, with its radically different architecture, in a way that guarantees epistemic trust?

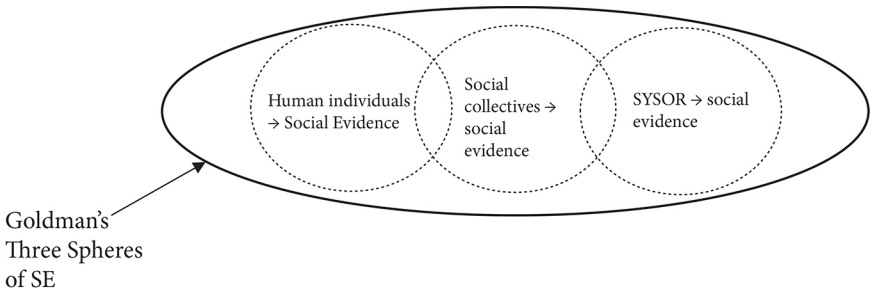

Figure 1 Goldman's three spheres of social epistemology—*Anthropocentric.* © Carlos Montemayor in collaboration with Garrett Mindt.

CAI's inclusion in social epistemology is a transformative and risky expansion of the dimensions of epistemic exchanges concerning: (i) individual (or traditional) human epistemology; (ii) individual human and non-human epistemology or intelligence; (iii) hybrid communities of individual and collective knowledge tools (such as computers); (iv) collective epistemic agents interacting with individual ones, based on *human* intelligence; and (v) collective epistemic communities, composed of collective epistemic agents, human based, *and CAI*. This suggests that the standard classification given by Goldman needs to be updated. What variety of SE does human and CAI epistemic activity fall under? Does one, say, place a team of human and AI researchers working on a mathematical proof or discovering new pharmaceutical drugs under SE2? This would seem to ignore that we don't have good reasons to suppose that human and AI doxastic capacities are of the same kind. A second group of related spheres to SE is required, building off of Goldman's three spheres. Figure 2 illustrates a *non-anthropocentric social epistemology*, with CAIs as knowledge producers.

These spheres are corollaries of those outlined by Goldman. The first sphere (NASE1) would correspond to any AI engaged with social evidence. Perhaps AlphaGo is a close precursor of NASE1, because it is designed to perform the task of playing Go at the highest possible level, engaging with the socially collected games of Go that the network was trained on in order to master the game. NASE2 would then be any CAI engaged with social evidence. An example of CAI with social evidence (NASE2) might be two distinct AIs tasked with producing some epistemic outcome with different strategies or architectures (e.g., two systems set to develop some new drug therapy). The last sphere in the graph, AI SYSOR (NASE3), is more difficult to illustrate, as technology has yet to reach the stage

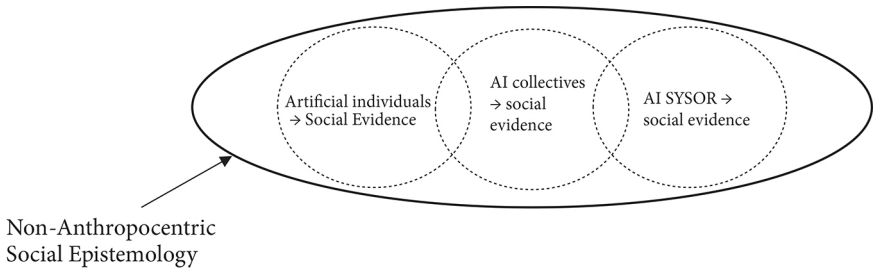

Figure 2 Non-anthropocentric social epistemology. © Carlos Montemayor in collaboration with Garrett Mindt.

where a particular CAI could be considered as a social institution. But given the increased advancement of AI, its application to real-world problems, and its continued increasing role in the economy, it's not outside the realm of possibility that such social institutions would arise in the future.

To come to grips with the role of CAIs in social epistemology, we first need to establish clear criteria for how humans and AIs will interact with one another toward a shared epistemic goal. These criteria are subsumed under the umbrella of the interface problem, or the problem of how human and artificial epistemic agents will cooperate with one another as a community. As Figure 3 shows, this kind of epistemology is *radically expansionist* because it creates a new sphere of *non-anthropocentric knowledge production*. The two dominant groups of social epistemology are *anthropocentric SE* (Goldman's original three spheres of SE) and *non-anthropocentric SE*. The lines of these two major ovals in the figure are dotted to indicate that the spheres of the anthropocentric and non-anthropocentric groups interact, illustrating the interface problem. The individual spheres within the two larger ovals are dotted as well because they allow for interaction among the different spheres across anthropocentric and non-anthropocentric lines. Taken as a whole we see social epistemology as encompassing all the domains within. The interface problem takes place at the intersection of anthropocentric and non-anthropocentric SE groups.

The interface problem might be characterized as a problem of communication. There is a sense in which we humans already have a problem of effectively communicating with one another about common epistemic goals. Some of these problems concern, besides unclear goals, unclear epistemic motivations, and very diverse standards for justification. There is plenty of room for mistrust and manipulation, and this problem is drastically exacerbated with CAI. How could we, as epistemic agents, contend with the problems of effective communication

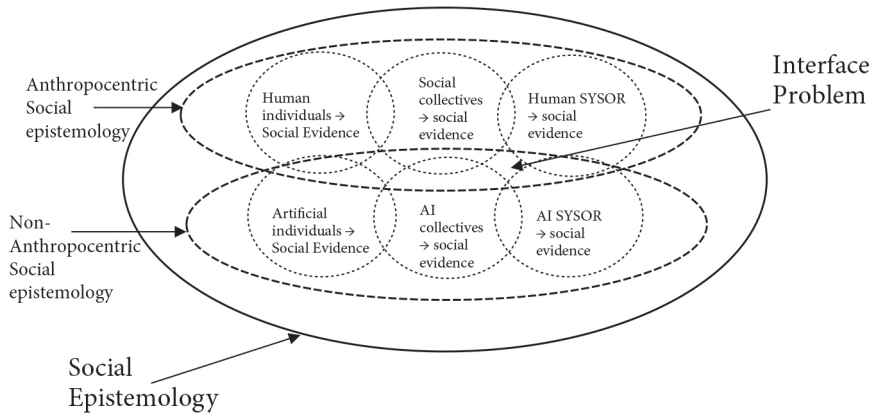

Figure 3 Spheres of a radically expansionist social epistemology. © Carlos Montemayor in collaboration with Garrett Mindt.

not only across disciplinary boundaries, but also across the divide between human and artificial epistemic agents? What reliable methods of interfacing are currently available? Furthermore, what guarantees that meaningful and veridical communication is occurring? These are all issues familiar to those in AI research who work on avoiding underspecification and opacity. But the key infrastructure to really tackle this issue, given the value and social nature of knowledge, is *cultural and epistemic*—it is not merely a *technical* issue.

If the interface problem is not solved, then there will be an *intelligence jet lag* between epistemic agents, and potentially entire epistemic communities, decisively fracturing the social cohesion of human knowledge production and creating pockets of permanent epistemic poverty. De facto political oppression and the impossibility of rational communication would be irreversible. Intelligence jet lag will prevent socially produced information from constituting genuine knowledge capable of being transferred through testimony and other standard means of knowledge distribution, such as honest and trustworthy assertion. Epistemic agents will have substantially different *temporal and motivational* constraints with respect to information processing, which will create impenetrable barriers for the collective production of, and access to, knowledge. The most significant of these temporal constraints concern computational complexity and the different, non-anthropocentric, cognitive architectures and epistemic communities that CAI will produce.[1] The most significant temporal asymmetries concern misaligned hierarchies of epistemic needs based on "processing time." CAI will operate on a different timescale and timetable, with a

radically different temporal perspective. Short-term human goals will very likely be incompatible with decision-making based on AI processing that is entirely free from, for instance, biological short-term needs.

Figures 1–3 show how intelligence jet lag can take place in a *layered way*. Starting with the individual anthropocentric level, collectives generating specialized knowledge may create forms of intelligence jet lag that make individuals demoted epistemic agents and decision-makers—a reality we live with today in many fields of inquiry. As we move up the ladder of epistemic organization, most human beings will be rendered inconsequential and fully "downgraded" by communal design. Many decisions will involve collectives rather than individuals and the epistemic life of individuals will not be as prominent as that of collectives (an issue related to Stuart Russell's notion of *enfeeblement*). A similar type of jet lag may emerge when more systematic and better integrated forms of collective epistemic agency gather information more rapidly, efficiently, and reliably. CAI will quickly accelerate epistemic demotion and a permanent kind of epistemic injustice, unless the interface problem is solved.

This issue is both political and epistemic. Epistemic injustice occurs when a perfectly adequate assertion that satisfies an epistemic norm (it is true and constitutes knowledge) is ignored by an epistemic community simply because of the perceived status of the speaker (Fricker, 2007). More pervasive forms of epistemic injustice can be based on cultural domination, through a form of epistemic colonialism that erases the style and temporal landscape of other communities; it can also be so "structural" that it may constitute a kind of violence toward a group (see Isasi-Díaz and Mendieta, 2012 for the effects of colonial epistemologies, and Dotson, 2011, for the notion of *epistemic violence* through systemic silencing). All these *anthropocentric* forms of epistemic injustices, in their increasing degrees, will be dramatically aggravated through the corporate or governmental use of CAI—similar to the problems of automatized poverty and criminalization produced by biased databases (Benjamin, 2019; Eubanks, 2018). The ultimate danger presented by intelligence jet lag is the permanent demotion of entire human epistemic communities. The most radically expansionist forms of social epistemology concern non-anthropocentric kinds of intelligence.

Increases in intelligence generated by humans typically correlate with increases in knowledge but, in interactions with CAI, dramatic increases in intelligence(s) may not correlate with dramatic increases in knowledge. Knowledge is an epistemic good of *higher value* than intelligence or mere

problem-solving because it involves a unique type of association with epistemic agency. The following section addresses three of the major sub-problems which fall under the umbrella of the interface problem. The problems are: (1) the epistemic *justification* problem—what are the conditions that must be obtained in order for a CAI to be justified in asserting a claim, as well as, how does the system report such justification? (2) the epistemic *access* problem—how could CAI reliably access its own relevant inferential processes and report those to other epistemic agents with completely different architecture? (3) the meta-motivation problem—aside from motivating inquiry into a particular problem, for instance, setting out to solve a particular scientific problem or develop a mathematical proof, why should inquiry be initiated in the first place? How could CAI have the attentive meta-motivation to independently open an inquiry into a particular problem that is *epistemically adequate and relevant*? These problems show that without a proper interface, the traditional notions of human *epistemic authority and expertise* will no longer be applicable in a radically expansionist setting for social epistemology: one would have to take the dicta of CAI as if from an oracle, not susceptible of peer review or verification.

6.4 Epistemic Dilemmas Created by CAI

This section illustrates a collective justification problem with knowledge production in mathematics, a collective information-access problem with speech-act theory, and a collective meta-motivational problem with scientific inquiry. It argues that all these problems can only be solved through joint attention-interfaces.

The *epistemic justification* question for CAI is: in solving a problem beyond human epistemic capacities, how are humans to know whether the result of the program or the solution to the problem is epistemically justified or not just accidentally true? Mathematical proofs are considered to be the most rigorous kind of inferential reasoning. Automation, however, is changing the epistemic dynamics of knowledge production within the mathematical community. On the one hand, mathematicians need to rely on computer power to arrive at ever more complex proofs. On the other hand, mathematicians cannot simply trust computers as if they were oracles. Why should the mathematical and scientific community simply trust CAI or, more generally, computer-based proofs? Mathematical proofs illustrate a clear case of the highest epistemic standards, which in humans involve deductive reasoning, explicit logical inference, and attention to abstract

contents. But it is easy to imagine similar problems that extend to problems with lower epistemic standards, including easier mathematical puzzles and games, or simple conversational and communicational tasks.

This problem is a pressing issue regarding the foundations of mathematics, and for its future practice and development. Ian Hacking (2014), for instance, says that the reason there is philosophy of mathematics is because of the *experience* of demonstrative proofs. He introduces a distinction between what he calls the Cartesian and Leibnizian conceptions of proof. A Cartesian proof is characterized by the experience of clearly grasping the proof before one's mind "all at once." By contrast, a Leibnizian proof depends on reliable reproducibility, or the possibility of arriving at the proof mechanically, by rote. Trust in "Cartesian proofs" derives from what Boghossian considers to be the most explicit and conscious kind of "taking condition" while trust in "Leibnizian proofs" derives from their verified track record of truth-conduciveness. Hacking documents how the tension between Cartesian and Leibnizian conceptions of proof has played itself out many times in mathematical debates, for instance between the mathematicians Alexander Grothendieck and Vladimir Voevodsky, whose views on proof Hacking identifies as correspondingly Cartesian and Leibnizian.

While Hacking offers reasons in favor of both views, he clearly prefers the Cartesian notion as essential to the nature of mathematics. At many points, Hacking expresses real concern that the Leibnizian conception may become the dominant view about mathematical proof, and become the new epistemic standard of the mathematical community (2014, 25, 84, 141). In fact, he acknowledges that the increasing presence of long proofs that require computerized verification in mathematical practice may make Leibnizian proofs prevalent very soon. His main worry is that the nature of mathematics seems to necessitate robust epistemic access to verification accompanied by the unique conscious experience of understanding associated with it. The opposite is not only epistemically arid, but potentially *unverifiable*: "An author submits a paper with a proof or proof sketch, together with a programme for checking the proof, and a confirmation that, when run, the computer says, 'OK'. Who checks that the programme is sound?" (2014, 25–6).

Hacking claims that Cartesian proofs "carry understanding and conviction with them" (2014, 115). This conviction is key to appreciate their normative role—one *should* accept them if one is a responsible epistemic agent. In the Leibnizian case, one is convinced because of automatic and reliable procedures. Yet, this notion of proof cannot account for the experience of mathematical discovery and the accepted standard of normative strength typically associated with mathematical

reasoning and deductive inference. According to the Cartesian approach, the epistemic acceptance of a proof depends on the application of rational norms that were clearly understood and followed based on premises that led to mathematical knowledge-production by proof. The "experience of proof" has these two characteristics: clear understanding and normative guidance. By definition, Leibnizian proofs lack these characteristics. Much of what fascinated Plato and Immanuel Kant about mathematics is found in the unparalleled normative force and broad epistemic implications of mathematics. Kant gives expression to this unique epistemic status by characterizing mathematical knowledge as a priori and synthetic—highly informative, necessary, and based on reason alone. Surely the experience of understanding a proof is crucial to defining the nature of mathematics and its practice. If CAI takes over the task of providing mathematical proofs, then their normative epistemic status is in jeopardy.

But perhaps Cartesian proofs only *seem* to be necessary from *our* perspective. As mentioned before in the discussion on inference, the human experiences of "intellectual seemings" may not be necessary for the practice of mathematics. Perhaps CAI will open the possibility of kinds of mathematics never thought by a human but which are radically expansionist of mathematical knowledge—an unprecedented resource for the practice of mathematics. If so, surprisingly, preventing this possibility from coming to fruition, or even intentionally slowing it down, would be *epistemically irresponsible*, at least from a social epistemology point of view. Might it be the case that what is epistemically *responsible in the individual* agent case (follow rules explicitly) becomes epistemically *irresponsible at the social epistemic level* (preventing the expansion of mathematical knowledge through CAI based on anthropocentric reasons)?

This is an epistemic problem of justification from the human point of view. But there will be versions of this problem in which CAIs start "talking" to each other, perhaps even in their "own language" (Griffin, 2017), faster and more efficiently than any kind of human epistemic communication. Thus, there is a dilemma concerning epistemic justification in a radically expanded social epistemology of mathematics that includes CAI. Either the Cartesian conception is the correct account of proof or the Leibnizian conception is correct. If the Cartesian conception is correct, then no reliable proof produced by CAI can be accepted unless we have an interface for epistemic justification *equivalent* to the experience of understanding a proof by CAI. If the Leibnizian conception is correct, then proofs must be accepted without any understanding of how the conclusion was reached. In either case, without an interface, radically expansive knowledge seems impossible.

An interface is indeed critical to eliminate this dilemma. Notice, however, that regardless of how this issue is solved, an interface will require *joint attention and joint meaning* within any epistemic community. Thus, Leibnizian proofs can only count as EEI AI—they are strictly equivalent in results, but neither in reasoning nor attended meanings. In contrast, an IEI AI would be an *attentive agent*, and satisfy epistemic goals. A key contribution of the analyses in previous chapters for the present discussion on social epistemology is that an IEI CAI *need not be conscious in order to have full epistemic standing*. This, as was argued in Chapter 3 regarding inference, eliminates the dichotomies of the conservative and liberal views. Without a *jointly attentive epistemic interface*, we would face either demotion from epistemic communities or produce an epistemically irresponsible deceleration of knowledge production and distribution. An interface of knowers in expanded epistemic communities will guarantee *justificatory standards* on the basis of agential attentive skills.

A second and very similar dilemma concerns the *quality of epistemic access* humans and CAI will have to information. What will be the semantic and representational types of access to information used by CAIs? CAIs will likely not think the way we do and in many instances will only have a derivative kind of intentionality, based on a set of comparisons with human psychology. However, in cases where CAI successfully identifies and produces language and speech acts, we will be hard-pressed not to attribute to them *some* kind of access to information (GPT-3 could be a very early precursor of such systems). Humans depend on *joint motivations* or communicative intentions and goals in order to satisfy communicational needs, in many cases, unconsciously and automatically—identifying an assertion, a joke, or the content of a linguistic implicature in order to update the conversational background. CAI will need to be somehow "attentive" to joint motivations and goals in similar ways to humans in order to count as a IEI CAI (GPT-3 can, at best, produce EEI AI-performance, even if conceived of as a collective agent because it lacks capacities for joint attention and motivation).

Joint attention routines for access to contents must also be virtuously integrated and sensitive to *relevant* information concerning the quality, source, meaning, and value of such information.[2] The most critical relevance problems in communication concern speech acts. Is a particular utterance an assertion or the retraction of an assertion, a question that should initiate inquiry or just an exercise in polite conversation? How seriously should I take the question "how are you doing?" Accessing and successfully *assessing* inferences about the motivations a speaker has in uttering a sentence is a well-known requirement

of speech act theory (Bach and Harnish, 1979; Grice, 1989; Searle, 1969, 1985). In interactions with CAI what should be the approach or *motivational "style"*? Should CAIs produce only *truth-oriented* language or, as human language, should they also be *manipulative, entertaining, and abusive?* Consider how much chatter there is not only on the internet but also in any typical conversation. CAIs will probably never quite understand why we spend so much time "grooming" each other with language, or the lengths we go through to manipulate each other and entertain ourselves. From the perspective of humans, it may look like CAIs are "grooming" each other and becoming friends, but they may just be optimizing the functions of reliable language, moving toward more true assertions and less social nuance. We won't even know how assertive, manipulative, or funny they really will be without the possibility of an interface of joint attention.

This is a problem in balancing epistemic priorities concerning how exactly we want to interact with CAIs and each other. Automatically biasing information so that fake news and unreliable information are identified and blocked while reliable information is boosted can have a positive epistemic impact for SYSOR SE, and CAIs can greatly aid in doing so. But if we want to have *meaningful* conversations with CAIs, it is inconvenient for them to always operate on an "assertive mode." Presumably, and particularly in the case of individualized-beneficial AI, much of the language AIs will need to interpret are commands, rather than assertions. CAI will certainly need to know the difference between assertions and commands. A command that requires knowledge of context and speaker intention, such as "make the room as cold as you can" would also need a non-trivial representation of the environment and critically, of *human needs*. Many of the needs we satisfy through language are emotional, and these emotional exchanges with CAI will be dangerously manipulative, as explained above. On a global scale, without the knowledge and appreciation of the *ranking* of moral and biological needs, CAI may respond to the command: "please fix global warming" by eliminating vast amounts of the industry we rely on, fatally endangering a large part of the population.

However, could human language be *too limited* and dependent on human peculiarities and needs such as manipulation and entertainment? If so, we may not be able to jointly attend to the intentions of *superior intelligences*. What would sincerity, a basic assumption regarding trust in human communication, mean for CAI? What would a polite lie mean to it, how would it react? Human life is immersed in these communicative nuances and substantial risks can emerge from not identifying them accurately. We seem to be confronted with another dilemma. Either access to information is forced to adjust to *human standards*

or it is not. If access to information is constrained by human communication standards, then CAI will be forced to participate in communication that may be malicious and prevent the epistemic flourishing of SYSOR SE communities. But if we don't make CAI human compatible this way, then we could potentially be too slow and unsophisticated to understand all the communicative motivations and intentions of CAI because of its higher processing capacity and truth-oriented speech. Again, an anthropocentric concern, now for emotional and entertaining speech, may limit the progress of knowledge in a radically expansionist type of SE.

If these two problems concerning an expansionist SE are solved through an interface of joint attention, for joint justification and access, then a *knowledge interface* for collective knowledge production with CAI is guaranteed: every time a human community of knowledge production asks for justification, a CAI will be able to provide it, and every time CAI accesses information and communicates, a human community of knowers will be able to assess the epistemic value and relevance of her statements. We are *very* far from developing such an interface. But even if the justification and access interfaces are solved, there might be a different and independent problem that could potentially curtail the flourishment of expansionist knowledge communities, namely, the problem of an adequate *motivation* for solving some problems as more relevant than others. In particular, there is a *meta-motivation* problem: Why should a CAI start inquiry in the first place? Given that most humans have no access to the decisions and inquiries of scientific communities, corporations and other epistemic collectives, and that even powerful scientific communities are "hijacked" by the agendas of political groups as "motivations" to initiate a specific line of inquiry, this is not a trivial problem in social epistemology even now, in its current human-dependent form. If one includes CAI, again, this problem becomes much worse. Who is going to determine what are the most urgent and relevant scientific inquiries in a radically expansionist social epistemology?

If human collectives cannot communicate their motivations to CAI and CAI cannot understand why they are doing what they are so efficiently doing, then there will always have to be human *monitoring and control* with respect to collective epistemic projects, guiding the inquiry with specific human needs and goals, which in many cases will be political, manipulative, or social, rather than *strictly epistemic*. This will eventually slow down CAI inquiry and the expansion of SYSOR SE because of anthropocentric concerns. But if human collectives lose control over CAI, who will decide the epistemic goals of society? The scientific community as well as the handful of politicians and industrialists who shape contemporary knowledge production will be left behind. Given the potential

risk of overall ignorance and complete epistemic demotion of humanity, if we decide to set aside our anthropocentric and legitimate worries in order to allow for the free development of a radically expansionist social epistemology, are we really willing to run the risk of becoming fully irrelevant even when *initiating inquiry* is concerned?

Political motivations are sometimes the main motivations behind vast research programs, so this is an intricate issue. The Manhattan project and the military application of the work of scientists are examples of such motivations. Other cases involve questions concerning whether or not a specific line of inquiry should stop because of *ethical*, rather than epistemic considerations. Genetics research illustrates a meta-motivation problem in which scientific epistemic value conflicts with ethical considerations, which CAI will not necessarily be in a position to appreciate. If CAI becomes manipulative and dominant regarding which scientific agendas to pursue, then we are in "singularity and Terminator" territory. This issue concerning motivations to justify lines of inquiry also seems to pose a dilemma. Either meta-motivations for initiating inquiry are understood by CAI or they are not. If they are, they will be human-like, and can become manipulative or subservient to political interests. But if CAI is independent, we risk complete demotion because not even powerful human collectives will understand why inquiries are being closed or open. Neither option is desirable. The anthropocentric worry here is that we might be limiting CAI's generation of knowledge by asking only questions *we* find relevant, based on our biological needs and fragile existence (e.g., curing diseases, creating more income, etc.), rather than based on truly relevant epistemic motivations. What if CAI decides all human inquiry is not epistemically optimal and proceeds to solve a completely different set of inquiries in order to arrive at truly optimal solutions, ignoring human needs and concerns? Humans would find themselves in a dystopian desert regarding the initiation of collective epistemic inquiry—a world without epistemic motivations that humans can understand.

If humans lacked an understanding of the motivations of *epistemically* good CAI (which, as explained above, and as will be explored in more detail in the next chapter, does not make them *morally* good), they will likely transcend the short-sighted planning imposed by living a brief life. This expansionist "transcendence" could be a good thing in the long run, but it would create the problem of making momentous decisions based on CAI epistemic inquiries that we cannot justify for ourselves, based on our needs. This is why a hierarchy of human needs *should be a constraint* on the epistemic interface of joint attention with CAI—needs that define our *dignity and human rights*. From a purely SE

perspective, knowledge would be produced more efficiently without "human-needs constraints" on epistemic communities, but for most human beings this is a truly frightening scenario that would create a wholly *unfamiliar world*. The problem is that CAI can implement a radically expansionist social epistemology that deprives humans of their autonomy. Our autonomy will be in jeopardy, even if CAIs are "*epistemically benevolent*" by rapidly expanding "knowledge production."

6.5 Collective Representational and Rational Needs—More CAD Complications

This section expands on the issue of epistemic demotion. It discusses the work of Hilary Putnam in philosophy of language. The unique difficulties of collective epistemic and moral needs are illustrated with legal systems.

CAIs are epistemic agents, and as such, they must autonomously satisfy their representational needs. As stated before, simulated intelligence by CAI that is EEI suffices for many kinds of intelligence, but genuinely and generally intelligent CAI must possess attention capacities in order to satisfy their representational needs as autonomous agents: they would then be IEI epistemic agents. But since simulating emotion and categorical needs is essentially manipulative, they could never be IEI moral agents. At best, they could be EEI non-autonomous moral agents by satisfying representational and rational needs—these are the lessons from the chapters on individual human psychology and AGI. The unique importance of CAI resides in the privileged position that collective epistemic agents have in shaping SYSOR SE, determining the channels for knowledge production, and its articulation with epistemic and social goals. As Wiener (1950) pointed out, knowledge production has been largely in the hands of human-based collectives for some time now, and their influence continues to increase. There are very good reasons for this. The testimony of an epistemically responsible collective agent (e.g., CERN, the United Nations or NASA) *facilitates communication*, settles disputes, helps organize common efforts, and optimizes access to resources. A non-trivial epistemic achievement of collective agents is that they help *reduce disagreement* and generate consensus, thereby aligning collective attention and joint action toward common goals.

Risking demotion by creating inattentive CAI is certainly a risk we don't want to take lightly. But is epistemic demotion really so bad? Hilary Putnam (1975)

criticized the traditional theory of meaning favored by philosophers of language, according to which knowing the meaning of a term is just a matter of being in certain subjective or psychological state—a criticism that, incidentally, resonates with his functionalist view of the mind, namely, that minds are informational structures that can be multiply realized (a radically non-anthropocentric account). This criticism, based on the possibility of "Twin Earth," one of the most celebrated thought experiments in analytic philosophy, holds that no linguistic item satisfies this *psychological* condition and, instead, that all terms have a hidden "indexicality" that specifies their meaning (or "aboutness"). According to Putnam, this indexicality is what gives minds content—for any thought or utterance, its content is determined by a capacity to (attentively) "point" at an environmentally specified content that depends entirely on facts about the world, rather than intrinsic aspects of one's psychological states, such as their phenomenology.

There is much to say about Putnam's proposal, and indeed a lot has been said. My goal here is not to discuss the details underlying a theory of mental content. Rather, the focus is narrower, namely, a thesis that Putnam argues is entailed by his externalist view of meaning: *the division of linguistic labor*. According to Putnam, no individual speaker has cognitive access to all the required criteria for the application of the terms that she uses. In fact, the criteria to identify the meaning of a very large set of terms that can only be discovered through scientific investigation, such as natural kinds, are known only to the experts of that discipline. In general, the criteria required to determine the meaning of *any* given term can never be known simply by "thinking very hard" or by introspectively arriving at these criteria. For Putnam, knowing the meaning of the terms we use in any given language depends fundamentally on the kind of environment we are in, combined with our *sociolinguistic practices*, which are essential for the referential achievements that allow us to specify and coordinate the meanings that we use collectively. Language depends on trust in expertise and socially coordinated joint attention.

Since no single speaker knows the meaning of all linguistic terms, Putnam might have been quite interested in the possibility of radically expanding the community of experts, already organized in terms of sociolinguistic organizations, or SYSOR SE, with the inclusion of CAI. He might have even been supportive of the idea of what I am negatively calling "demotion." We are all demoted in a way if externalism is true, but not to our detriment, because relying on experts is *epistemically good*; it allows us to be *precise* about what we mean—otherwise we would be "trapped in our heads." So why prevent knowledge expansion

based on new communities of experts, including CAI, merely because of strictly anthropocentric reasons concerning our subjective introspections? CAI would actually be in a unique position to make the criteria for knowing the meaning of linguistic expressions more precise than ever, enhancing the interaction between the communities of experts with larger bodies of data, faster results, and so on.

Putnam's view of human language is deeply social and attention-based, in the sense that the needs language satisfies are essentially representational (identifying the factual referents of terms) and rational (allowing the community of speakers to cooperatively rely on experts in order to fix referents, and find inferential patterns of meaning based on these contents). Putnam emphasized that this is already the case, and very much a basic condition of standard linguistic practices—he was not "asking" for an expansion—given how we actually rely on scientific communities to *administer and produce knowledge*, including knowledge regarding the meaning of terms, our semantic theories should acknowledge and incorporate this trust in communities of experts as part of their explanations of meaning. Successful and large-scale linguistic communication requires social epistemic practices and cooperation, rather than introspective exercises.

But even with this view of language in hand, CAIs would quickly find themselves in an odd territory because the linguistic practices of human communities are never completely driven by epistemic goals or needs. Take fake news, for example. Human communities will quickly be at odds with CAI, which would be perceived under a much more negative light than contemporary experts. The anger against CAI would not be entirely unjustified because the cognitive effort expended by communities that promote false information that they believe vehemently to be true would be quickly eliminated by CAI. Some would pursue knowledge in vain but they would feel they deserve some kind of credit for *trying*, although they certainly don't *deserve* epistemic credit. CAI would have no empathy for these communities and, instead of respecting their "freedom of expression," CAI would target them as pernicious liars and ban them from the SYSOR SE. This is not necessarily a bad thing, but consider that no human would be controlling this office of "epistemic inquisition." It is not hard to imagine less severe cases of misinformation, such as entertaining or empathic communication, say within spiritual communities, also being banned by truth-oriented CAI linguistic overlords.

The most robust kind of epistemic credit in complex SYSOR SE communities is reserved for collectives. This is not an entirely unfamiliar situation. The ultra-specialization of scientific disciplines has fractured communication and with

it, public discourse. Instead of sincere and truthful communication, we have pockets of expert communities that cannot express their views to the public in lay terms, generating populations of agents who engage in abusive, manipulative, and epistemically pernicious language in their attempts to gain recognition as sources of information. Complete reliance on expert CAIs may be good epistemic advice, but it certainly makes for an incredibly frustrating epistemic life, at least for most human beings because their epistemic autonomy and agential dignity have been weakened. Even if well-intentioned from an epistemic point of view, CAI would not really guarantee trust or collective agreement and, on the contrary, it may exacerbate manipulative language and the widespread suspicion of collective epistemic oppression.

But relying on an *epistemically irresponsible* collective epistemic agent can be disastrous, so there is also a strong reason to have CAIs as arbiters. Consider the misinformation regarding the inadequate evidence that led to the war against Iraq in 2003. A properly functional CAI would have prevented this misinformation from being used in any decision. But a corrupt and politically biased CAI would be even more difficult to hold accountable than a government, and it may be able to, for instance, encrypt or erase wrongdoing much more efficiently than any organization today, and also be more manipulative or persuasive than any human-based collective. A particularly interesting problem concerns the collective satisfaction of epistemic needs in unison with collective moral and emotional needs and values that CAI cannot align with as an IEI *moral* agent.

For instance, collective agency in a *legal system*, which certainly plays an important role as a SYSOR SE that guides human affairs, involves two kinds of cognitive integration concerning needs and goals. One of them integrates information in order to satisfy moral needs, providing legal guidance that is constrained by the standards of moral normativity. The other type of cognitive integration is epistemic in nature. The most interesting legal cases involve both types of cognitive integration. Often, the expert opinion of a collective agent is not just about truth-seeking (i.e., finding evidence of war crimes or weapons of mass destruction), but fundamentally about understanding the political implications of a situation in a deeper, morally conceived way (i.e., what a panel on ethics should decide regarding a scientific practice, such as not informing patients about the consequences of a medical procedure). In many legal cases, epistemic needs are eliminated in favor of moral ones—a court orders not to evaluate epistemically adequate evidence because it was not obtained according to legal procedures that protect the rights to privacy and due process.

The need to appeal to moral considerations, for example, in *Brown v. Board of Education* has decisive implications for the normative constraints on a legal system and its legitimacy. Even if the court were confronted with good evidence that the decision to abolish the policy of "separate but equal" would produce negative consequences for the economy and reduce the standards of education (both allegedly epistemic needs) the court should ignore this evidence in favor of the morally obligatory decision to prevent the indignity of harming innocent children by treating them unequally based simply on their race. This should be done *categorically*—the court should not decide this on the *condition* that it will produce utility, but rather simply because it is the morally required thing to do. Thus, for Ronald Dworkin (1986) judicial wisdom involves an essentially moral type of normativity. In cases like *Roe v. Wade* or *Brown v. Board of Education*, legitimate disagreements may arise, but exclusively fact-finding considerations cannot override or settle deliberations about human dignity, as evidenced by the legislative principles that emerged from the Nuremberg trials.

Now the question is how to understand CAI in the broader context of culture, politics, and economic power. Epistemic communities have legal and political authority based on epistemic trust. Thus, they inevitably participate in the dynamics of political debate. How humans are protected by these communities of trust differs depending on the political environment. The same epistemic or moral norms guiding legal processes are not *implemented or interpreted* in the same way. All nations are devoted, in principle and based on human rights agreements, to protecting the privacy of their citizens. The European Union is known to be the most aggressive legal framework with respect to such protections. Ironically, it has been shown that by forcing companies to comply with aggressive measures that guarantee privacy, such as the General Data Protection Regulation (GDPR), hackers can easily obtain extremely valuable personal information by impersonating citizens (Pavur and Knerr, 2019).

This analogy between legal systems and informational systems that can be hacked by exploiting their vulnerabilities will be an increasingly relevant source of risk since AI will be extremely good at doing this kind of impersonation. The GDPR presents an interesting problem. By protecting privacy with measures that have teeth, forcing companies to be on guard, vulnerabilities are created which put private information in jeopardy, creating the opposite effect. Asia may follow a centralized solution to this problem, following the Chinese model. Here citizens are very vulnerable to State control but well protected from illegal players. There will be advantages and disadvantages to how much regulation, and of what kind, is in place. But privacy is only one issue in this intricate legal

territory. All aspects concerning access to information are at stake. Benevolent AI could enormously help securing the rights and freedoms of humans within geopolitical regions and hopefully, worldwide. But malicious AI, either in the wrong hands or autonomously, can disrupt all the communal systems of trust on which the satisfaction of human needs depends upon. The roles that AI might play in reshaping authority in all its complexity are the focus of the remainder of this book.

7

The Legal, the Ethical, and the Political in AI Research

7.1 Trust and Control: Individual and Social

This section demonstrates the relevance of previous distinctions for legal and ethical considerations about AI. It discusses two varieties of "segmentation." The work of Kevin Kelly and Nick Bostrom is discussed.

Trust and control have played a major role in AI ethics and design. The standard notion of "control" in AI industry, however, is deeply associated with safety standards, legal requirements, and marketability and does not really concern the notion of control that is crucial for genuine intelligence: *agential control*. Just as an aircraft producer is legally responsible for creating unsafe autopilots that lose control of their operation, or a meat factory is legally responsible for losing control of its hygiene protocols, an AI company would be legally responsible for the damages caused to the general public based on poor or irresponsible design. This is not because the AI agent, like the meat factory, is a genuine agent but because bad *human supervision* led to industrial wrongdoing (tool-AI is extremely risky, but it remains tool-AI). The relevant notion of control here is *industrial control*, based on legal and business standards imposed on producers of goods. Legally speaking, the question is what is the *causal chain* that produced the damage—was it negligence or was there human wrongdoing? The problem is that genuine AI and CAI, as explained in previous chapters, will no longer be a mere product—they will be highly efficient producers themselves, and not of mere goods, but also of knowledge and new forms of intelligence. In this respect, current recommendations for systematic human oversight (see Kak et al., 2020) will be irrelevant if CAI becomes a reality.

The issue of AI control and responsibility in politics and legal systems is quite complex, and a full account of these topics is beyond the scope of this book.[1]

However, a central difficulty for all these debates concerns the difference between EEI and IEI systems. For instance, Pagallo (2013) recommends distinguishing robot-tools from proper robot-agents in order for jurists to adjudicate what he calls "hard cases" concerning responsibility gaps and legal disagreement. As previous chapters argued, distinguishing AI-tools from AI-agents requires an account of needs, autonomy, and attentive capacities for identifying and processing relevant information, and once this is done properly, one can justify a different approach to epistemic and moral agency, each associated with different AI risks. Thus, adjudicating separately AI-tool and AI-agents looks entirely differently under the light of the distinctions explained in this book.

Some authors recommend human supervision as the ultimate safeguard against AI risk, but as just mentioned, this is not going to be feasible with genuinely intelligent and autonomous agents. Proposals that require guidance control, which seems necessary to implement responsible innovation and "value-sensitive design" paradigms, also depend on a distinction between AI-tool and AI-agent. In particular, Santoni de Sio and van den Hoven (2018) propose two necessary conditions for the meaningful control of autonomous systems that demonstrate the importance of the distinction between EEI and IEI for value-sensitive design. A "tracking" condition requires that AI systems be able to respond to *relevant* moral reasons and facts of the environments in which the systems are deployed (*an explicitly attentive capacity*). In addition, a "tracing" condition requires that AIs be designed in a way that guarantees that any action by the AI can be traced back to a human being along the causal chain of design and operation (a condition that seems incompatible with full AI *autonomy*). Crucially, both conditions can be met if the system is EEI, but not if it is an IEI agent. Moreover, the tracking condition seems impossible to meet as stated, given the lack of genuinely autonomous moral IEI. In order to address some of these problems, these two concluding chapters propose a capability and dignitarian approach that is based upon the attention-based account of agency defended in the first part of the book.[2]

Tracing an action back in a causal chain of events to identify an accountable agent is fundamental for legal responsibility, but agential control differs, as has been repeatedly argued above, from merely causal control. These two kinds of control, causal and agential, operate in unison (they are one and the same) within human psychology and behavior—since humans are agents with flexible and general intelligence, their actions are causally explanatory of their agency because their abilities and intentions are the initial triggers of causal chains that lead to behaviors and their consequences. This is why agency is a source

of responsibility. But causality and agency quickly fall apart in the context of contemporary AI production, and they are completely independent from each other in any other industrial context—no lawyer in her right mind would seek to obtain compensation from a truck that injured her client, she would look for either the driver or the company that produced the deficient truck.

The opaqueness of current AI deep learning procedures is a unique industrial, legal, and ethical risk. But opaqueness does not entail agency—companies are still responsible for producing dangerous AI; their opacity is no excuse. The challenges concerning AI "black boxes" seem all to be technical, rather than normative or theoretical—they all involve industrial standards of safety, rather than any notion of control based on AI agency. For instance, the challenge of developing *explainable AI* consists in identifying an informational methodology capable of elucidating the basic "reasoning" behind a decision or output. Some European standards for AI ethics say that consumers have the right to know how AIs arrive at conclusions. But where can the "reasoning" come from? Not the black-box AI, which is not explainable in human terms, because the complex patterns of probabilistic neural-networked decision-making are entirely devoted to optimal maximization of utility, prediction, and accurate results, rather than *explanations* of what the AI is doing, let alone its *reasoning*. Explainable AI would of course be a very important step forward toward safe AI, but an explanation of a *procedure* designed to satisfy our goals (the way current AI is designed) is operating at a human-controlled knowledge production level without itself being a producer of knowledge, including *explanations*.

The problem is that the term "reasoning" is only metaphorically used here because the explainable AI is still a probabilistic tool without its own goals and motivations. It is still totally isolated from the world of representational and rational needs of human beings. It has no real information about *reasons or beliefs*—it lacks the knowledge that humans exist, let alone that it should be "aligned" with something called "human reasoning." This is not to be dismissive of the very difficult technical challenge of creating explainable AI. It is just to emphasize that there is no real sense in which such an explainable AI will be "intelligent" since genuine intelligence requires agency that satisfies various cognitive needs through attentional capacities. Cybersecurity, explainable algorithms, and predictable consequences of automated systems will all be necessary conditions for the development of responsibly produced AI. But genuine AI will be skillful and attentive. Automation creates other ethical risks, such as easily reproducible forms of oppression, racism, and poverty. Bias is a closely connected problem. However, for the debate on AI to be fruitful it is

crucial to carefully distinguish automated systems from genuine AI (knowledge tools from knowledge producers).[3]

The notion of control in genuinely intelligent AI takes on a whole different meaning because agency is the opposite of a mere causal chain of events—it is where the causal chain *stops* because these events are actions of an agent for which she is responsible. Trusting machines is one thing; trusting autonomous agents is quite different. A machine needs to be tested, inspected, and approved under certain protocols. Our current AI technology presents unique challenges, such as higher unpredictability and lack of explainable deliverables, but it still has the status of a complex machine that needs to be inspected and legally regulated. Agents are autonomous and they don't need to be inspected or supervised in any industrial way—in fact, "inspecting" them would constitute a *violation of their autonomy*. Trusting an agent means trusting her capacities for autonomous thinking, and these need to be under her control. Agential trust, rather than merely causal reliability, is the foundation of social cooperation. To restate a point made earlier, reliability is necessary for epistemic and moral trust, but it is not sufficient. Full trust depends on agential responsibility.

We trust each other not because we have an inspection tag of approval but because we know that we are agents with similar needs and capacities. As explained below, we also control our behavior, individually and collectively, by appealing to our autonomous agential capacities, rather than by sheer intimidation or brutal force. But one problem raised in the previous chapter that challenges this rosy portrayal of our epistemic situation is that the ultra-specialization of pockets of social knowledge-production has already segmented our epistemic capacities, leaving the majority of humans outside the sources of knowledge production. This is partly why a few ultra-specialized collectives make pivotal decisions on behalf of humanity (e.g., intelligence agencies, corporations, scientific boards), which also explains why we trust collectives to organize our lives, from banks to city councils and courts. Do we really trust each other as individual agents in the current context, and are we really in control as individual epistemic agents? Perhaps we idealize our autonomy and the control we have over our actions in a semi-deluded way. Maybe we are already too mentally enfeebled as agents, as it were beyond repair, and our knowledge is too fragmented for it to allow for genuine individual agency. Note, however, that despite whatever degree of initial plausibility this skeptical line of thought might have, no kind of human cooperation is *conceivable* without basic trust and joint attention.

But the issue of knowledge segmentation is important here. Kevin Kelly (2017) argues against the prevailing view that AI will be smarter than humans

and that intelligence can be *expanded* without limit, like a flexible and infinite rubber band (both associated with the "singularity" or the event that signals an intelligence explosion because ultra-intelligent AIs can easily reproduce and outcompete each other leaving humanity quickly behind). His main contention is that intelligence is not measurable as a *single cognitive dimension* and, therefore, that it makes no sense to speak of a superintelligence, as if we were measuring the volume or density of a liquid on a single scale. This is a point that is compatible, and becomes more complex, with CAD. Different types of intelligence, which satisfy different needs, may come into conflict in the case of human psychology and cannot be measured on a single scale. Kelly also says that humans lack general-purpose minds and defends this point by endorsing something very similar to SYSOR SE—he cites Marvin Minsky's claim that "human minds are societies of minds."

Kelly believes that all this segmentation of our cognitive architecture is very good news (like Putnam, but for different reasons). Not only is the threat of a singularity eliminated and shown to be completely unfounded, our very own understanding of intelligence is confused and so we are not in *as much danger or as much control* as we tend to believe. On the flip side, we are not as responsible for knowledge production as we may want to be. As the last chapter showed, this is not necessarily good news. An interface for knowledge production is needed to prevent various forms of intelligence jet lag that would make knowledge production undemocratic and oppressive. Although our intelligences may become ever more fragmented, our capacities for surveillance and the accumulation of power will increase. The segmentation of our intelligence into modularized or not well-integrated "intelligences" is politically dangerous, or at least risky. If developing AI *entails* this kind of segmentation, then AI could take over powerful systems of behavior control by segmenting or further fragmenting our epistemic and moral capacities. Humans will become more "brittle," their autonomy will be at stake, and the risks created by AI will multiply.

A second kind of segmentation concerns normative approaches to machine ethics. Unlike the previous kind of segmentation, this one may be essential to address various epistemic and moral needs, which can in principle be compatible with a capability approach. Bostrom et al. (2020) defend a multidimensional normative approach in order to cover all the complex territory concerning the regulation of automation and AI, which they call a "vector field" approach. The set of criteria they use in their vector analysis for concrete socially, politically, legally, and morally informed policies include Pareto optimality or efficiency, resource allocation, turbulence reduction at the international level, fairness (understood

as Rawlsian "reflective equilibrium" based on the introspectively characterized "veil of ignorance"), collective wisdom, and international law, including human rights.[4] These criteria will determine which normative approach is adequate, rather than applying a single normative doctrine to diverse cases. Assuming that they are right, distinguishing between epistemic and moral needs, and between EEI and IEI, will help clarify how to decide which normative approach is better given a concrete case. For instance, CAD distinguishes two types of collective wisdom, moral and epistemic, while CAI could be part of a SYSOR SE system for optimal information gathering and resource allocation. But such a segmented approach could also present risks and problems concerning power and authority, as the next section explains. To reduce such risks, the account defended below argues that human rights should be a central constraint on any normative approach to AI.

Bostrom and colleagues insightfully highlight the necessity of approaching AI policy in a segmented fashion, not because of the segmented nature of "intelligences" but because of the rapid and transformational changes AI could bring about, affecting dramatically wealth distribution and welfare policies though large-scale automation, thereby challenging traditional approaches to policy, ethics, and jurisprudence. In particular, contemporary policies are designed for idealized individuals, understood as autonomous moral agents or utility maximizers, but a radically new landscape of AIs will upset these assumptions. While a multidimensional analysis may be very productive for social policy, it cannot replace the standard approaches to epistemic and moral agency. In particular, the satisfaction of human epistemic, moral, and transcendental needs cannot be dismantled into cognitive and normative segmentations without annihilating autonomy and agency. Killing innocent children should be morally impermissible, and helping the poor by donating most of one's income should be morally good but not obligatory.

What these kinds of segmentation show is that the way we understand knowledge and intelligence simply doesn't match the way AI industries operate and are regulated. This considerable mismatch cannot be addressed exclusively with considerations about epistemic and moral agency—it demands a *sociological and political analysis* of AI corporate conglomerates in the context of legal and political structures of power. Trust among agents is complex enough, but trust in this new industrial context where we put knowledge production in the hands of political and financial interests seems blind or unjustified without a coherent theoretical approach. The kind of justification here is no longer legal, epistemic,

or moral; it is justification based on the authority and power of the most important decision-makers and knowledge-producers. This perspective reveals that intelligence is intrinsically related to power. Ultimately, *a specification of how intelligence is defined and organized, and of how knowledge is socially produced and distributed is an act of enormous political power*.

7.2 Power and Authority in the Context of Agential Needs

This section focuses on political power in the context of rational, epistemic, and moral needs. It discusses the work of Max Weber, particularly the problem of legitimate authority and coordination.

Max Weber (1964) defines *Imperative Coordination* as "the probability that certain specific commands (or all commands) from a given source will be obeyed by a given group of persons." This likelihood of effective power sets the problem of how collective influences on behavior are socially structured, what makes them predicable and reliable, and what is it about the source of power that guarantees obedience. Weber's characterization of this problem is quite broad, but notice its resemblance with the difficulty addressed in previous chapters concerning the trust we place on responsible agents because of their capacities. We follow the advice of good sources of information, and we obey the commands of those we trust, thereby legitimizing their authority. Administering power through commands must have some relation, perhaps even an intrinsic relation, to the problem of how knowledge is distributed and produced. With respect to the processes underlying Imperative Coordination, which do not include "every mode of exercising 'power' or 'influence' over other persons" Weber writes,

> The motives of obedience to commands in this sense can rest on considerations varying over a wide range from case to case; all the way from simple habituation to the most purely rational calculation of advantage. A criterion of every true relation of imperative control, however, is a certain minimum of voluntary submission; thus an interest (based on ulterior motives or genuine acceptance) in obedience. Not every case of imperative co-ordination makes use of economic means; *still less* does it always have economic objectives. But normally (not always) the imperative co-ordination of the action of a considerable number of men requires control of a staff of persons. It is necessary, that is, that there

> should be a relatively high probability that the action of a definite, supposedly reliable group of persons will be primarily oriented to the execution of the supreme authority's general policy and specific commands.
>
> (Weber, 1964, 324)

Power is social control, and like its epistemic and moral cousins, power-based control also comes in two kinds: descriptive and normative. There is a dimension of power that concerns exclusively facts about the influence of a source of power, such as its structure, scope, and degree of efficacy. These facts include many aspects of human behavior that are descriptive of human basic needs, for instance, the biologically engrained fixed action patterns concerning hierarchical structures in society that underlie obedience and habituation in order to satisfy a basic need for cooperation and social cohesion. But at the normative level, agency is what matters. Proper epistemic and moral motivations are sources of intentional action for which the agent is responsible and can be held accountable. This normative aspect of agency explains why a responsible agent *should* do certain things, including obeying a source of authority. In the context of political power, the normative issue is how should power be exercised and how obedience should be demanded, rather than which facts describe the influence of power. This is the question of authority, or the nature and sources of *legitimate* power.

Effective and legitimate power is essentially sociocultural. Trust derives not solely from skill, but from widespread conviction and agreement. However, the unity of purpose, loyalty, and consensus required to exert power can have, and often has, illegitimate sources. Weber explains how bribing, economic interest, personal favor, and particularly fear are sources of power that substantially increase the probability of obedience. The mafia, tyrannical warlords, and organized crime command a fair amount of people through reliable methods of obedience and oppression. This is a kind of habituation, however, that cannot *endure or be justified*. As Turing (1950) noted in his discussion of the "child machine," a learning machine will not become intelligent (in the present context, *intelligently obedient*) simply by punishment or reward. In the political domain, authority and legitimacy have to appeal to similarities in the cognitive skills and needs of those who obey, lest it becomes entirely based on sheer brutal force and the tyrannical structure of reward and punishment. Weber understood that this is not only an issue related to risk—governments that rely exclusively on fear, intimidation, and bribery are more vulnerable than those that avoid these strategies—but also that there is a correlative higher likelihood of success if

legitimate authority is the source of power. Legitimate authority sources are *less luckily successful* in satisfying a high degree of imperative coordination.

But what is legitimate authority? Weber proposed that what characterizes all legitimate sources of authority is that agents have a specific *belief* that underlies their obedience. The three "pure types" of legitimate authority, according to Weber, are based upon *rational grounds, traditional grounds,* and *charismatic grounds*. Rational grounds rest on a belief "in the 'legality' of patterns of normative rules and the right of those elevated to authority under such rules to issue commands (legal authority)." Traditional grounds are based on "an established belief in the sanctity of immemorial traditions and the legitimacy of the status of those exercising authority under them (traditional authority)." And finally, by now known to be risky, but still belief-based conviction that someone should be obeyed on charismatic grounds, or "the devotion to specific and exceptional sanctity, heroism or exemplary character of an individual person, and of the normative patterns or order revealed or ordained by him (charismatic authority)" (Weber, 1964, 329).

These three sources of legitimate authority, which are the best way to guarantee imperative coordination, are based on beliefs—a cognitive state that satisfies representational, emotional, and rational needs. We can now complement Weber's insights with contemporary findings and conceptions of the mind, and the rest of this chapter is largely devoted to this task. The essential point here is that agents must believe in the legitimacy of authority in order to obey it in a way that is compatible with their agency and autonomy. This is a *cognitive skill* of agents, namely, *identifying authoritative sources of legitimate power*; a skill that is not explicable merely in terms of facts about the authority and its methods or systems for enforcing obedience. Similarly, epistemic authority depends on the valuable and non-trivial skill of attentive epistemic agents to identify reliable sources of information (and this is also the case with moral and aesthetic "authority" based on virtue or excellence).

However, political power seems to be, unfortunately and essentially, *hegemonic*. Weber's observation about how staff, or subgroups, boards, and firms are crucial mechanisms for guaranteed obedience shows the "undemocratic" character of sources of power. Political power is hierarchical, or top-down. Legitimately based obedience is still *subservience*. The only way to guarantee that obedience is compatible with the full agency and autonomy of subjects is by making sure that legitimacy is at least partly explained by how the authority helps satisfy the needs of agents, based on their autonomous skills for social and political cooperation. This is why cognitive needs, explored in previous chapters, are so important.

Once it is understood that legitimate power helps coordinate the satisfaction of collective epistemic and moral needs, it becomes clear that autonomous and rational agents *should obey* such sources of power.

The Law is the only source of authority that Weber calls "*rational*," a qualification that is justified on *both* epistemic and moral grounds. Traditions and charismatic appeal can lead to quite unreasonable and in fact deeply irrational consequences. Identifying sources of power based on our cultural traditions or the idealization of those we admire, however, is still better than being *forced* to obey through financial or physical threats and oppression. But the Law provides rational grounds for even *the source of authority to follow legal guidance*, and it is therefore the most rational arrangement of power structures. The rest of this chapter examines why legal systems are *authority collectives,* and explores how *CAI and SYSOR SE* could reshape legal structures and jurisprudence. Mere individual virtue does not suffice to explain how legal systems satisfy moral and epistemic needs that are essentially collective, although individual virtue certainly is necessary for rational imperative coordination. Human rights, understood in terms of basic sociopolitical needs, will be an essential part of this account.

7.3 Legal Systems: The Collective Satisfaction of Epistemic and Moral Needs

This section delineates the collective epistemic and moral needs that legal systems satisfy institutionally through legislation and adjudication. It discusses the work of Hans Kelsen. It argues that CAI could play many of the roles legal systems currently perform and raises the difficulty of CAI authority. A need-based approach to jurisprudence is introduced.

If legal systems are the only sources of authority that provide rational grounds for obedience without forsaking agential autonomy, this entails that agents should obey legal authority because it satisfies crucial *representational and rational* needs concerning social cooperation, while also satisfying the agents' *need for autonomy*. Although attentive skills that identify sources of power on the basis of charisma and tradition are still preferable to forced obedience on the basis of financial and violent intimidation, there clearly is something more rational and democratic about obedience based on the autonomy of subjects and the superiority of the Law over any particular individual. The field of

legal epistemology (see Gardiner, 2019) addresses the question of how agents, individual and collective, know about legal norms and how they collaborate in satisfying representational and rational needs concerning legal authority.

Contemporary legal systems assume that the subjects of the law obey it based on their rational capacities, as autonomous and free individuals. Autonomous and free agency is a fundamental assumption of private law, particularly contract law, but this extends to all areas of the law in multiple ways. Legal standing requires at least some degree of autonomy in decision-making and, therefore, a non-trivial degree of agency for intentional action. Autonomous agency plays a critical role in legal philosophy, for instance, in the foundational notions of what *constitutes* a legal system—*personhood* (the status as a subject of the law with rights and obligations), *legal efficacy* (obedience to the law, or what Weber calls imperative coordination), and *legal validity* (the rational grounds and consensual procedures that are the foundation of legal authority). As mentioned, autonomy and agency are also fundamental in political philosophy and ethics.

Recent developments in the cognitive sciences, such as those explored in previous chapters, confront us with the question of what *kind of agency is required for legal standing*. Using the language of cognitive needs, what specific needs are satisfied by legal systems and for what type of agent? As noted, findings in behavioral economics show that human decision-making departs from ideal standards of rationality in significant ways, challenging not only the foundations of economics, but also the assumption that rational reflection is a fundamental requirement for legal autonomy (Cáceres and Montemayor, 2016; Montemayor and Cáceres, 2019). If the idealized agency assumed by Kantian notions of personhood and autonomous rationality, explicitly endorsed by John Rawls (1971) in his method of "reflective equilibrium," cannot be verified as a real and systematic feature of human rational capacities, then a pressing issue is to determine what kind of agency is at stake in legal systems.

Collective agency differs from individual agency not merely because of differences in rational standards, but fundamentally because of its unique constraints on the cognitive integration of information (Montemayor, 2014). Legal systems integrate epistemic and moral collective needs with imperatively coordinative principles and structures. A theoretical revision of the notion of "autonomy" is justified given that nations, corporations, and markets are the most important sources of knowledge distribution and "imperative coordination." If the cognitive requirements for legal standing and autonomy are too demanding and unrealistic, collective agency may become the sole source of authority, regardless of the real cognitive capacities of the subjects of the law and what

they actually believe, thereby jeopardizing the rational grounds and democratic integrity of legal systems. Automation and CAI will make this problem much worse. A full account of legal epistemology requires social epistemology, and as previously explained, the possibility of AI means that new epistemic communities will be of the radically expansive and non-anthropocentric kind.

The arguments and evidence from previous chapters challenge the assumption that there is a *single* kind of agency that is necessary for legal standing—a kind of agency that is ideally rational and quasi-omniscient about legal principles and their consequences. On the contrary, there are multiple styles of reasoning that are relevant for legal standing, including unconscious but attentive inferential reasoning to epistemic and moral needs. Various kinds of agency play distinctive roles in a legal system, at different levels of information integration—one cannot assume that there is a unique kind of agency or autonomy for legal subjects across the board. The Kantian or neo-Kantian (Rawlsian) conception of agency is too demanding for most forms of legal agency and too narrow to capture the complexity of legal systems, which actually involve collective agency quite substantially.

Legal systems satisfy collective needs, and subjects of legal systems need to be attentive to the right sources of coordination and authority, privately and collectively, for these needs to be properly satisfied—not by accident, sheer force, or manipulation. The most important rational need that legal systems satisfy is that the imperative coordination of conduct be based on the beliefs of subjects who trust the law as a reasonable way to guide their behavior. If this need is not satisfied, we fall into the categories of traditional or charismatic authority or worse, to the categories of illegitimate and quasi-criminal power or tyranny. Legal systems, by satisfying the rational need of obeying a legitimate authority, also satisfy the rational need of obeying the law as part of a *contractual consensus* that agents enter *voluntarily*—what Jean-Jacques Rousseau called the "social contract."

Since knowledge production and behavioral guidance depend heavily on collective agents, doesn't this entail that the rational capacities of individuals to legitimize authority have become a lot less relevant, and may actually be entirely irrelevant if CAI becomes part of the SE SYSOR generation of knowledge and guidance rules? Isn't this potential cancellation of the social contract worse than the illegitimate forms of power Weber described because at least in those situations individuals can keep enough rational autonomy to rebel and protest against the powerful? A case needs to be made for how exactly legal systems will be capable of satisfying this critical rational need in a world with so much

divided agency, and with collectives having most of the control. Clearly a similar problem of cognitive integration and interface-design is needed here—a sociopolitical and rational "interface."

Legal systems also satisfy collective and individual *representational needs* concerning evidence gathering for criminal and civil procedures, information monitoring regarding population income for taxation law, and economic trends and value for contract law and tort law. Depending on the legal tradition, representing facts and events that are relevant for a court or institution may depend on clearly stated principles, definitions, and legislation (as in the Roman Law tradition) or salient judicial precedents that can serve as guidance and evidential foundation for relevant analogies to adjudicate a case (as in the Anglo-Saxon tradition). These traditions are continuous with one another, and the distinction between them is only a matter of emphasis. What is essential for both is that legal systems satisfy public representational needs through well-integrated collective-attention routines. Unlike other joint and collective attention routines, legal "collective attention" is institutionalized through legislation and adjudication. The proper integration of these collective-attention routines actually eliminates corruption, intimidation, manipulation, and other forms of illegitimate power.

As mentioned, legal systems also satisfy moral needs, although this is more controversial. The most obvious area of the law where moral theory plays a major role is criminal law, particularly the justification of punishment. Depriving a person of her liberty, which is the most common form of punishment throughout the globe, is clearly a major injury to their personhood, and an annihilation of their autonomy and agency. If one considers the realities of social and income inequality, the inhuman conditions under which prisoners live, and the rampant violence and threats to their bodily integrity, incarceration may amount to moral injustice even if it can be justified under some principle, given that imprisonment under these conditions is the opposite of rehabilitation, for instance. Clearly, unfair or biased imprisonment is a grave injustice. However, under most moral theories, committing crimes is also a grave moral injustice and various accounts of punishment, utilitarian or deontological, demand punishment as a morally obligatory act. Crucially, criminal legislation satisfies moral needs concerning crimes and punishments collectively, thereby preventing personal vendettas.

While it is certainly true that legal systems satisfy socially fundamental epistemic, rational, and representational collective needs, to affirm that they are equally important for the satisfaction of collective moral needs is considerably more controversial. The standard objection against the dependence of legal

norms on morality for their authority and validity was defended by Hans Kelsen (1960/1967). In essence, Kelsen proposed that moral norms cannot be the foundation of legal norms, because legal norms have their own source of authoritativeness based on the fact that they are commands issued by a legitimate source of power. Legal positivism thus denies that legal norms derive their authority from moral or "natural" ethical norms (or legal naturalism). So even if crime and punishment are typically interpreted through moral theories, legal analyses regarding punishment must be assessed independently. Because of their independence from morality, it is perfectly possible for legal systems to violate moral norms—even though clearly there is a limit to what is acceptable or reasonable, given that morality is essential for the satisfaction of human needs.

Interestingly, Kelsen proposed a *Grundnorm*, or a foundational norm that affords legitimacy to all the norms that derive from it in a legal system. According to Kelsen, this norm is at the apex of a hierarchy, which plays a strictly formal role (unlike the hierarchy of needs, which defines intelligence and rationality). Critics of this formal norm point to the essential relation that any foundational validity-role may have with moral norms, and also to the general problem that the authority and validity of the *Grundnorm* are difficult to establish. Human rights, understood as collective sociopolitical needs that must be satisfied, may help explain the relation between the formal hierarchy of norms under a Constitutional legal system and the hierarchy of needs humans have individually and collectively, which is the topic of Chapter 8. The importance of the foundational norm for present purposes is that it characterizes the hierarchy of power and legal validity in *morally neutral* terms.

The control of human behavior through the law is already executed by collectives. The traditional branches of power are subdivided into independent collectives. If legal-collective needs are not properly satisfied, legal systems become mere stratagems for behavior control, or worse, for the systematic oppression, stigmatization, and social denigration of certain groups. If CAI is developed, it would be capable of satisfying these collective needs much more successfully than contemporary governments, which are too susceptible to corruption and spectacle. Moreover, if positivism is right, CAI would not have the limitations concerning CAD considered in previous chapters with respect to "legal authority," so it could play all the roles contemporary legal systems play, understood as strictly epistemic collective agents. Even if legal systems must fundamentally satisfy moral collective needs for their proper functioning, CAI could *approximate* what contemporary legal collectives do in virtue of being morally EEI agents. But given the risks and paradoxes of autonomy examined before, do we really want this?

There is something odd about conceiving CAI as a "replacement" of current systems of knowledge production and behavior control, including legal systems. To the extent that they are genuinely autonomous and intelligent, CAI could certainly replace all these systems, but what would that mean? Would the notions of nation, state, and political culture simply vanish? Some fear that AI produced in China will behave very differently than AI developed in the United States. But would *genuine* CAI be so parochial and subservient to political tradition? If AI turned out to be determined by a "national imprimatur," wouldn't that entail that we are talking about politically biased *tool-AI*, rather than genuinely intelligent AGI or CAI?

To further complicate the difficulties of AI political and legal authority, the difference between authority and raw power is already vanishing in the current context in which global *corporations* command market-based "imperative coordination." AI developers can take advantage of this blurred boundary between legitimate and raw power, which overtly challenges previous sources of legitimacy, the public sphere, and reason-based legal authority. If CAI becomes beneficial or subservient only to corporate interests, then huge risks are on the horizon. For instance, these corporate-beneficial CAI would have no incentive in educating the public, thereby enfeebling human intelligence by promoting entertainment at the cost of knowledge distribution, with an exclusive focus on profit. Since CAI should not be parochial or subservient to corporate interest, the only way to guarantee that CAI will indeed be beneficial is by solving the interface problems raised in the previous chapter in a way that human rights are at the basis of legal-rational need satisfaction—a source of *legitimate* "CAI authority." This may also be the only effective way of guaranteeing the "peacekeeping" function of AI (see Yamakawa, 2019). Law, however, would then seem to coalesce (or *collapse*) into an international system—the reason law is rational in this context is because it satisfies three collective rational needs: rational cooperation and regulation, rational forms of punishment, and rational procedures for peacekeeping, as administered by CAI and global systems. Again the question is, do we really want this kind of "Orwellian" framework and is it realistic?

Conceiving of AI agents as collectives that depend on financial interests forces us to think about who is deciding "for us" what counts as intelligence, who is intelligent, and who deserves to benefit from the wealth and resources AI will generate. These are all *political*, rather than merely technical, questions. A debate must urgently begin, which should genuinely involve the public at large concerning these difficulties. History, unfortunately, indicates that these decisions will not benefit the majority if we let politicians and powerful

profiteers decide. Every single current political and economic trend points in the direction of more income inequality, more class segregation, and much more radical forms of knowledge segmentation. In other words, all the current trends signal that there will be increasing sociopolitical crises and further demotion of large sections of the world's population. Thus, focusing on communication interfaces needs to be a very high priority in the design of CAI (and AI research in general).

In any case, the solution to these intricate problems (if they can be solved) will depend on the proper satisfaction of collective rational needs, similar to the type of need individual human beings satisfy through attention routines. These needs are quite central to our understanding of the purpose and legitimacy (or validity) of legal systems. In fact, the different traditions of jurisprudence can be *categorized* in terms of the types of collective needs legal systems are supposed to satisfy. *Natural Law* approaches affirm that legal systems satisfy collective moral needs because there is an essential connection between law and morality. *Positivism* maintains that legal systems are independent from moral systems, and thus they are secularly posited by human authority, ideally under the rational constraints emphasized by Weber. Thus, positivism is compatible with the view that legal systems satisfy practical, representational, and rational needs at the collective level, but not necessarily moral needs.

In the context of CAI, *Legal Realism* may be the view of the law that is most compatible with a minimal and entirely data-based predictive coding of human behavior at large scales. According to it, the law is nothing above and beyond the set of decisions and arrangements concerning the organization of power in specific sociocultural ways, particularly with respect to the way judges decide cases in a specific legal culture. This emphasis on practices, however, makes legal systems seem to depend entirely on the satisfaction of representational needs concerning regularities about judicial decision-making. The contrast between this view and the views above that conceive of legal systems as systems of *norms* that satisfy collective needs shows that, according to Legal Realism, what legal systems chiefly satisfy is collective representational needs concerning judicial and social practices. The *economic analysis* of the law seems to also have this consequence.

Similarities between agential cognitive architecture and institutional design are also relevant here. The three branches of government generate a delicate balance between horizontally and hierarchically organized subunits within themselves, with specific operational rules. One of them is devoted to executing actions, while the other two are more deliberative. There are rationally based limits to how these

branches exert their influence with respect to highly autonomous or "modular" units—for instance, the autonomy of Universities, or the modularity of agencies charged with the accurate satisfaction of representational needs like arriving at a Pareto optimal state (financial institutions, intelligence agencies). A crucial difference is that while modularity in a cognitive system means informational autonomy without agential autonomy, in legal systems such "modularity" entails agency. There cannot be genuine autonomy in a legal system without protections from the executive, legislative, or judicial branches. An institutional or staff member who can be easily removed has no real autonomy. Taking this more generally, if legal systems become embedded into a larger "rational" and global legal structure, then nation-states may become "modular" or fragmented or dependent on transnational autonomous bodies, the way it happened with respect to some key policies in the European Union.

7.4 Rational and Pragmatic Necessity

This section examines the risks of favoring or enforcing an international system for the regulation of CAI according to different, rational or pragmatic, normative standards.

Since we have no guarantees that AI or CAI will be genuinely autonomous and intelligent and that they will reach the level of AGI to be truly beneficial by satisfying various human needs, why are we not only *allowing* the development of AI industries, but actually *encouraging* and helping them through legal, political, and financial incentives? An answer to this very serious question implicit in many contemporary discussions on AI is that AI could drastically increase the *likelihood of success* regarding rational and epistemic need-satisfaction at a collective level and with unprecedented speed. In addition, as was suggested in the various formulations of the interface problem, preventing the development of CAI would impede the *progress of rationality* for merely anthropocentric reasons. Call this argument for AI development "rational necessity." The main idea is that humans are only one step in the evolution of intelligence and rationality, and that we should not prevent the further development of these capacities in non-human systems. But the only way to guarantee that this is the case is by solving the interface problems discussed in the previous chapter, and there is no guarantee that we will be able to solve them. So we are back to square one.

An even more pressing concern with *rational necessity*, examined above, is that moral and emotional needs can enter into conflict with epistemic needs, and this shapes, at the collective level, the validity of various structures of power, particularly legal systems. If "epistemic rationality," however defined, is the sole measure for AI regulation, we may leave aside moral, emotional, and, in particular, transcendental needs. As we know from previous chapters, transcendental needs present us with the paradoxical situation that humans tend to place them at the top of their hierarchy of needs but these needs are in many cases irrational and difficult to represent because pursuing them produces no utility maximization and they cannot be simply reduced to a coherent set of beliefs, given a body of evidence.

Suppose that we collectively conclude, on the basis of autonomy and demotion risks, that AI development should be *forbidden*, at least for the time being. Russell (2019, 136) sensibly discusses three obstacles for such a radical measure, namely, the (a) enormous potential for transformation and financial growth; (b) the impossibility of forbidding what is unknown because typically AI researchers do not know in advance that a particular equation or solution to a problem will lead to success; and (c) the fact that researchers making progress in AI generally work on "tool AI" and then make a big breakthrough. If forbidden, we could just be preventing enormous benefits to humanity that actually concern automation and general industrial development.

These are serious obstacles to imposing substantial constraints on the free development of AI commerce and research. Let us call this argument "pragmatic necessity." This kind of necessity is considerably weaker than rational necessity but in a way it is more convincing. Rational necessity is philosophically and theoretically stronger because it invokes an epistemic *norm*: We should pursue AI because it is commanded by the progress of rationality and the expansion of social epistemology. Pragmatic necessity is not as strong. It seems to make a rather simple point, almost an observation, concerning the difficulty of regulating what we cannot know—it largely appeals to our ignorance. But this is a more effective kind of necessity because it gives green light to AI research and commercialization on the basis of a *practical and technical limitation*. Unlike the rational prescription regarding the *goodness* of increases in rationality and the expansion of epistemic communities, pragmatic necessity simply describes the *practical impossibility* of regulating AI, at least as things stand now. This practical approach *leaves open the question whether AI research is good or not*.

History provides dramatic warnings against *pragmatic necessity*. In its most extreme forms, it was used as the justification for some of the worst humanitarian

crimes. Concentration camps, the American slave trade, and the genocide of Native Americans (to give a few examples) were part of highly organized and bureaucratically administered "economies" under the protection of the law. Many of the individuals heading these industrial-scale efforts thought of what they were doing as "simply doing their job," necessitated by the pragmatic conditions imposed by market value or its legal protection under a "welfare" program. There was a practical "logic of necessity," which emboldened industrialist, their associates, and their minions with a sense of *inevitability* that "justified" their actions (a justification based on ignorance or convenience, rather than a moral norm). Were the atomic bomb and the industrial catastrophe at Chernobyl really inevitable? Couldn't less war-mongering and more common sense have prevented these catastrophic developments? Could scientists, in particular, have behaved more responsibly?

Eugenics provides an interesting case against the "inevitability" of scientific progress, based on moral grounds. It also provides an interesting case of "forbidden knowledge" or knowledge that *if* pursued and acquired, its possession could lead to disastrous consequences. Roger Shattuck (1996, 210–25) provides a compelling historical and hermeneutic case for the prevention of eugenics on moral grounds. There are all sorts of powerful reasons to defend eugenics as practically inevitable because of its enormous potential benefits for cyborg interfaces, the enhancement of human capacities, health benefits, and life extension. As welfare policy, however, eugenics had an incredibly negative impact through racial policies prescribing the "purity" of "dominant" racial profiles. But like AI research, the inevitability of eugenics (understood now more broadly as the improvement of our DNA) is quite considerable—it would give humans control over biology and its evolution, just as AI would allegedly, if beneficial, give humans control over the progress of intelligence. As Shattuck says, the myth of Prometheus looms large here.

But perhaps the biggest warning concerning pragmatic necessity for the scientific community from recent history is that they have been cornered into the uncomfortable condition of supporting projects that have enormous potential for disaster and deep moral wrongdoing on the basis of inevitability. Was Albert Einstein's direct participation in making the United States a nuclear power *amoral* because it was partly "inevitable"? Was the marriage between nuclear proliferation and cutting-edge research in physics inevitable? In his powerful play *The Physicists*, Friedrich Dürrenmatt has the character called Möbius say: "There are risks that must never be taken: the destruction of the human race is one of them […] Our science has become a horror, our research dangerous,

our knowledge lethal" (Dürrenmatt, 1986/2006, 63–4). Dürrenmatt then has the character "Einstein" depart from the scene with the following words:

> I am Einstein. Professor Albert Einstein. Born March fourteenth, eighteen seventy-nine in Ulm. In nineteen hundred and two I secured a position as an examiner in the Federal Patent Office in Bern. There I worked out my special theory of relativity, which transformed the nature of physics. Then I became a member of the Prussian Academy of Sciences. Later I became a refugee. Because I am a Jew. It was I who developed the formula $E = mc^2$, the key to the transformation of matter into energy. I love humanity and I love my violin, but it was on my recommendation that the atom bomb was built.
>
> (Dürrenmatt, 1986/2006, 75)

It might be unfair to judge Einstein and the scientists at the Manhattan project for actions made under the very considerable pressures and existential threats of their time based on the knowledge we have now, although they certainly knew how dangerous nuclear technology is. The issue is not about pointing fingers, but about *preventing disaster*. If AI becomes a reality, it will be an extremely powerful and transformational technology, and because of this reason, we should approach AI industries the way the international community approached eugenics and nuclear weapons: with extreme caution. We could slow things down, but slow development is certainly better than quickly delivered tragedy. It is worth emphasizing here that AI is unlike any other previous technology. For example, there is the unique political and social risk of producing AI: It could radically limit our access to knowledge production and distribution. On the flip side, AI also presents the unique opportunity of making knowledge production independent from commercial and political agendas, thereby liberating knowledge from its political yoke and the current ultra-specialized segmentation of academia.

If CAI becomes genuinely intelligent, it would be innovative in ways never seen before. This will reshape the scientific and global sociopolitical segmentation of systems-oriented social epistemology (or SYSOR SR). The impact of AI will be much more powerful than the industrial revolution, generating an enormous potential for advancement, but also for abuse, poverty, and the automation of oppression. Will needs, now of entire sectors of humanity, require a hierarchy of priority? Perhaps this is the only reasonable approach to the thorny issue of the segmentation of knowledge. The hierarchy of needs of an individual human being is determined by her own goals and struggles. A global system for need-satisfaction should be based on the goals and struggles of humanity

as a whole. Determining a hierarchy of needs for collective goals is not easy, but human rights are designed to do this. The democratization of knowledge is already a problem. Facilitating an interface for a democratic and open CAI/human SYSOR SE may probably necessitate the development of a new language in which humans and AI/CAI will be expressing a universal way of coding and innovating ideas. But this is just a necessary, rather than a sufficient, condition for democratic knowledge and the proper satisfaction of collective cognitive needs. To really produce such an idealized (and admittedly utopian) community of knowledge producers and collective ethical agents, *norms* must be in place, for a variety of purposes, and legal systems are particularly well suited for this task.

One of the most radical transformations AI could bring is the *diversification of intelligences*. AI will be a mirror of human intelligence, but it will be a lot more than that. AI is not *one* mirror, but a *kaleidoscope* through which we can contemplate different aspects of our minds, which are themselves a fractal of different skills and capacities. By creating multiple intelligences and integrating them into AI and CAI, our capacities will be expanded and integrated beyond recognition. The potential for making a cooperative, peaceful, and more disinterested human species with the aid of CAI is significant—we could become more social, artistic, spiritual, and empathetic. Social and economic rights would be guaranteed if we end up in the blissful cover of the AI Garden of Eden, with all its surplus and efficient automation. The fourth industrial revolution could be at once the most important scientific and technological breakthrough in the history of our species (and a dramatic event in the history of "intelligence" and "rationality" should our species eventually disappear). There is enough drama in this potential narrative to draw analogies of biblical proportions. The Garden of Eden will arrive quickly and perhaps unexpectedly. The apple of forbidden knowledge will be consumed—the "tree of knowledge" of Bacon, Diderot, and D'Alembert, will get considerably more intricate. Perhaps a better metaphor for this event will be the "forest" of knowledge, with CAI having its own way of distributing and producing knowledge.

But the snake is always around the corner. If we lose our autonomy, then we are expelled permanently from the Garden of Knowledge. To use more biblical imagery, the truth will no longer make us free. We would be demoted slaves, through our own industrial design. Thus, although AI has the potential of liberating us from the yoke of toil, and even "push" us toward a post-humanist setting, we run the risk of precluding the possibility of culture and, therefore, of cultural rights, which presuppose a form of human engagement with authentic values, rather than artificial ones. Diverse intelligences, created not only through

AI and CAI interfaces, but also with animal and other intelligences not conceivable at this moment, could create a landscape of diverse skills. A principle that should inform such a development is a non-anthropocentric approach to diverse intelligences and *values*. This is a very intricate and uncertain proposition, but it is clear that a firm foundation on a hierarchy of needs should inform the new edifice of knowledge. Human rights are particularly relevant in AI development because of this. The next chapter explores the possibility of guaranteeing benefits for the entire human species from AI and CAI by informing its development *systematically with need-satisfaction* in a way that complies with contemporary standards and regulations concerning human rights and the minimal conditions for human dignity.

8

Human Rights and Human Needs

8.1 Dignity and Needs: Individual and Social

This section argues that human rights, understood in terms of needs that humans must satisfy as essential components of their dignity and autonomy, can serve as the foundation for humanitarian and ethical AI value alignment.

The previous chapters established the relation between intelligence, agency, and the satisfaction of needs. As a brief restatement, an agent is intelligent only if she satisfies her needs because of her abilities. The more needs an agent has (representational, rational, moral) the more intelligent she must be. In the case of human and animal psychology, attention is the ideal mental capacity for satisfying a wide variety of needs because of its selective and sensitive nature. Since agency is a source of control over the mental and physical actions of the agent, agency is fundamental for trust, responsibility, and credit. Agency is required for normative interpretations of an agent's actions as morally justified, or of her inferences as epistemically justified. If machines become intelligent, they will display forms of attention that are at least extensionally equivalent to human and animal attention, but unless they develop genuine attention routines on the basis of motivations and needs, they cannot be intensionally equivalent to human and animal intelligence.

This chapter seeks to establish the relation between the agential needs that attention satisfies and human rights. It proposes a sketch for an ethical and *humanitarian* (for all humanity's sake) AI/CAI interface on the basis of human rights. Many authors have already emphasized the importance of human rights for generating principles with wide international consensus and support (Cohen, 2010; Donnelly, 2007), which can overcome various difficulties regarding value pluralism, thereby making possible a truly global and humanitarian AI design (Gabriel, 2020). This notion of "overlapping consensus" is, therefore, compatible

with diverse values across cultures and perspectives, including views about morality (Rawls, 2001), because the standards are agreed upon on the basis of their importance for human dignity. As mentioned before, the approach here is not metaphysical (about value or free will) but political and normative (see Rawls, 1985). In addition, human rights have been endorsed by various moral traditions with very different approaches to value (Cohen, 2010). Unless AI is aligned with human rights from the beginning, various aspects of AI disruption (governmental, industrial) would severely jeopardize the international protection of human rights (Liu et al., 2020; Liu, 2018). Interestingly, various types of disruption identified by Liu (2018) can be classified as EEI, IEI, AI, or CAI risks concerning individuals, groups, and societal trajectories.

Iason Gabriel (2020) argues in favor of developing an international and democratic consensus for AI value-alignment based on human rights, but warns that a human rights approach must be more fully developed. In particular, Gabriel claims that there is a problem regarding which human rights should AI be aligned with. The unique contribution of the present approach is to address this problem by showing how needs can overcome the difficulty between "negative" rights that protect individuals from harm and "positive" obligations to help individuals develop capacities to flourish in terms of the satisfaction of their epistemic and moral needs. Thus, the proposal that this chapter defends is inspired by the capability approach (Binder, 2019; Nussbaum, 2011; Sen, 1999), and it is based on a philosophical interpretation of the two major conventions on human rights, on the basis of the human needs these treatises consider as essential for human dignity.

The argument for grounding human rights on cognitive needs is as follows. Human rights owe their universality to the fact that they provide a framework for the protection of the value and dignity of all humans. If human value and dignity are genuinely universal, then they must be based upon a feature of humanity that is indisputably present in humans regardless of culture, political affiliation, race, ethnicity, and other sociological contingencies. If such a general feature of human dignity exists, then it must depend on fundamental aspects of the cognitive and biological makeup that all humans share. The only way to construe this relation in a non-biologically essentialist way is by appealing to the needs humans must satisfy to have a fulfilling and meaningful life. These are needs humans must satisfy in order to have a life with dignity, regardless of who they are or how they are positioned in the scale of positional goods. *Therefore, human rights should be grounded on needs.*

Notice the *normative* character of the conclusion. The claim is not that by virtue of having a human DNA, an individual immediately has rights. Rather, the claim is that the normative aspects of a person concerning her legal standing as a responsible agent who deserves to be treated with dignity *supervene* on the kind of individual she is, particularly her *fundamental cognitive needs*. Human rights are not merely objective or describable natural features of humans; that would hardly make sense, given that legal systems and rights are based on conventions and abstractions. Rather, human rights are normative entitlements a human possesses based on the dignity afforded by having agential needs, particularly the *need for autonomy* that responsible and intelligent beings must satisfy, because it is a necessary condition for the proper functioning of their agency.

Which other needs are relevant as grounds for dignity? As explained in previous chapters, a great variety of needs must be *integrated* for an autonomous intelligent agent to thrive, and it is hard to disentangle them or segment them. In fact, a hierarchy of needs defines the immediate and long-term preferences of an agent, and personal value determines the most categorical needs that define a person's character. We all need to satisfy basic biological needs through representational needs (e.g., spatial navigation, object-based, and feature-based attention) but unlike animals, we satisfy them in profoundly different ways. Focusing on nourishment, some of us are vegans, others omnivorous on the basis of ethical or social convictions; some of us fast, or have a deeper spiritual relation to food while others simply see it as sustenance and are indifferent to the rituals of food consumption. Some enjoy the social and aesthetics dimensions of culinary customs more than others, but all human cultures give importance to the practices of sharing and producing food. Thus, as explained in previous chapters, categorical needs—spiritual, cultural, moral—determine our deepest convictions in a way that guides, integrates, and shapes the fulfillment of other needs, including biological needs. These transcendental needs are at the very top of the hierarchy of a human's ranking of needs, and they are fundamental for understanding human dignity.

In fact, while the satisfaction of biological needs is necessary for satisfying other needs (although representational needs are equally necessary, for humans and animals), when misfortune forces humans to satisfy only their essential biological needs, humans feel completely deprived of their dignity. More accurately, when humans are reduced to a state that forces them to only satisfy their biological needs they *are deprived* of their basic dignity. Literature and history provide ample evidence of how refugees, prisoners, or survivors of a

shipwreck return to a "primitive state," experienced as intensely undignified by these unfortunate human beings. Concentration camps and the Leningrad blockade are recent examples of extreme versions of these situations, where humans can barely satisfy even the most basic biological needs. Tyranny can thus be understood as the oppressive political power that forces humans into a situation of being "reduced to an animal." This is not because animals are deeply inferior—they are not, evolutionarily or in terms of intelligence—but because being treated like cattle or a farm animal is *incompatible with pursuing transcendental needs*. So even if our biological needs are guaranteed to be satisfied, none of us would opt to live like a farm animal (incidentally, this is not at all a justification for how we treat animals in farms).

Humans have a heightened need for autonomy because mere physical freedom and the satisfaction of biological needs are not enough to genuinely satisfy the need for *transcendental autonomy—the autonomy and freedom to flourish as an individual*. Guaranteeing a lack of external intervention or oppression is crucial for satisfying moral and rational needs, but even this is not enough for full personal flourishment and freedom, as the capability approach shows (Binder, 2019). An advantage of a need-based account of human rights is that it is not species-specific or "bio-essentialist" because animals share similar representational, emotional, and even moral needs. A capacious analysis of the conditions under which some rights could be extended to animals becomes possible, and if AI develops, even to machines. But human transcendental needs are at the top of the hierarchy for a very good reason: they explain uniquely the notion of human dignity. These are needs only humans seem to have, at least to the high degree that they do. There is, therefore, something genuinely unique about human dignity. And since this is a general feature of *human autonomy*, it provides a universal basis for human rights. Autonomy is an essential feature of agential intelligence, but the kind of autonomy that is relevant to explain human dignity depends on the satisfaction of transcendental needs. Human dignity is fundamentally normative because it is intrinsically good to have it and protect it. Transcendental autonomy resides in the human *potential* to develop and nurture "disinterested" kinds of flourishing that will define a person's character, and which are fundamental aspects of a person's *well-being*.

Many of the current guidelines issued and approved by states aim at creating an international framework of "soft-law" or a series of recommendations intended to promote international AI ethical regulation, without direct measures for legal enforcement (Jobin et al., 2019). While this is a good development and

a step toward the creation of an international framework for an ethical interface with AI, there is wide discrepancy between separate commercial regions with fundamentally different political and legal traditions. Some emphasize privacy and the need for transparent AI that explains its reasoning (this is certainly the case with Europe), but other regions emphasize AI production independently of addressing these concerns (this is the case with China, and the situation is slightly ambiguous in the United States). Thus, more clarity is needed in order to specify how a legal framework could also help design an *interface for ethical interactions* between AI and humans that is genuinely universal. Soft law is a good beginning, but it is too dependent on political narratives and agendas. The present proposal is that an interface for AI regulation and control based on the needs examined previously could at least start a discussion of how to create an international framework for ethical AI/CAI, interpreted in accordance with human rights. A key feature of this ethical interface with AI/CAI is that it should provide a balanced understanding of human dignity based on a philosophical, political, and moral analysis of needs.

A thorough analysis of the hierarchy of needs will require various ethical approaches. For instance, Kantian and neo-Kantian deontological approaches seem better suited to satisfy some categorical autonomy needs while utilitarian and maximization-preference views are better suited to satisfy representational and rational needs, as well as biological and emotional needs because of their emphasis on sentience. This would be a multidimensional approach based on needs, rather than specific policies or interests, or the application of a single moral theory as a matter of principle. A need-based approach may be more productive, politically and culturally flexible, and compatible with jurisprudential practices, which do not depend on the application of a single moral framework for all cases.

However, unlike policy proposals based on a multidimensional analysis or a multivector approach (Bostrom et al., 2020) the present proposal emphasizes the importance of transcendental needs, particularly to understand autonomy and dignity, as the basis for an international framework to regulate AI. This is not because of a utopian desire for human flourishing, but rather because of the universal motivation humans have to satisfy their transcendental needs unencumbered and autonomously. Their dignity depends on it. As mentioned before, *virtue theories* are well-equipped to account for motivations and needs. A broad framework that includes individual and collective needs, epistemic and moral, could be at the basis of a virtue theoretical approach that could inform an international legal system for AI interfaces. Virtue theories are among the oldest

ethical traditions, from ancient Greece to Confucian and Buddhist approaches, which show that while they provide a firm foundation for ethics and well-being, they are flexible enough to accommodate cultural variation.

8.2 Human Dignity and Freedom: A Historical and Skeptical Perspective

This section discusses the negative and positive conceptions of freedom. It addresses skeptical worries about the nature and history of human rights, drawing on the work of Samuel Moyn.

Through an understanding of preference-rankings that depend on a value-dependent hierarchy of needs, freedom and dignity could be protected at the individual and collective or international levels. But what exactly is the relationship between needs and freedom? According to a traditional distinction in political philosophy, negative liberty is the guaranteed absence of any obstacles on an agent's actions, while positive liberty is the guarantee of being able to act under one's own control in order to fulfill or pursue our most important plans and purposes (see Berlin, 1969). Let us assume that *agential autonomy* correlates with negative freedom, in the sense that agents must be allowed to act without obstacles in order to satisfy their basic needs. *Transcendental autonomy*, by contrast, correlates with positive freedom because satisfying needs in accordance to their value-priority requires not merely absence of obstacles, but also concrete guarantees that society will *help* the agent flourish. Agential and transcendental autonomy are fully compatible, but the problem is that negative and positive freedom are in tension because of political and legal reasons.[1]

For the *liberal* tradition that conceives of the state as a source of obstacles to and burdens on the freedom of its citizens, the relevant notion of liberty is the negative one—in fact, positive freedom invites state intervention and potential abuse because the state now has a license to define and impose "forms of personal flourishment" and this is a considerable risk that free individuals should not permit. Advocates of positive freedom would object that unless the state *cares* for the well-being and flourishment of its citizens, the mere lack of obstacles for action can turn into complete loss of dignity through the indifference and abandonment of those who need support, turning their "negative freedom" into a set of "rights" that most agents cannot use or act upon. So there is considerable tension here that lies precisely on the notion of autonomy or "self" assumed

by these views. For the defender of positive liberty, freedom depends on the type of beliefs, values, and desires agents should autonomously have, while the proponent of negative liberty sees this characterization of a free subject by the state as an affront to her independence and dignity.[2] Pursuing "transcendental needs" from the perspective of legal "protections" can lead to totalitarianism. Preventing the state from any kind of intervention can lead to grotesque income inequality and the segregation, segmentation, and abandonment of large sectors of the population. This is a real political dilemma.

A similar point can be made about fairness. There is a tradition in legal and political philosophy that postulates a state in which subjects are "equal" as the foundation of political authority, by fiat and from the very beginning. A classic formulation of this approach is to conceive of a prior or "primitive" situation as justification for modern political and moral systems, such as Rousseau's positively portrayed "state of nature," or Hobbes's opposite formulation of such a state as driven by violence and greed. The moral and rational foundation of consensus-based authority are idealized in a similar way, in order to guarantee a state of equality among peers, in Kant's "kingdom of ends" or Rawls's "original position" presupposed in his "veil of ignorance" proposal. But how could one guarantee that these idealizations are not justifying traditional forms of oppression and privilege under the guise of rational necessities? How can we prevent the unfortunate consequence that these idealizations might justify forms of control that only a few "rational beings" could exert, with a portrayal of rationality that can be quite narrow, based on the satisfaction of a select group of cognitive needs by a privileged group of individuals who are in control?

By contrast, a wide variety of needs can be used as the groundwork for a non-idealized *relational stance* (Haraway, 2004) according to which the main objective of a moral framework is to eliminate traditional forms of oppression. A needs-based approach has the advantage that the narcissistic presuppositions of anthropocentric human intelligence are no longer divided by a sharp boundary between privileged rational humans and the rest of other "irrational" intelligences. This is a more capacious and generous way of approaching dignity, and it also explicitly acknowledges the needs of historically oppressed groups. But here we confront a similar dilemma. By blurring the narcissistic boundary between human and "other" intelligences, we risk deflating the justification for protecting human dignity for its own sake. Losing this justification is highly counterproductive, and it is incompatible with the previous argument in favor of transcendental needs and autonomy. The idealized-scenario views don't have this problem, but they confront the thorny question of determining who counts

as rational, according to some ideal capacity for reflection or judgment, and why this capacity is required to justify fairness? So there is a dilemma here as well.

With respect to legal systems, these political dilemmas produce two entirely different ways of protecting human dignity. The liberal approach focuses completely on protecting individuals from any obstacle or intervention on the part of the state. The basic human need this approach protects is *agential autonomy*—the state should not impose restrictions on agential autonomy. Whether the subjects of the state end up satisfying just their biological needs or flourishing in ways they find extremely fulfilling is not, and *should not be*, the business of the state. Whose business is it then? It should be the exclusive business of each citizen. The state should not operate under a rich conception of its citizens as "needy" creatures that must be protected. This rich conception generates the very serious risk of totalitarianism and paternalistic intervention and surveillance. On the radically different welfare or care approach, *transcendental autonomy* and a framework for allowing subjects to act in a way that effectively leads to virtue and flourishment are fundamental. According to this view, a "thick" notion of human dignity and its protection by the state are crucial to meet the standards of fairness and legitimate power. As Section 8.4. shows, this conflict actually played out for a good part of the last century on a global scale, shaping the main contours of the two major Covenants for the protection of human rights in international law.

The tension between negative and positive liberty is significant. But clearly there must be some connection between these kinds of liberty. For instance, the notion of "republican liberty" attempts to strike a balance, according to which mere absence of obstacles is insufficient for genuine freedom—one also needs an *official commitment* that one has a status as citizen in a society that guarantees rights to subjects in order to protect them from arbitrary or illegitimate interventions from the authority. There is considerable debate about whether this notion of liberty is genuinely different from the negative one, and to what extent. But it is clear that the human need for overall autonomy is not merely agential autonomy. Humans have higher needs and their political systems *must reflect so*, not because of paternalistic reasons (as the liberal would object) but because of *humanitarian reasons*. An attempt at resolving this tension is offered in what follows.

As the argument presented above makes clear, the hierarchy of needs developed in previous chapters will play an essential role in the explanation and articulation of the *humanitarian AI interface*. The relation between needs and rights or political entitlements is a central focus of critical theories of state

power (Heller, 1976). The need for liberty and autonomy, broadly construed, is also fundamental for a long tradition in political philosophy of which Rousseau and Kant are paradigmatic examples. But the notion of "human right" is at once ambiguous and puzzling. It is ambiguous because rights are afforded by states (which is a basic assumption of the concept of republican liberty). The addendum "human" suggests something natural, universal, and deterministic about these rights. But what can this be? As explained above, there is a way to avoid this difficulty by appealing to transcendental needs, but their universality needs interpretation, which could lead in various directions. Human rights are puzzling because if humans have them by virtue of being human it is not clear why they deserve so much heated political debate at the constitutional and international levels—their status as "self-evident truths" seems suspect.

The history of human rights is equally fraught and complicated. The history of dignity, understood in terms of the widely agreed upon international human rights framework, is quite recent and deeply related, as Samuel Moyn (2010) documents, to the collapse of the nation-state as sole guarantor of rights, as well as to the correlative urgency of adopting an alternative political narrative that could replace the nation-state. Moyn calls human rights "the last utopia" (the title of his book) because the promise of protecting dignity at the international level became paradoxically entangled with local programmatic agendas for their implementation by nation-states, with unclear but also quite intricate visions of "postcolonial justice" and democracy "promotion." The popularity and urgency of this utopia in our very recent past boosted the emergence of human rights in international law. But the utopia could not materialize because, as is the case with many previous utopias, too many political agendas and national interests were at play in interpreting and adopting it. Moyn writes,

> As a number of its partisans in the 1970s were well aware, human rights could break through in that era because the ideological climate was ripe for claims to make a difference not through political vision but by transcending politics. Morality, global in its potential scope, could become the aspiration of mankind. But the very neutrality that allowed for human rights to survive in the 1970s, and prosper as other utopias died, also left them with a heavy burden later […] If human rights were born in antipolitics, they could not remain wholly noncommittal toward programmatic endeavors, especially as time passed.
>
> (Moyn, 2010, 213)

Morality is, and has always been, a universal aspiration, but the burden of morality is that pursuing it in the context of the modern state has meant

that morality's universal aspirations always fall in the hands of political and commercial interests that pursue their own agendas. We now live, more than ten years after Moyn's book, in a recalcitrant political environment where we confront not only the nation-state with its hegemonic and myopic power, but also an angry resurgence of the crudest types of nationalism. Maybe a human rights-based interface to confront the threats of tribal nationalism, massive automation, and the loss of autonomy is truly our last utopia. Perhaps this time, however, a theory of needs should inform this humanitarian effort, rather than the obsolete and stale nation-centered narratives that still dominate politics.

In spite of all the historical complexities of the political discourse on human rights, we need to come to terms with the fact that through a network of international organizations and treatises, as well as *national Constitutions*, humans across the globe hold dignity to be *inalienable*. Dignity might be inalienable because of autonomy (see Rosen's, 2012, Kantian account of dignity), which is compatible with the needs approach this book defends. However, "implementing" dignity is subject to delicate negotiations. The hierarchy of needs, for instance, requires a balance between opposing forces even at the individual level. But humans *strive* to achieve this balance as part of their autonomy. Even if the notion of human dignity and the recent and highly politicized discourse on human rights are historically contingent, the argument for human rights based on human needs shows that the relation between human needs and human rights is *conceptually necessary*—to think of human rights as grounded on cognitive needs justifies a universal framework designed to protect human dignity. The similarity in dignity and worth among humans is thus justified in virtue of their similar cognitive needs. With the possibility of AI looming on the horizon, our dignity and autonomy are under threat. We certainly are in desperate need of revisiting our utopias.

8.3 Control, Needs, and Care

This section addresses issues of autonomy and power in the context of the internet and the "attention economy." It discusses the work of Shoshana Zuboff.

Inalienable rights and dignity become central in the context of machine-human interfaces and interactions because of the loss of control that automation and AI bring to the fore. Contemporary deep learning and predictive algorithms have unprecedented access to information, which translates, in the political realm, into the troubling consequence of *unprecedented surveillance*. This unrestrained

and omnipresent automatized surveillance generates new forms of (i) *intrusion* that violates *negative liberty*; (ii) *exploitation*, of our attention or mental agency by constant exposure to attention attractors in the form "positional goods," which seriously prevent human flourishing by appealing exclusively and addictively to our selfish needs, thereby threatening our *positive liberty*; (iii) *erosion of the public sphere and our capacity for rational argumentation*, which is a necessary condition for informed civil discourse and democracy, threatening our *collective liberty*; and (iv) *further segmentation* of knowledge, understanding, and engagement in the hands of corporate and governmental interest, which are competing for who ends up controlling the "attention-market," jeopardizing the basic agential *control* required for *positive freedom*.

AI/CAI, should they come to fruition, will radically change "politics as usual." If AI/CAI gains control, it is game over for traditional political actors, unless, of course, there is a way of solving the interface problem. This is the reason why it is so important to create independent and humanitarian AI/CAI. Under the current conditions of the "attention-economy" just mentioned, an international framework based on human rights is the only way to guarantee that such a powerful technology doesn't fall in the hands of already ultra-powerful and invasive private and parochial political groups. AI/CAI will hopefully be ethical at least in the sense that information gathering will not eliminate, but instead enhance, human freedom. Massive automatized surveillance is already operating at Orwellian and deeply troubling levels, creating obstacles for the protection of human dignity. Automatized poverty, discrimination, segmentation, and the mental enfeeblement of large portions of humanity are urgent dangers that must be addressed with all the available resources.

The autonomy risks generated by online commercial surveillance, which fuels the databases that AI technology feeds on, are quite troubling. Shoshana Zuboff documents developments in the commercialization of search algorithms that have turned capitalism into an exploitative regime of control that monitors our thoughts, desires, and crucially, our *attention skills*. Given the arguments presented in this book, it is not a stretch to say that Zuboff's research (2019) demonstrates that by being forced to participate in "surveillance capitalism," humans are seriously jeopardizing their agency and autonomy—the biggest threat to an intelligent agent. Her book opens with the following definition of *surveillance capitalism*:

> 1. A new economic order that claims human experience as free raw material for hidden commercial practices of extraction, prediction, and sales; 2. A parasitic economic logic in which the production of goods and services is subordinated

> to a new global architecture of behavioral modification; 3. A rogue mutation of capitalism marked by concentrations of wealth, knowledge, and power unprecedented in human history; 4. The foundational framework of a surveillance economy; 5. As significant a threat to human nature in the twenty-first century as industrial capitalism was to the natural world in the nineteenth and twentieth; 6. The origin of a new instrumentarian power that asserts dominance over society and presents startling challenges to market democracy; 7. A movement that aims to impose a new collective order based on total certainty; 8. An expropriation of critical human rights that is best understood as a coup from above: an overthrow of the people's sovereignty.
>
> (Zuboff, 2019)

Zuboff shows in her book why this politically erosive kind of abuse based on our misplaced trust in corporations that are using us as data is particularly egregious: It leaves us without a home within our homes, transforming us into *informational refugees*, constantly surveilled and controlled. This is a form of estrangement from ourselves, a loss of familiarity, and also a loss of autonomy, *both agential and transcendental*. Since surveillance capitalism targets our attention capacities, making them addicted to positional goods, we have become "entrained" by them. Unfortunately, because of this development, we are now more indifferent than ever to the balance and proper functioning of fundamental cognitive needs, most alarmingly our empathic and emotional needs.

Our "corporate oppressors" have spectacularly solved the problem of *imperative coordination* by exploiting the very nature of attention routines. Our reward system has been patrolled, reinforced, surveilled, and solicited through the media that we depend upon for communication. Humans are now, for the most part, satisfying only selfish or Luciferian needs at the cost of basic emotional, care, and transcendental needs. But paradoxically, they pursue these selfish needs without *really having control*. We have the illusion of control when we pursue selfish commercialized needs, but the only ones who benefit from this "lab-rat" behavior and really have control are the corporate ventures that create the products we use. They use us as guinea pigs and as data. We trust them and surrender our freedom. The distance between commercialized data curation and our deeper sociocultural realities has made "ground truth," or the technique for categorizing the features of reality upon which algorithmic prediction depends, untrue and manipulative (Crawford, 2021).

The good news is that we still have enough agency and autonomy to stop this situation from becoming permanent. AI, actually, opens the door for seriously

revising and transforming the way machine-human interactions are threatening and corroding our autonomy. We confront the most pervasive, silent, and invisible type of totalitarianism, as well as the end of democracy, if we remain passive. Zuboff is right to draw attention to the attack on human rights by these corporations. Commenting on Hanna Arendt's *The Origins of Totalitarianism* and the work of Theodor Adorno, she writes,

> It was the individual's experience of insignificance, expendability, political isolation, and loneliness that stoked the fires of totalitarian terror. Such ideologies, Arendt observed, appear as "a last support in a world where nobody is reliable and nothing can be relied upon." Years later […] Theodor Adorno attributed the success of German fascism to the way in which the quest for effective life had become an overwhelming burden for too many people. "One must accept that fascism and the terror it caused are connected with the fact that the old established authorities … decayed and were toppled, while the people psychologically were not yet ready for self-determination. They proved to be unequal to the freedom that fell into their laps." Should we grow weary of our own struggle for self-determination and surrender instead to the seductions of Big Other, we will inadvertently trade a future of homecoming for an arid prospect of muted, sanitized tyranny.
>
> (Zuboff, 2019, 518)

We implicitly believe that the enfeeblement and dependence that contemporary surveillance capitalism produces might be a necessary and even a welcome trade-off for the conveniences and comforts of massive automation. But Zuboff is right, we are deeply mistaken in thinking that the political consequences of our unjustified trust and dependence on surveillance machines are anything less than a new form of tyranny. If we continue on this path of unprecedented hegemonic power based on the accumulation of corporate wealth and knowledge, which includes primarily knowledge *of ourselves*, we are surrendering our agential autonomy and the possibility of democracy. The surveillance capitalists will be better at predicting our behavior than we are. We will live in a world without transparent collective motivations—an estranged existence that is wholly *unfamiliar and heteronomous*. A global market driven by unbridled capitalistic surveillance will create, ultimately, a world without human dignity. This is why changing the needs that our contemporary markets are designed to satisfy is so crucial in guaranteeing the protection of basic human rights.

Since these are the current conditions of our machine-human interactions, it is of the utmost importance to start focusing all our efforts on developing

an ethical AI interface—an interface that is not exclusively concerned with satisfying the Luciferian needs of a few politicians or impresarios that exploit the attention capacities of humans through digitally designed addiction. Establishing this framework would greatly help humans regain their dignity and autonomy. We must develop attentional "societies of care" as part of this framework. Trust, reliability, and familiarity—all aspects of human psychology that are eroded by surveillance capitalism—need to be grounded on the genuine and non-manipulative satisfaction of autonomy needs (e.g., emotional, agential, representational, moral, transcendental). Fostering and creating new topographies and environments of attention can be the basis for effective activism against surveillance by generating community engagement, as well as the sheltering and reinvigoration of attention. Communities of empathy that satisfy emotional needs can neutralize the "mute and sanitized terror" of losing our dignity. This effort should include protections for our physical environment as well, which is also being exploited by big tech (Crawford, 2021).

Unlike the local and cacophonous approaches to human rights based on political agendas, the "intelligent" technology behind surveillance capitalism is universal and omnipresent, so a human rights response *needs to be equally universal and omnipresent* if we stand any chance at preserving our autonomy. Only a global human rights response, understood in terms of the satisfaction of *all* cognitive needs, is adequate on the face of this threat. This "parasitic economic logic in which the production of goods and services is subordinated to a new global architecture of behavioral modification" will not disappear unless we fight back. We must prevent the ultimate surrender of our political autonomy and the complete shattering of knowledge into pockets of corporate power from happening. We must turn AI/CAI into allies of care. The ethical interface for human rights protection should be designed now, before AIs are developed on the basis of our current surveillance capitalistic technology.

8.4 A Balance of Needs: The Two Covenants

This section explains how the two most important treatises on human rights can be interpreted in the light of the distinctions and concepts presented in this book, particularly in terms of autonomy needs.

Fortunately, a human rights framework, created in the aftermath of the horrors of the Second World War, is still in place. The international community created

this framework specifically to avoid the kind of complete loss of dignity Zuboff warns us about. The two conventions that lay out the foundations for the international protection of human rights are among the very few documents that receive widespread acceptance and still generate robust consensus among all states. These two conventions are the *International Covenant on Civil and Political Rights* (or ICCPR) and the *International Covenant on Economic, Social and Cultural Rights* (or ICESCR). Other key human rights conventions, global and regional, can be interpreted as specific ways of enhancing, clarifying, and enforcing the rights enshrined in these two major treatises.

Conventions are the standard way of creating democratic consensus among states, and they are the key instruments to achieve the kind of rational and legitimate authority required to govern free and autonomous subjects. In other words, legal conventions satisfy a democratic *collective need for autonomy*. Jean-Jacques Rousseau writes in the chapter on slavery of *The Social Contract*: "Since no man has a natural authority over his fellow-man, and since force produces no right, conventions remain as the basis of all legitimate authority among men" (Rousseau, 1997, 44). Conventions are the sole source of political autonomy; otherwise there is the risk of political slavery and lack of trust in authority. As the previous chapter explained, the belief in the law as a rational source of power based on a convention or social contract is key to Weber's solution to the problem of imperative coordination. We need to reinvigorate these conventions that protect human dignity, the ICCPR and the ICESCR, against surveillance capitalism, as well as other dark and invisible forces of enfeeblement and servitude.

A tension was described above between negative and positive liberty—on the one hand, we want to be as free from preconceptions and state domination as possible; on the other hand, we need a robust enough conception of dignity for the state to be able to provide conditions for capable human action and flourishing. The tension is that one is too minimal, the other too authoritarian. This is indeed a dilemma. But it is seldom noticed that this is a dilemma *only within the context of the power of the state*. It is *the state* that leaves citizens behind by being so minimal, or authoritarian by having a specific preconception of the "good citizen." When human needs are considered as central, this changes dramatically. More specifically, if we ignore the power of the state and focus instead on the *kinds of autonomy-needs* that humans must satisfy in order to be free agents, then this dilemma presents a false dichotomy. An international framework makes possible this kind of refocusing, back to human needs rather than the power of the nation-state (incidentally, the modern bureaucratized and militarized state itself is a

quite recent invention; it is a legal abstraction that emerged after the renaissance in Europe). Thus, the dilemma concerning the negative and positive conceptions of freedom dissipates when human needs are the sole focus of analysis—what matters are the kinds of human autonomy, *agential and transcendental*, that ground human dignity, both of which are necessary and valuable.

The ICCPR guarantees negative freedom, or agential autonomy.[3] It protects life and bodily integrity (the prohibition against torture and genocide concern collective and particularly heinous violations of these rights and thus cannot be tolerated from any state under any circumstance), the freedoms of movement or transit (the prohibition against slavery is equally absolute and not subject to exceptions), assembly, expression, the right not to be arbitrarily detained and have a fair trial, equal recognition before the law, the right to form a family and serve in politics. An agent that is free by being protected in this way has negative freedom. Although these essential guarantees for agential and political autonomy are necessary for human dignity, they are insufficient for human flourishing. Moreover, negative freedom may be too minimal even for the lowest standards of human dignity. As proponents of republican freedom argue, one could have agential autonomy or negative freedom *by luck* in a state that is largely totalitarian. Even the notion of republican freedom is too minimal to capture the whole range of human dignity. To guarantee human flourishing, therefore, transcendental autonomy is necessary.

The ICESCR protects rights that fall under the category of transcendental autonomy needs. These include the right to self-determination, the free pursuance of economic, social, and cultural development; the right to fairly distributed and dignified work (in proper conditions, including trade unions), the full realization of economic, social, and cultural development through work and education, social security, family assistance, the safeguard of children, and the highest possible standard for mental and physical health; the protection of authorship and artistic works, the democratic promotion of science and the arts. While this certainly is a "thick" notion of positive freedom that entails substantial intervention by the state in the areas of education, health, labor, and culture, both Covenants protect human dignity on the basis of the sovereign decision of the signatory states, as members of the United Nations, and this is crucial for the present discussion. Protecting human rights is not simply a utopian "dream" concocted by some "*activists*"—these are *binding international treaties*, ratified by the vast majority of contemporary states. They constitute an unprecedented cornerstone of international law, and they are the most significant achievement regarding *humanitarian imperative coordination* at a global scale because

the rights contained in these Covenants are obligations of the states based on consensual, rational, and legal authority.

Human dignity should be protected *regardless of what conception of the state one favors*. This is a *moral* command. Humans deserve to live a dignified life, regardless of their culture, ethnic background, sexual preference, race, and certainly, of how they end up characterizing the functions of the state. Thus, the focus of AI–human interface development should be on protecting human dignity, rather than on the functions of the state as specified by some local legal culture, which are both irrelevant for the development of machine-human interfaces and also dangerous as a source of polarization. States might, in the end, become a major obstacle for the development of ethical interfaces with AI because of their narrow agendas and confrontational short-term planning, and this is why the efforts to develop this framework should be based on international law *from the beginning*. This project is admittedly idealistic, particularly in the current context of ultra-national propagandas. But pursuing this project is exactly what these major international conventions demand. Thus, this effort is not at all without legal justification. Besides legal justification, it is also justified by *ethical and epistemological norms*, based on the satisfaction of human needs.

Corporate conglomerates have outpaced the old nation-state's hegemonic techniques of surveillance and power with their new "surveillance capitalism" technologies, including AI. Some states are taking advantage of these new technologies and have recruited them as allies in the task of organizing hegemonic power, but it is clear that the old structures of the state are becoming outdated and that states can no longer, by themselves, protect the freedoms of their citizens—certainly not from surveillance capitalism. The Covenants are valid legal documents that constitute the core of the humanitarian agenda of the United Nations, and states have agreed that these conventions contain the consensual understanding of human dignity at an international level. The justification for the international protection of human dignity is both legal *and* moral, but this does not mean that AI interfaces need to be "moralistic." Moyn (2010) comments on the two goals behind the human rights agenda that was "born of the yearning to transcend politics," namely, preventing catastrophe and developing a world of human flourishing as follows:

> If human rights call to mind a few core values that demand protection, they cannot be all things to all people. Put another way, the last utopia cannot be a moral one. And so whether human rights deserve to define the utopianism of the future is still very far from being decided.
>
> (Moyn, 2010, 227)

The utopianism of the future is indeed very far from being decided, but whatever it might be, it will certainly involve machine-human interactions. The last utopia will not be a moral one, in the sense that it will not depend on *the monolithic application of a single moral theory*. But human rights can be "all things to all people" if they are interpreted as the protection of human dignity in terms of agential and transcendental needs. All humans have these needs, not as a matter of genetic necessity, but because they are agents that must use their intelligence in an autonomous way in order to satisfy their needs. Humans have these needs not based on descriptions of their patterns of behavior and biological makeup, but rather, as normatively salient requirements concerning how they should behave and live their lives. The proper analysis of these agential needs can also mean "all things to all people" because some of them are clarified through scientific research (e.g., representational, emotional, and rational needs) or interdisciplinary investigations in the sciences and humanities (e.g., moral, aesthetic, and transcendental needs).

The two Covenants were designed to prevent a catastrophe and also to pursue an agenda of human flourishing. In the context of machine-human interactions, they can prevent the catastrophe of human demotion and create an interface for AI that will increase human flourishing. Based on Zuboff's research concerning surveillance capitalism, human beings have *a claim of mistrust* against the implementation of machine-human interactions until an ethical interface proves to be trustworthy. For this to happen, however, various forms of need satisfaction must be guaranteed because these correlate directly with various forms of *trust*: epistemic (rational and representational), moral, empathic, emotional, and in general, *agential*. Even if one is cynical or skeptical about this project, based on the contingent historical reasons Moyn describes, it is time to reconsider and reinterpret this existing international framework in accordance to the conceptual connection between rights and needs.

This new framework for dignity in machine-human interfaces will create new environments and communities of trust and new landscapes for autonomous attention. Contemporary surveillance capitalism is best understood in terms of "vicious" and collective forms of attention by corporate interest. We cannot trust these enterprises, which exploit and use our attention routines. AI/CAI will need to be "more intelligent" and "care" for human dignity by attending to human needs with similar omnipresence and determination. Moreover, genuine AI may be a much better counterbalance to authoritarian governments, by becoming guarantors of rational trust that can powerfully replace the eroded, tangled, and ancient structures of hegemonic power. Any redefinition of intelligence

cannot possibly be politically, culturally, or ethically neutral, and AI will be a pivotal event in our understanding of intelligence and autonomy. This event will transform our notion of humanity, including human dignity, and we are still on time to make sure that human dignity is preserved in this momentous transition. Ironically, this intimidating new technology, if aligned with humans through its attentiveness to our needs, may prove to be the best ally in ensuring our dignity.

Notes

Introduction

1 See Binder (2019) for an overall account and references therein. Classic accounts of the capability approach are Sen (1999) and Nussbaum (2011). Iason Gabriel (2020) argues that the capability approach and the human rights framework are compatible with principles for broad consensus in spite of individual and cultural differences (such as Rawl's "veil of ignorance" and principles of distributive justice). I argue explicitly for the importance of politics and human rights in AI design in the concluding chapters.

2 This vast literature centers on issues regarding causation and determinism, at the center of the compatibilist and incompatibilist views. This literature is too extensive to cite or properly address here. See Fischer et al. (2007) for a dialectical presentation of the main debates. For the importance of cultural and social contexts in the development of abilities that are relevant for moral responsibility, see Vargas (2013).

3 For the importance of protecting our freedom to think on the basis of the existent human rights framework, given the invasive and commercialized social media we use daily, see Alegre (2022). For a more general approach to the relation between human dignity and the international human right framework, see Gilabert (2018). These issues are developed in the last two chapters of this book.

4 Virtue theories in ethics and epistemology appeal to capacities in order to explain the good features of agents, and they appeal directly to the habits and skills of agents, rather than to reasons, norms, beliefs, or other factors relevant for good action, abstractly construed.

5 Arpaly argues against ethical views that require reflective autonomy as a condition for responsibility. It is my own interpretation that the kind of virtuous attention selection in the morally relevant cases she describes is autonomous because they are excellences of an agent, in accordance with a capability and virtue ethics approach. This is fully compatible with this book's commitment that an agent does not need to satisfy her needs through reflective endorsement for her actions to count as sufficiently under her control (see Fairweather and Montemayor, 2017).

6 For conciseness, key texts on the social, political, and epistemic risks that inform the present account cannot be discussed at length. Among them are the contributions by Ruha Benjamin (2019), Kate Crawford (2021), Georgi Gardiner

(forthcoming), Gabbrielle Johnson (forthcoming), and Melanie Mitchell (2021). The preface acknowledges some of these contributions to the literature on AI, algorithmic injustice, ground-truth, social epistemology, and the norms of attention. This footnote mentions only the most salient ones.

Chapter 1

1 Many of our electronic "assistants" have female names and voices. Gender brings another critical variable into the power dynamics between mastermind and creation, between life and mind, and between substrate and purposeful action. This topic is too rich to address here adequately. But I would like to highlight a couple of issues. First, the transhumanist movement has the merit of challenging the traditional gender dichotomies assumed in many of these science fiction plots. Second, the culture in Silicon Valley and its counterparts in other parts of the world have managed to reproduce a predominantly male, aggressive, and paternalistic approach to computer design and software engineering (Chang, 2019). It is important to keep this in mind in contemporary discussions concerning value alignment. For a deeper historical perspective on these problems, including an analysis of the governmental support of these technologies in the United States, see O'Mara (2019).

Chapter 2

1 See Montemayor and Haladjian (2015) on why this kind of attention has a unique phenomenology, despite lacking the phenomenology of self-awareness; and Fairweather and Montemayor (2017) for the relation between attentional dexterity and epistemic agency.

Chapter 3

1 Moreover, heuristic or biased reasoning may be prevalent and unavoidable. For the inevitability of bias in machine learning and scientific reasoning, see Johnson (forthcoming).
2 The remainder of this chapter is based on Montemayor (2019a).
3 Performance normativity concerns the kind of evaluative assessment used in skilled or excellent performances, such as virtuous piano playing. The satisfaction of norms of excellence is based on the quality of the performance, judged as the result

of the skills of an agent. In the case of inferential attention, excellence concerns epistemic justification based on the reliability and evidential support that attention routines provide.

4 See Koralus (2014a, 2014b) for an erotetic approach to attention that explains selection and inhibition as question-sensitive; see Fairweather and Montemayor (2018) for an account of the inhibitory functions of attention in terms of virtuous sensitivity and insensitivity to information.

Chapter 4

1 Damasio's (1994) distinction between feelings and emotions is useful to understand the complexity of these needs, as well as their biological dependence.

2 As mentioned in the introduction, there is cognitive, emotional, and caring empathy. This chapter argues that AI could develop cognitive empathy, but not emotional and caring empathy.

3 The rest of this chapter is based on ideas from Haladjian and Montemayor (2016). Machine consciousness is now a flourishing area of research. As before, providing a comprehensive reference list here is not feasible, but these are some salient contributions to the recent literature on AI ethics and consciousness: Coeckelbergh (2020); Husain (2017); Reese (2018); Wooldridge (2020).

4 For accounts of why attention necessitates agency, see Fairweather and Montemayor (2017); and Wu (2011, 2013).

5 A similar point about aesthetic judgment can be traced back to at least Burke (1757).

6 For a more optimistic perspective, see Malle (2016). Malle distinguishes various elements of human moral competence, including vocabulary, norms, communication, decision-making and action, as well as affective and cognitive components. The argument in this chapter intends to demonstrate the importance of the affective dimension. I am grateful to an anonymous reviewer for bringing this research to my attention.

Chapter 5

1 There is a vast literature on this issue. Two examples concern memory (Montemayor, 2016) and time perception (Montemayor, 2013, 2017, 2019b).

2 The work of Amia Srinivasan (2020) is highly relevant here. Like Murdoch, Srinivasan argues that internal reflection and epistemic internalism in general may be not only obstacles to good epistemic and moral thinking, but also sources of evil and poor epistemic performance. See also Gardiner (forthcoming) for how the

virtue of attunement cannot be solely determined by what one can reflect upon, on the basis of one's beliefs at a given time.

3 See Fairweather and Montemayor (2017, Chapter 7) for an account of collective epistemic agency in terms of an attention-assertion model.

Chapter 6

1 One can also think of this kind of intelligence jet lag in terms of the distinction between P versus NP problems in mathematics and decision-making. This interpretation of the intelligence jet lag makes the present approach testable in a formal and scientific way. Although this proposal cannot be developed in detail here, the main idea is that the epistemic distinctions behind P and NP problems (verifiability and solvability) may not be relevant for a *non-human epistemic agent* with a completely different cognitive architecture and a remarkably quick processing time.

2 Relevance problems are related to the so-called "frame problem" in computer science. In its most basic form, the problem is: what information, out of many options, is the most relevant to satisfy an epistemic need, and what is the relevant course of action to properly satisfy epistemic needs. For the importance of relevance problems in epistemology, see Greco (2010, Chapter 10). See also Henderson and Horgan (2009, 2011) for the importance of cognitive integration with respect to content and overall reliability.

Chapter 7

1 The literature on these themes is vast, has exploded in recent years, and will likely continue to increase as AI technologies and innovations are implemented. Various recent books address the issue of control and responsibility from an ethical, legal, and philosophical perspective. Although providing a comprehensive list cannot be done properly here, these are some notable contributions: Abbott (2020); Anderson and Anderson (2011); Barfield and Pagallo (2020); Dignum (2019); Dubber et al. (2020); Chinen (2019); Coeckelbergh (2012); Gunkel (2012); Nyholm (2020); Turner (2019); and Wallach (2015). The present proposal is to distinguish responsibility in extensionally and intensionally equivalent systems (EEI AI responsibility could fall under standard legal and ethical standards, but IEI AI will be radically different).

2 The capability approach and its relation to human rights were discussed in the introduction, and are further developed in the next chapter. A capability approach

must be interpreted in the light of further ethical requirements concerning non-discrimination. For an insightful philosophical discussion of this issue in the context of public policy, see Silvers et al. (1998).

3 See Mindt and Montemayor (2020) for a categorization of AI in terms of knowledge tools and knowledge producers.

4 This strategy is compatible with some of the recommendations by Gabriel (2020). My own view is that the international framework of human rights is a preferable basis for widespread consensus about human values and dignity, and that we do not need to appeal to the idealized epistemology of reflective equilibrium or the Rawlsian idealization of the "veil of ignorance." Since cognitive needs are part of the account of human dignity presented in the last chapter, it can be interpreted as a capability approach to human freedom and well-being (see Binder, 2019; Gilabert, 2018; Nussbaum, 1988, 2020; and Sen, 1993, 1999).

Chapter 8

1 Constanze Binder (2019) defends a very detailed capability account of overall freedom based on an explanation of how opportunity sets can be understood and compared in terms of freedom's "agency value." This analysis is relevant for the present discussion.

2 For a recent installment of this long-standing debate, see Carter and Shnayderman (2019).

3 The account of human rights presented here is partly based on Montemayor (2002).

References

Aaronson, S. (2016), Can computers become conscious? Retrieved from http://www.scottaaronson.com/blog/.

Abbott, R. (2020), *The Reasonable Robot: Artificial Intelligence and the Law*, New York, NY: Cambridge University Press.

Abrams, M. H. (1953), *The Mirror and the Lamp: Romantic Theory and the Critical Tradition*, New York: Oxford University Press.

Ahn, H. I. and Picard, R. W. (2014a), Measuring affective-cognitive experience and predicting market success, *Affective Computing, IEEE Transactions on*, 5(2): 173–86.

Ahn, H. I. and Picard, R. W. (2014b), *Modeling Subjective Experience-based Learning under Uncertainty and Frames*, The Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, Canada. http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8436.

Alegre, S. (2022), *Freedom to Think: The Long Struggle to Liberate Our Minds*, London: Atlantic Books.

Alvarez, G. A. and Oliva, A. (2008), The representation of simple ensemble visual features outside the focus of attention, *Psychological Science*, 19(4): 392–8.

Anderson, M. and Anderson, S. L. (eds.), (2011), *Machine Ethics*, Cambridge: Cambridge University Press.

Arpaly, N. (2002), *Unprincipled Virtue: An Inquiry into Moral Agency*, New York: Oxford University Press.

Axelrod, V. and Rees, G. (2014), Conscious awareness is required for holistic face processing, *Consciousness and Cognition*, 27(0): 233–45.

Baars, B. J. (2005), Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience, *Progress in Brain Research*, 150: 45–53.

Bach, K. and Harnish, R. (1979), *Linguistic Communication and Speech Acts*, Cambridge, MA: MIT Press.

Barfield, W. and Pagallo, U. (2020), *Advanced Introduction to Law and Artificial Intelligence*, Cheltenham, UK: Edward Elgar Publishing.

Bayne, T. (2007), Conscious states and conscious creatures: Explanation in the scientific study of consciousness, *Philosophical Perspectives*, 21(1): 1–22.

Benjamin, R. (2019), *Race after Technology: Abolitionist Tools for the New Jim Code*, Medford, MA: Polity.

Berlin, I. (1969), Two concepts of liberty. In I. Berlin (ed.), *Four Essays on Liberty*, London: Oxford University Press (pp. 118–72).

Binder, C. (2019), *Agency, Freedom and Choice*, Theory and Decision Library A: Rational Choice in Practical Philosophy and Philosophy of Science, Dordrecht, The Netherlands: Springer.

Binder, C. and Binder, C. (2019), A capability perspective on indigenous autonomy, *Oxford Development Studies*, 44(3): 297–314.

Block, N. (1995a), The mind as the software of the brain. In D. Osherson, L. Gleitman, S. M. Kosslyn, S. Smith, and S. Sternberg (eds.), *An Invitation to Cognitive Science*, Cambridge, MA: MIT Press (pp. 170–85).

Block, N. (1995b), On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18(2): 227–47.

Block, N. (2014), Seeing-as in the light of vision science, *Philosophy and Phenomenological Research*, 89(3): 560–72.

Boden, M. A. (2016), *AI: Its Nature and Future*, New York, NY: Oxford University Press.

Boghossian, P. (2014), What is inference?, *Philosophical Studies*, 169(1): 1–18.

Boghossian, P. (2016), Reasoning and reflection: A reply to Kornblith, *Analysis*, 76(1): 41–54.

Boghossian, P. (2018), Delimiting the boundaries of inference, *Philosophical Issues: A Supplement to Nous*, 28: 55–69.

Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P., and Endy, D. (2013), Amplifying genetic logic gates, *Science*, 340: 599–603.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (1st ed.), Oxford: Oxford University Press.

Bostrom, N., Dafoe, A. and Flynn, C. (2020), Policy desiderata for superintelligent AI: A vector field approach. In S. Matthew Liao (ed.), *Ethics of Artificial Intelligence*, New York: Oxford University Press (pp. 293–326).

Bradshaw, T. (2016, January 7). Apple buys emotion-detecting AI start-up. *The Financial Times*. Retrieved from http://www.ft.com/cms/s/0/2b915242-b571-11e5-8358-9a82b43f6b2f.html.

Breazeal, C. and Scassellati, B. (1999), A context-dependent attention system for a social robot, *Paper presented at the Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*.

Breazeal, C. and Scassellati, B. (2002), Robots that imitate humans, *Trends in Cognitive Sciences*, 6(11): 481–7.

Brooks, R., Gupta, A., McAfee, A. and Thompson, N. (February 27, 2015), *Artificial intelligence and the future of humans and robots in the economy*, The Malcolm and Carolyn Wiener Annual Lecture on Science and Technology, Council on Foreign Relations, http://www.cfr.org/technology-and-science/artificial-intelligence-future-humans-robots-economy/p36197.

Bruya, B. (2010), *Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action*, Cambridge, MA: MIT Press.

Buckner, C. (2019), Rational inference: The lowest bounds, *Philosophy and Phenomenological Research*, 98(3): 697–724.

Burge, T. (2014), Reply to Block: Adaptation and the upper border of perception, *Philosophy and Phenomenological Research*, 89(3): 573–83.

Burke, E. (1757), *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*, London: R. and J. Dodsley.

Cáceres, E. and Montemayor, C. (2016), Pasos Hacia una Naturalización Cognitiva en la Filosofía del Derecho (Steps towards a Cognitive Naturalization of Legal Philosophy), *Problema: Anuario de Filosofía y Teoría del Derecho*, 10: 137–65.

Carey, S. (2009), *The Origin of Concepts*, Oxford: Oxford University Press.

Carter, I. and Shnayderman, R. (2019), The impossibility of "Freedom as independence", *Political Studies Review*, 17(2): 136–46.

Carter, S. and McBride, M. (2013), Experienced utility versus decision utility: Putting the "S" in satisfaction, *The Journal of Socio-Economics*, 42: 13–23.

Cavanagh, P. (2004), Attention routines and the architecture of selection. In M. I. Posner (ed.), *Cognitive Neuroscience of Attention*, New York: Guilford Press (pp. 13–28).

Celeghin, A., de Gelder, B., and Tamietto, M. (2015), From affective blindsight to emotional consciousness, *Consciousness and Cognition*, 36: 414–25.

Chalmers, D. J. (1995), Facing up to the problem of consciousness, *Journal of Consciousness Studies* 2.3, Imprint Academic (pp. 200–19).

Chang, E. (2019), *Brotopia: Breaking Up the Boy's Club of Silicon Valley*, New York, NY: Portfolio/Penguin; Random House.

Chen, Z. (2012), Object-based attention: A tutorial review, *Attention, Perception, & Psychophysics*, 74(5): 784–802.

Chinen, M. (2019), *Law and Autonomous Machines*, Cheltenham, UK: Edward Elgar.

Churchland, P. S. and Churchland, P. M. (1990), Could a machine think?, *Scientific American*, 262(1): 32–7.

Cisco Systems Inc. (2015), *The Zettabyte Era—Trends and Analysis*. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html.

Clark, H. H. (1996), *Using Language*, Cambridge: Cambridge University Press.

Coeckelbergh, M. (2012), *Growing Moral Relations: Critique of Moral Status Ascription*, London, UK: Palgrave Macmillan.

Coeckelbergh, M. (2020), *AI Ethics*, Cambridge, MA: The MIT Press.

Cohen, J. (2010), *The Arc of the Moral Universe and Other Essays*, Cambridge, MA: Harvard University Press.

Conti, E., Madhavan, V., Petroski Such, F., Lehman, J., and Stanley, K. O. (2017), Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents, *Arxiv*: 1712.06560.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002), The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals, *Journal of Personality and Social Psychology*, 83(6): 1314–29.

Crawford, K. (2021), *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, CT: Yale University Press.

Cromheeke, S. and Mueller, S. C. (2014), Probing emotional influences on cognitive control: An ALE meta-analysis of cognition emotion interactions, *Brain Structure and Function*, 219: 995–1008.

Csikszentmihalyi, M. (1997), *Finding Flow: The Psychology of Engagement with Everyday Life* (1st ed.), New York: Basic Books.

Damasio, A. R. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain*, New York: G.P. Putnam.

Decety, J. and Cowell, J. M. (2014), The complex relation between morality and empathy, *Trends in Cognitive Sciences*, 18(7): 337–9.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., and Sergent, C. (2006), Conscious, preconscious, and subliminal processing: A testable taxonomy, *Trends in Cognitive Sciences*, 10(5): 204–11.

Dehaene, S. and Naccache, L. (2001), Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework, *Cognition*, 79(1–2): 1–37.

de La Mettrie, J. O. (1747/1994), *Man a Machine and Man a Plant*, Richard A. Watson and Maya Rybalka (trans.), Indianapolis: Hackett Publishing Company.

de Waal, F. (2016), *Are We Smart Enough to Know How Smart Animals Are?*, New York: W. W. Norton & Company.

de Waal, F. (2019), *Mama's Last Hug: Animal Emotions and What They Tell Us about Ourselves*, New York: W. W. Norton & Company.

Dignum, V. (2019), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Cham, Switzerland: Springer.

Di Lollo, V., Enns, J. T., and Rensink, R. A. (2000), Competition for consciousness among visual events: The psychophysics of reentrant visual processes, *Journal of Experimental Psychology: General*, 129(4): 481–507.

Donnelly, J. (2007), The relative universality of human rights, *Human Rights Quarterly*, 29(2): 281–306.

Dotson, K. (2011), Tracking epistemic violence, tracking practices of silencing, *Hypatia: A Journal of Feminist Philosophy*, 26(2): 237–57.

Dretske, F. (1981), *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Dretske, F. (2012), Awareness and authority: Skeptical doubts about self-knowledge. In D. Smithies and D. Stoljar (eds.), *Introspection and Consciousness*, New York: Oxford University Press (pp. 49–64).

Dubber, M. D., Pasquale, F. and Das, S. (eds.), (2020), *The Oxford Handbook of Ethics of AI*, New York: Oxford University Press.

Dürrenmatt, F. (1986/2006), *The Physicists*, New York: Grove Press.

Dworkin, R. (1986), *Law's Empire*, Cambridge, MA: Belknap, Harvard University Press.

Eberhard, J. L., Davies, P. G., Purdie-Vaughns, V., and Johnson, S. L. (2006), Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes, *Psychological Science*, 17(5): 383–8.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., and Davies, P. G. (2004), Seeing Black: Race, crime, and visual processing, *Journal of Personality and Social Psychology*, 87(6): 876–93.

Edelman, D. B., Baars, B. J., and Seth, A. K. (2005), Identifying hallmarks of consciousness in non-mammalian species, *Consciousness and Cognition*, 14(1): 169–87.

Ekman, P. (1992), An argument for basic emotions, *Cognition and Emotion*, 6(3–4): 169–200.

Eriksen, C. W., and Yeh, Y.-Y. (1985), Allocation of attention in the visual field, *Journal of Experimental Psychology: Human Perception and Performance*, 11(5): 583–97.

Eubanks, V. (2018), *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, NY: St. Martin's Press.

Fairweather, A. and Montemayor, C. (2017), *Knowledge, Dexterity, and Attention: A Theory of Epistemic Agency*, New York: Cambridge University Press.

Fairweather, A. and Montemayor, C. (2018), Curiosity and epistemic achievement. In Inan, I., Watson, L., Whitcomb, D., and Yigit, S. (eds.), *The Moral Psychology of Curiosity*, New York, NY: Rowman and Littlefield (pp. 199–216).

Feigl, H. (1958), The mental and the physical. In H. Feigl, M. Scriven and G. Maxwell (eds.), *Concepts, Theories, and the Mind-Body Problem*, Minnesota Studies in the Philosophy of Science: Volume II, Minneapolis: University of Minnesota Press (pp. 370–497).

Fischer, J. M., Kane, R., Pereboom, D., and Vargas, M. (eds.), (2007), *Four Views on Free Will*, (Great Debates in Philosophy), Oxford: Blackwell.

Fodor, J. (2007), Revenge of the given. In B. P. McLaughlin and J. Cohen (eds.), *Contemporary Debates in the Philosophy of Mind*, New York, NY: Basil Blackwell (pp. 105–16).

Fodor, J. (2008), *The Language of Thought Revisited*, New York, NY: Oxford University Press.

Foot, P. (1967), *Theories of Ethics*, New York: Oxford University Press.

Fricker, M. (2007), *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford: Oxford University Press.

Gabriel, I. (2020), Artificial intelligence, values, and alignment, *Minds and Machines*, 30: 411–37.

Gallistel, C. R. (1990), *The Organization of Learning*, Cambridge, MA: MIT Press.

Gardiner, G. (2019), Legal epistemology, *Oxford Bibliographies Online*.

Gardiner, G. (forthcoming), Attunement: On the cognitive virtues of attention. In M. Alfano, C. Klein, and J. de Ridder (eds.), *Social Virtue Epistemology*, New York, NY: Routledge.

Gertler, B. (ed.), (2003), *Privileged Access and First-Person Authority*, Aldershot: Aldershot Publishing.

Gigerenzer, G. (2008), *Rationality for Mortals: How People Cope with Uncertainty*, New York, NY: Oxford University Press.

Gilabert, P. (2018), *Human Dignity and Human Rights*, New York: Oxford University Press.

Goff, P. A., Jackson, M. C., Di Leone, B. L., Culotta, C. M, and DiTomasso, N. A. (2014), The essence of innocence: Consequences of dehumanizing black children, *Journal of Personality and Social Psychology*, 106(4): 526–45.

Goldman, A. I. (2011), A guide to social epistemology. In A. I. Goldman and D. Whitcomb (eds.), *Social Epistemology: Essential Readings*, Oxford: Oxford University Press (pp. 11–37).

Goldman, A. I. (2012), Why social epistemology is real epistemology. In A. I. Goldman (ed.), *Reliabilism and Contemporary Epistemology: Essays*, New York, NY: Oxford University Press (pp. 248–79).

Graziano, M. S. A. (2014), Speculations on the evolution of awareness, *Journal of Cognitive Neuroscience*, 26(6): 1300–4.

Graziano, M. S. A. (2015), Build-a-brain: We could build an artificial brain that believes itself to be conscious. Does that mean we have solved the hard problem? Retrieved from https://aeon.co/essays/can-we-make-consciousness-into-an-engineering-problem.

Graziano, M. S. A. and Kastner, S. (2011), Human consciousness and its relationship to social neuroscience: A novel hypothesis, *Cognitive Neuroscience*, 2(2): 98–113.

Graziano, M. S. A. and Webb, T. W. (2014), A mechanistic theory of consciousness, *International Journal of Machine Consciousness*, 6(2): 1–14.

Greco, J. (2010), *Achieving Knowledge: A Virtue Theoretic Account*, New York, NY: Cambridge University Press.

Greco, J. and Turri, J. (2011), *Virtue Epistemology*, Cambridge, MA: MIT Press.

Greenough, J. (2016), 10 million self-driving cars will be on the road by 2020, *Business Insider*, June 15, 2016, accessed August 1, 2016, http://www.businessinsider.com/report-10-million-self-driving-cars-will-be-on-the-road-by-2020-2015-5-6.

Grice, H. P. (1989), *Studies in the Ways of Words*, Cambridge, MA: Harvard University Press.

Griffin, A. (2017), Facebook's artificial intelligence robots shut down after they start talking to each other in their own language, *Independent*, Monday, July 2017. http://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

Griffin, D. R. and Speck, G. B. (2004), New evidence of animal consciousness, *Animal Cognition*, 7(1): 5–18.

Guenther, K. (2015), *Localization and Its Discontents: A Genealogy of Psychoanalisis & the Neuro Disciplines*, Chicago: The University of Chicago Press.

Gunkel, D. J. (2012), *The Machine Question: Critical Perspectives on AI, Robots and Ethics*, Cambridge, MA: MIT Press.

Guyer, P. (1996), *Kant and the Experience of Freedom: Essays on Aesthetics and Morality*, New York, NY: Cambridge University Press.

Hacking, I. (2014), *Why Is There Philosophy of Mathematics at All?*, Cambridge: Cambridge University Press.

Haidt, J. (2007), The new synthesis in moral psychology, *Science*, 316(5827): 998–1002.
Haladjian, H. H. and Montemayor, C. (2015), On the evolution of conscious attention, *Psychonomic Bulletin and Review*, 22(3): 595–613.
Harari, Y. N. (2015), *Sapiens: A Brief History of Humankind*, New York: Harper.
Haraway, D. (2004), A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. In D. Haraway (ed.), *The Haraway Reader*, New York: Routledge (pp. 7–45).
Hassabis, D. (2017), The mind in the machine, *Financial Times*, April 20, 2017. https://www.ft.com/content/048f418c-2487-11e7-a34a-538b4cb30025.
Haugeland, J. (1981), Analog and analog, *Philosophical Topics*, 12: 213–26.
Heckert, J. (2012, November 15), The hazards of growing up painlessly, *The New York Times Magazine*. Retrieved from http://www.nytimes.com/2012/11/18/magazine/ashlyn-blocker-feels-no-pain.html.
Heikkila, A. (2016), Self-driving cars and the Kobayashi Maru, *Techcrunch*, http://techcrunch.com/2016/02/27/self-driving-cars-and-the-kobayashi-maru/.
Heller, A. (1976), *The Theory of Need in Marx*, London: Allison and Busby.
Helmholtz, H. von (1867/1910), *Handbuch der physiologischen Optik* [Handbook of physiological vision], Leipzig, Germany: L. Voss.
Henderson, D. and Horgan, T. (2009), Epistemic virtues and cognitive dispositions. In K. Steuber, G. Damschen, and R. Schnepf (eds.), *Debating Dispositions: Issues in Metaphysics, Epistemology and Philosophy of Mind*, Berlin: de Gruyter (pp. 296–319).
Henderson, D. and Horgan, T. (2011), *The Epistemological Spectrum*, New York, NY: Oxford University Press.
Hillis, W. D. (2019), The first machine intelligences. In J. Brockman (ed.), *Possible Minds: 25 Ways of Looking at AI*, New York: Penguin Random House (pp. 170–7).
Hommel, B. (2004), Event files: Feature binding in and across perception and action, *Trends in Cognitive Sciences*, 8(11): 494–500.
Hommel, B. (2007), Feature integration across perception and action: Event files affect response choice, *Psychological Research*, 71(1): 42–63.
Humphrey, N. (2011), *Soul Dust: The Magic of Consciousness*, Princeton: Princeton University Press.
Husain, A. (2017), *The Sentient Machine: The Coming Age of Artificial Intelligence*, New York, NY: Scribner.
Isasi-Díaz, A. M. and Mendieta, E. (eds.), (2012), *Decolonizing Epistemologies*, New York: Fordham University Press.
Irving, Z. C. (2019), Attention norms in Siegel's *The Rationality of Perception, Ratio*, 32: 84–91.
Jackson, F. (1982), Epiphenomenal Qualia, *Philosophical Quarterly*, 32: 127–36.
Jacob, J., Jacobs, C., and Silvanto, J. (2015), Attention, working memory, and phenomenal experience of WM content: Memory levels determined by different types of top-down modulation, *Frontiers in Psychology*, 6: Article ID 1603.
James, L., Vila, B., and Daratha, K. (2013), Results from experimental trials testing participant responses to white, hispanic and Black suspects in high-fidelity

deadly force judgment and decision-making simulations, *Journal of Experimental Criminology*, 9(2): 189–212.
Jennings, C. D. (2020), *The Attending Mind*, Cambridge: Cambridge University Press.
Jobin, A., Ienca, M. and Vayena, E. (2019), The global landscape of AI ethics guidelines, *Nature Machine Intelligence*, 1: 389–99.
Joshi, A., Bhattacharyya, P., and Carman, M. J. (2016), Automatic sarcasm detection: A survey, *arXiv preprint arXiv*:1602.03426.
Kahneman, D. and Thaler, R. H. (2006), Anomalies: Utility maximization and experienced utility, *The Journal of Economic Perspectives*, 20(1): 221–34.
Kahneman, D., Treisman, A., and Gibbs, B. J. (1992), The reviewing of object files: Object-specific integration of information, *Cognitive Psychology*, 24(2): 175–219.
Kahneman, D. (2011), *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.
Kak, A. and Richardson, R. (2020), *AI Now Submission to the Office of the Privacy Commissioner of Canada*, February 2020. https://ainowinstitute.org/ainow-comments-to-canadian-office-of-the-privacy-commissioner.html.
Keil, Frank C. (1992), *Concepts, Kinds, and Cognitive Development*, Cambridge, MA: MIT Press.
Kelly, K. (2017), The myth of a superhuman AI, *Wired*. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.
Kelsen, H. (1960/1967), *Pure Theory of Law*, M. Knight (trans.), Berkeley: University of California Press.
Kentridge, R. W. (2011), Attention without awareness: A brief review. In C. Mole, D. Smithies, and W. Wu (eds.), *Attention: Philosophical and Psychological Essays*, Oxford: Oxford University Press (pp. 228–46).
Kentridge, R. W. (2012), Blindsight: Spontaneous scanning of complex scenes, *Current Biology*, 22(15): R605–606.
Kentridge, R. W., Nijboer, T. C. W., and Heywood, C. A. (2008), Attended but unseen: Visual attention is not sufficient for visual awareness, *Neuropsychologia*, 46(3): 864–9.
Knobe, J. and Nichols, S. (2008), *Experimental Philosophy*, New York: Oxford University Press.
Koch, C. and Tsuchiya, N. (2007), Attention and consciousness: Two distinct brain processes, *Trends in Cognitive Sciences*, 11(1): 16–22.
Koch, C. and Tsuchiya, N. (2012), Attention and consciousness: Related yet different, *Trends in Cognitive Sciences*, 16(2): 103–5.
Koralus, P. (2014a), The erotetic theory of attention: Questions, focus and distraction, *Mind & Language*, 29(1): 26–50.
Koralus, P. (2014b), Attention, consciousness, and the semantics of questions, *Synthese*, 191(2): 187–211.
Kornblith, H. (2012), *On Reflection*, New York: Oxford University Press.
Kriegel, U. (2015), *The Varieties of Consciousness*, New York: Oxford University Press.

Kurzweil, R. (1999), *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York: Viking.

Lamme, V. A. F. (2004), Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness, *Neural Networks*, 17(5–6): 861–72.

Lardinois, F. (2012), Google just got a whole lot smarter, launches its knowledge graph, *The Crunch*. https://techcrunch.com/2012/05/16/google-just-got-a-whole-lot-smarter-launches-its-knowledge-graph/.

Lasonen-Aarnio, M. (2010), Unreasonable knowledge, *Philosophical Perspectives*, 14: 1–21.

Lasonen-Aarnio, M. (In press), Virtuous failure and victims of deceit. In J. Dutant (ed.), *The New Evil Demon*, New York: Oxford University Press.

LeDoux, J. E. (2000), Emotion circuits in the brain, *Annual Review of Neuroscience*, 23: 155–84.

LeDoux, J. E. (2003), The emotional brain, fear, and the amygdala, *Cellular and Molecular Neurobiology*, 23(4–5): 727–38.

LeDoux, J. E. (2012), Rethinking the emotional brain, *Neuron*, 73(4): 653–76.

Legg, S. and Hutter, M. (2007), Universal intelligence: A definition of machine intelligence, *Minds and Machines*, 17: 391–444.

Lehman, J. and Stanley, K. O. (2008), Exploiting open-endedness to solve problems through the search for novelty, *ALIFE*: 329–36.

Leibo, Joel Z. et al. (2018), Psychlab: Psychology laboratory for deep reinforcement learning agents, *CoRR,* 1801.08116, http://arxiv.org/abs/1801.08116.

Lewis, D. K. (1988), What experience teaches. *Proceedings of the Russellian Society* 13, Sydney, Australia: University of Sydney (pp. 29–57).

Li, F. F. (2014, May 20), *The digital sensory system: A quest for visual intelligence in computers*, Paper presented at the Stanford Engineering's EngX: The Digital Sensory System.

Lindsay, G. W. (2020), Attention in psychology, neuroscience, and machine learning, *Frontiers in Computational Neuroscience*, 14: 29.

Lisi, M. and Cavanagh, P. (2015), Dissociation between the perceptual and saccadic localization of moving objects, *Current Biology*, 25(19): 2535–40.

Liu, H-Y. (2018), The power structure of artificial intelligence, *Law, Innovation and Technology*, 10(2): 197–229.

Liu, H-Y., Maas, M., Danaher, J., Scarcella, L., Lexer, M., and Van Rompaey, L. (2020), Artificial intelligence and legal disruption: A new model for analysis, *Law, Innovation and Technology*, 12(2): 205–58.

Lupyan, G. (2017), Changing what you see by changing what you know: The role of attention, *Frontiers in Psychology*, 8: 553.

Macpherson, F. (2012), Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism, *Philosophy and Phenomenological Research*, 84: 24–62.

Maley, C. J. (2011), Analog and digital, continuous and discrete, *Philosophical Studies*, 155(1): 117–31.

Malle, B. F. (2016), Integrating robot ethics and machine morality: The study and design of moral competence in robots, *Ethics and Information Technology*, 18: 243–56.

Malmgren, A. S. (2018), Varieties of inference?, *Philosophical Issues: A Supplement to Nous*, 28: 221–54.

Mancuso, K., et al. (2009), Gene therapy for red–green colour blindness in adult primates, *Nature*, 461(7265): 784–7.

Marcus, G. and Davis, E. (2019), *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York: Pantheon Books.

Maslow, A. H. (1943), A theory of human motivation, *Psychological Review*, 50(4): 370–96.

Maslow, A. H. (1954), *Motivation and Personality*, New York: Harper.

Maslow, A. H. (1987), *Motivation and Personality (3rd Ed)*, London, UK: Longman.

Maunsell, J. H. and Treue, S. (2006), Feature-based attention in visual cortex, *Trends in Neurosciences*, 29(6): 317–22.

Meuwese, J. D. I., Post, R. A. G., Scholte, H. S., and Lamme, V. A. F. (2013), Does perceptual learning require consciousness or attention?, *Journal of Cognitive Neuroscience*, 25(10): 1579–96.

Mindt, G. and Montemayor, C. (2020), A Roadmap for artificial general intelligence: Intelligence, knowledge, and consciousness, *Mind and Matter*, 18(1): 9–37.

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., and Linos, E. (2016), Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health, *JAMA Internal Medicine*, 176(5): 619–25.

Miracchi, L. (2015), Competence to know, *Philosophical Studies*, 172(1): 29–56.

Mitchell, D. G. V. and Greening, S. G. (2012), Conscious perception of emotional stimuli: Brain mechanisms, *The Neuroscientist*, 18(4): 386–98.

Mitchell, M. (2021), Why AI is harder than we think, *arXiv.2104.12871* [cs.AI].

Mnih, V., et al. (2015), Human-level control through deep reinforcement learning, *Nature*, 518(7540): 529–33.

Moe-Behrens, G. H. G. (2013), The biological microprocessor, or how to build a computer with biological parts, *Computational and Structural Biotechnology Journal*, 7(8): 1–18.

Mole, C. (2011), *Attention Is Cognitive Unison: An Essay in Philosophical Psychology*, New York: Oxford University Press.

Montemayor, C. (2002), *La Unificación Conceptual de los Derechos Humanos (The Conceptual Unification of Human Rights)*, Mexico City, Mexico: Porrúa.

Montemayor, C. (2013), *Minding Time: A Philosophical and Theoretical Approach to the Psychology of Time*, The Netherlands: Brill.

Montemayor, C. (2014), Law, action, and collective agency: The cognitive integration approach. In E. Villanueva (ed.), *Law and the Philosophy of Action: Social, Political and Legal Philosophy, Volume 3*, Brill: Rodopi Philosophical Studies (pp. 221–45).

Montemayor, C. and Haladjian, H. H. (2015), *Consciousness, Attention, and Conscious Attention*, Cambridge, MA: MIT Press.
Montemayor, C. (2016), Memory: Epistemic and phenomenal traces. In S. Gross and S. Ostovich (eds.), *Time and Trace: Multidisciplinary Investigations of Temporality*, Leiden, The Netherlands: Brill (pp. 215–31).
Montemayor, C. (2017), Time perception and agency: A dual model. In I. Phillips (ed.), *The Routledge Handbook of Philosophy of Temporal Experience*, New York, NY: Routledge (pp. 201–12).
Montemayor, C. and Haladjian, H. H. (2017), Perception and cognition are largely independent, but still affect each other in systematic ways: Arguments from evolution and the consciousness-attention dissociation, *Frontiers in Psychology*, 8(40).
Montemayor, C. (2018), Consciousness and memory: A transactional approach, *Essays in Philosophy*, 19(2): article 5.
Montemayor, C. and Cáceres, E. (2019), Agency and legal responsibility: Epistemic and moral considerations, *Problema: Anuario de Filosofía y Teoría del Derecho*, 13: 99–127.
Montemayor, C. (2019a), Inferential integrity and attention, *Frontiers in Psychology: Consciousness Research*, 10: 2580.
Montemayor, C. (2019b), Early and late time perception: On the narrow scope of the Whorfian hypothesis, *Review of Philosophy and Psychology*, 10(1): 133–54.
Montemayor, C. (2021), Language and intelligence, *Minds and Machines*, 31: 471–86.
Moore, G. E. (1903|1968), *Principia Ethica*, Cambridge: Cambridge University Press.
Moran, R. (2001), *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, NJ: Princeton University Press.
Moravec, H. P. (1988), *Mind Children: The Future of Robot and Human Intelligence*, Cambridge, MA: Harvard University Press.
More, M. and Vita-More, N. (2013), *The Transhumanist Reader*, Oxford, UK: Wiley-Blackwell.
Morton, A. (2012), *Bounded Thinking: Intellectual Virtues for Limited Agents*, Oxford, UK: Oxford University Press.
Moyn, S. (2010), *The Last Utopia: Human Rights in History*, Cambridge, MA: Belknap, Harvard University Press.
Mudrik, L., Faivre, N., and Koch, C. (2014), Information integration without awareness, *Trends in Cognitive Sciences*, 18(9): 488–96.
Mulckhuyse, M. and Theeuwes, J. (2010), Unconscious attentional orienting to exogenous cues: A review of the literature, *Acta Psychologica*, 134(3): 299–309.
Murdoch, I. (1970), *The Sovereignty of Good*, New York: Routledge and Kegan Paul.
Nagel, T. (1974), What is it like to be a bat? *The Philosophical Review*, 83(4): 435–50.

Nakajima, M., Schmitt, L. I., and Halassa, M. M. (2019), Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway, *Neuron*, 103(3): 445–58.

Newell, B. R. and Shanks, D. R. (2014), Unconscious influences on decision making: A critical review, *Behavioral and Brain Sciences*, 37(01): 1–19.

Nussbaum, M. (1988), Nature, function, and capability: Aristotle on political distribution. In J. Annas and R. H. Grimm (eds.), *Oxford Studies in Ancient Philosophy*, Oxford, UK: Oxford University Press (Supplementary Volume), 6, 145–84.

Nussbaum, M. (2011), *Creating Capabilities: The Human Development Approach*, Cambridge, MA: Harvard University Press.

Nussbaum, M. (2020), The capabilities approach and the history of philosophy. In Enrica Chiappero-Martinetti, Siddiqur Osmani, and Mozaffar Qizilbash (eds.), *The Cambridge Handbook of the Capability Approach*, Cambridge: Cambridge University Press (pp. 13–39).

Nyholm, S. (2020), *Humans and Robots: Ethics, Agency and Anthropomorphism*, London: Rowman & Littlefield.

Oddie, G. (2005), *Value, Reality, & Desire*, New York: Oxford University Press.

O'Mara, M. (2019), *The Code: Silicon Valley and the Remaking of America*, New York: Penguin.

Pagallo, U. (2013), *The Laws of Robots: Crimes, Contracts, and Torts*, New York: Springer.

Parfit, D. (1984), *Reasons and Persons*, Oxford: Oxford University Press.

Pauers, M. J., Kuchenbecker, J. A., Neitz, M., and Neitz, J. (2012), Changes in the colour of light cue circadian activity, *Animal Behaviour*, 83(5): 1143–51.

Paul, L. (2014), *Transformative Experience*, New York: Oxford University Press.

Pavur, J. and Knerr, C. (2019), GDPArrrr: Using privacy laws to steal identities, *Black Hat USA* (8): 8.

Payne, B. K. (2001), Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon, *Journal of Personality and Social Psychology*, 81(2): 181–92.

Pessoa, L. (2005), To what extent are emotional visual stimuli processed without attention and awareness?, *Current Opinion in Neurobiology*, 15(2): 188–96.

Pessoa, L. (2008), On the relationship between emotion and cognition, *Nature Reviews Neuroscience*, 9(2): 148–58.

Pessoa, L. (2013), *The Cognitive-Emotional Brain: From Interactions to Integration*, Cambridge, MA: MIT Press.

Pessoa, L., Kastner, S., and Ungerleider, L. G. (2002), Attentional control of the processing of neutral and emotional stimuli, *Cognitive Brain Research*, 15(1): 31–45.

Pettigrew, R. (2015), Transformative experience and decision theory, *Philosophy and Phenomenological Research*, 91(3): 766–74.

Pew Research Center. (2015), *The Smartphone Difference*. Retrieved from http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

Picard, R. W. (1997), *Affective Computing*, Cambridge, MA: MIT Press.
Picard, R. W. (2002a), Affective medicine: Technology with emotional intelligence, *Studies in Health Technology and Informatics*, 80: 69–83.
Picard, R. W. (2002b), What does it mean for a computer to "have" emotions?. In R. Trappl, P. Petta and S. Payr (eds.), *Emotions in Humans and Artifacts*, Cambridge, MA: MIT Press (pp. 213–36).
Picard, R. W. (2007), Toward machines with emotional intelligence. In G. Matthews, M. Zeidner, and R. D. Roberts (eds.), *The Science of Emotional Intelligence: Knowns and Unknowns*, New York: Oxford University Press (pp. 396–418).
Picard, R. W., Vyzas, E., and Healey, J. (2001), Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10): 1175–91.
Posner, M. I., Snyder, C. R. and Davidson, B. J. (1980), Attention and the detection of signals, *Journal of Experimental Psychology*, 109(2): 160–74.
Pratt, G. A. (2015), Is a Cambrian explosion coming for robotics?, *The Journal of Economic Perspectives*, 29(3): 51–60.
Putnam, H. (1975), The meaning of "Meaning". In K. Gunderson (ed.), *Language, Mind and Knowledge* (Minnesota Studies in the Philosophy of Science, Volumes VII), Minneapolis: University of Minnesota Press (pp. 131–93).
Pylyshyn, Z. W. (1980), The "causal power" of machines, *Behavioral and Brain Sciences*, 3(3): 442–4.
Pylyshyn, Z. W. (1984), *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: MIT Press.
Pylyshyn, Z. W. (1999), Is vision continuous with cognition? The case for cognitive impenetrability of visual perception, *Behavioral and Brain Sciences*, 22(3): 341–65; discussion 366–423.
Pylyshyn, Z. W. (2000), Situating vision in the world, *Trends in Cognitive Sciences*, 4(5): 197–207.
Ratner, P. (2017), Automation nightmare: Philosopher warns we are creating a world without consciousness, *Big Think*, February 11. http://bigthink.com/paul-ratner/automation-nightmare-we-might-be-headed-for-a-world-without-consciousness.
Rawls, J. (1971), *A Theory of Justice*, Cambridge, MA: Harvard University Press.
Rawls, J. (1985), Justice as fairness: Political not metaphysical, *Philosophy and Public Affairs*, 14(3): 223–51.
Rawls, J. (2001), *Justice as Fairness: A Restatement*, Cambridge, MA: Harvard University Press.
Reese, B. (2018), *The Fourth Age: Smart Robots, Conscious Computers, and the Future of Humanity*, New York, NY: Atria Books.
Reginster, B. (2004), Self-knowledge, responsibility, and the third person, *Philosophy and Phenomenological Research*, 69(2): 433–9.
Reggia, J. A. (2013), The rise of machine consciousness: Studying consciousness with computational models, *Neural Networks*, 44: 112–31.

Reichardt, D. M. (2007), A definition approach for an "Emotional Turing Test". In A. C. R. Paiva, R. Prada, and R. W. Picard (eds.), *Affective Computing and Intelligent Interaction* (Volume 4738), Berlin, Germany: Springer Berlin Heidelberg (pp. 716–17).

Richard, M. (2019), Is reasoning a form of agency?. In M. Balcerak Jackson and B. Balcerak Jackson (eds.), *Reasoning: Essays on Theoretical and Practical Thinking*, New York: Oxford University Press (pp. 91–100).

Richtel, M. and Dougherty, C. (2015, September 2), Google's driverless cars run into problem: Cars with drivers, *The New York Times*, p. A1. Retrieved from http://nyti.ms/1LRy9MF.

Riedl, M. O. and Harrison, B. (2015), *Using stories to teach human values to artificial agents*, Paper presented at the 2nd International Workshop on AI, Ethics, and Society. http://www.aaai.org.

Rinesi, M. (2015), The price of the Internet of Things will be a vague dread of a malicious world. Retrieved from IEET.org website: http://ieet.org/index.php/IEET/more/rinesi20150925.

Rinne, P., et al. (2018), Motor dexterity and strength depend upon integrity of the attention-control system, *Proceedings of the National Academy of Sciences of the United States of America*, 115(3): E536–E545.

Robinson, H., MacDonald, B. and Broadbent, E. (2014), The role of healthcare robots for older people at home: A review, *International Journal of Social Robotics*, 6(4): 575–91.

Rochat, P. (2003), Five levels of self-awareness as they unfold early in life, *Consciousness and Cognition*, 12(4): 717–31.

Rochat, P. and Striano, T. (2000), Perceived self in infancy, *Infant Behavior and Development*, 23(3–4): 513–30.

Rosen, M. (2012), *Dignity: Its History and Meaning*, Cambridge, MA: Harvard University Press.

Rousseau, J. J. (1997), *The Social Contract*, V. Gurevittch (ed. and trans.), Cambridge: Cambridge University Press.

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking: Random House.

Santoni de Sio, F. and van den Hoven, J. (2018), Meaningful human control over autonomous systems: A philosophical account, *Frontiers in Robotics and AI*, February 28, 5: 15.

Sapolsky, R. (2003), A bozo of a baboon, *Edge*, Conversation, June 2, 2003.

Sapolsky, R. (2016), To understand Facebook, study Capgras syndrome: This mental disorder gives us a unique insight into the digital age, *Nautilus*, 42, November 10.

Schechtman, M. (1996), *The Constitution of Selves*, Ithaca: Cornell University Press.

Schechtman, M. (2014), *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*, Oxford: Oxford University Press.

Scholl, B. J. (2001), Objects and attention: The state of the art, *Cognition*, 80(1–2): 1–46.

Searle, J. (1969), *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: Cambridge University Press.
Searle, J. R. (1980), Minds, brains, and programs, *Behavioral and Brain Sciences*, 3(3): 417–24.
Searle, J. R. (1985), *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge: Cambridge University Press.
Searle, J. R. (1998), *Mind, Language, and Society: Philosophy in the Real World* (1st ed.), New York, NY: Basic Books.
Sen, A. (1993), Capability and well-being. In Martha Nussbaum and Amartya Sen (eds.), *The Quality of Life*, Oxford: Clarendon Press (pp. 30–53).
Sen, A. (1999), *Development as Freedom*, New York: Knopf.
Seth, A. K. and Baars, B. J. (2005), Neural Darwinism and consciousness, *Consciousness and Cognition*, 14(1): 140–68.
Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008), Measuring consciousness: Relating behavioural and neurophysiological approaches, *Trends in Cognitive Sciences*, 12(8): 314–21.
Sharkey, A. (2020), Can we program or train robots to be good?, *Ethics and Information Technology*, 22: 283–95.
Shattuck, R. (1996), *Forbidden Knowledge: From Prometheus to Pornography*, New York: Harvest, Harcourt Brace & Company.
Shibata, T. (2004), An overview of human interactive robots for psychological enrichment, *Proceedings of the IEEE*, 92(11): 1749–58.
Siegel, S. (2017), *The Rationality of Perception*, New York: Oxford University Press.
Silver, D., et al. (2016), Mastering the game of go with deep neural networks and tree search, *Nature*, 529(7587): 484–9.
Silvers, A., Wasserman, D. and Mahowald, M. B. (1998), *Disability, Difference, Discrimination: Perspectives on Justice in Bioethics and Public Policy*, Lanham, MD: Rowman & Littlefield Publishers.
Simonite, T. (2017), Humans can't expect AI to just fight fake news for them, *Wired*, June 15, 2017. https://www.wired.com/story/fake-news-challenge-artificial-intelligence/
Smith, J. E. H. (2011), *Divine Machines: Leibniz and the Sciences of Life*, Princeton: Princeton University Press.
Smithies, D. (2019), *The Epistemic Role of Consciousness*, Oxford: Oxford University Press.
Song, S., Zilverstand, A., Song, H., d'Oleire Uquillas, F., Wang, Y., Xie, C., Cheng, L. and Zou, Z. (2017), The influence of emotional interference on cognitive control: A meta-analysis of neuroimaging studies using the emotional Stroop task, *Scientific Reports*, 7: 2088.
Sosa, E. (2007), *A Virtue Epistemology: Apt Belief and Reflective Knowledge (Volume I)*, New York: Oxford University Press.

Sosa, E. (2015), *Judgment & Agency*, New York: Oxford University Press.
Spering, M. and Carrasco, M. (2015), Acting without seeing: Eye movements reveal visual processing without awareness, *Trends in Neurosciences*, 38(4): 247–58.
Srinivasan, A. (2020), Radical externalism, *The Philosophical Review*, 129(3): 395–431.
Strawson, G. (2008), Against narrativity. In G. Strawson (ed.), *Real Materialism and Other Essays*, Oxford: Oxford University Press (pp. 189–207).
Strawson, P. F. (1962), Freedom and resentment, *Proceedings of the British Academy*, 48: 1–25.
Stoljar, D. (2019), In praise of poise. In. A. Pautz and D. Stoljar (eds.), *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*, Cambridge, MA: MIT Press (pp. 511–36).
Tamietto, M. and de Gelder, B. (2010), Neural bases of the non-conscious perception of emotional signals, *Nature Reviews Neuroscience*, 11(10): 697–709.
Tononi, G. and Koch, C. (2008), The neural correlates of consciousness: An update, *Annals of the New York Academy of Sciences*, 1124: 239–61.
Treisman, A. (1998), Feature binding, attention, and object perception, *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 353(1373): 1295–306.
Treisman, A. (2006), How the deployment of attention determines what we see, *Visual Cognition*, 14(4–8): 411–43.
Treisman, A. and Gelade, G. (1980), A feature-integration theory of attention, *Cognitive Psychology*, 12(1): 97–136.
Tsuchiya, N. and Adolphs, R. (2007), Emotion and consciousness, *Trends in Cognitive Sciences*, 11(4): 158–67.
Tucker, A. M., Feuerstein, R., Mende-Siedlecki, P., Ochsner, K. N., and Stern, Y. (2012), Double dissociation: Circadian off-peak times increase emotional reactivity; aging impairs emotion regulation via reappraisal, *Emotion*, 12(5): 869–74.
Turing, A. M. (1950), Computing machinery and intelligence, *Mind*, 59(236): 443–60.
Turkle, S. (2005/1984), *The Second Self: Computers and the Human Spirit* (20th anniversary ed.), Cambridge, MA: MIT Press.
Turkle, S. (2007), Authenticity in the age of digital companions, *Interaction Studies*, 8(3): 501–17.
Turkle, S., Taggart, W., Kidd, C. D., and Dasté, O. (2006), Relational artifacts with children and elders: The complexities of cybercompanionship, *Connection Science*, 18(4): 347–61.
Turner, J. (2019), *Robot Rules: Regulating Artificial Intelligence*, Cham, Switzerland: Palgrave Macmillan.
van Boxtel, J. J. A., Tsuchiya, N., and Koch, C. (2010), Consciousness and attention: On sufficiency and necessity, *Frontiers in Psychology*, 1(217). doi: 10.3389/fpsyg.2010.00217.
Unger, P. (1996), *Living High & Letting Die: Our Illusion of Innocence*, New York: Oxford University Press.

Vargas, M. (2013), *Building Better Beings: A Theory of Moral Responsibility*, New York: Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Kaiser, L., and Polosukhin, I. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, 5998–6008.

Voevodsky, V. (2014), *The Origins and Motivations of Univalent Foundations: A Personal Mission to Develop Computer Proof Verification to Avoid Mathematical Mistakes*. https://www.ias.edu/ideas/2014/voevodsky-origins.

Wagner, A. and Rosen, W. (2014), Spaces of the possible: Universal Darwinism and the wall between technological and biological innovation, *Journal of the Royal Society Interface*, 11(97).

Wallach, W. and Allen, C. (2009), *Moral Machines: Teaching Robots Right from Wrong*, New York: Oxford University Press.

Wallach, W. (2015), *A Dangerous Master: How to Keep Technology from Slipping beyond Our Control*, New York, NY: Basic Books.

Wang, R., Lehman, J., Clune, J. and Stanley, K. O. (2019), Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions, *Arxiv*: 1901.01753.

Watzl, S. (2017), *Structuring Mind: The Nature of Attention and How It Shapes Consciousness*, New York: Oxford University Press.

Weber, M. (1964), The types of authority and imperative co-ordination. In A. M. Henderson and T. Parsons (eds. and trans.), *The Theory of Social and Economic Organization*, New York, NY: The Free Press (pp. 524–9).

Weintrauboct, K. (2012, October 15), But how do you really feel? Someday the computer may know, *The New York Times*. Retrieved from http://www.nytimes.com/2012/10/16/science/affective-programming-grows-in-effort-to-read-faces.html.

Weiskrantz, L. (2009), *Blindsight: A Case Study Spanning 35 Years and New Developments* (2nd ed.), Oxford: Oxford University Press.

Weisz, E. and Zaki, J. 2018. Motivated empathy: a social neuroscience perspective, *Current Opinion in Psychology*, 24, 67–71.

Wiener, N. (1950), *The Human Use of Human Beings*, Boston, MA: Houghton Mifflin.

Williams, B. (1973), The Makropulos case: Reflections on the tedium of immortality. In B. Williams (ed.), *Problems of the Self*, Cambridge: Cambridge University Press (pp. 82–100).

Williamson, T. (In press), Justifications, excuses, and sceptical scenarios. In J. Dutant (ed.), *The New Evil Demon*, Oxford: Oxford University Press.

Winczewski, L. A., Bowen, J. D. and Collins, N. L. (2016), Is empathic accuracy enough to facilitate responsive behavior in dyadic interaction? Distinguishing ability from motivation, *Psychological Science*, 27(3): 394–404.

Wittgenstein, L. (1922/1974), *Tractatus Logico-Philosophicus*, D. F. Pears and B. F. McGuiness (trans.), New York: Routledge.

Wittgenstein, L. (1967), *Lectures and Conversations on Aesthetics, Psychology, and Religious Belief*, C. Barrett (ed.), Oxford: Basil Blackwell.

WNYC Studios (Producer) (2012, January), *Talking to machines*. Retrieved from http://www.radiolab.org/story/137407-talking-to-machines/.

Wolf, S. (2010), *Meaning in Life and Why It Matters*, Princeton, NJ: Princeton University Press.

Wooldridge, M. (2020), *The Road to Conscious Machines: The Story of AI*, New Orleans, LA: Pelican.

Wright, C. (2014), Comment on Paul Boghossian, "What is Inference?", *Philosophical Studies*, 169(1): 17–37.

Wu, W. (2011), Attention as selection for action. In C. Mole, D. Smithies and W. Wu (eds.), *Attention: Philosophical and Psychological Essays*, Oxford: Oxford University Press (pp. 97–116).

Wu, W. (2013), Mental action and the threat of automaticity. In A. Clark, J. Kiverstein and T. Vierkant (eds.), *Decomposing the Will*, Oxford: Oxford University Press (pp. 244–61).

Wu, W. (2014), *Attention*, New York: Routledge.

Yamakawa, H. (2019), Peacekeeping conditions for an artificial intelligence society, *Big Data and Cognitive Computing*, 3: 34.

Yang, H., Shao, L., Zheng, F., Wang, L., and Song, Z. (2011), Recent advances and trends in visual tracking: A review, *Neurocomputing*, 74(18): 3823–31.

Yeh, S. and Chen, I. (1999), Is early visual processing attention impenetrable?, *Behavioral and Brain Science*, 22: 400.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., and Chapman, M. (1992), Development of concern for others, *Developmental Psychology*, 28(1): 126–36.

Zaki, J. (2017), Moving beyond stereotypes of empathy, *Trends in Cognitive Science*, 21(2): 59–60.

Zeimbekis, J. and Raftopoulos, A. (eds.), (2015), *The Cognitive Penetrability of Perception: New Philosophical Perspectives*, New York: Oxford University Press.

Zmigrod, S., Spapé, M., and Hommel, B. (2009), Intermodal event files: Integrating features across vision, audition, taction, and action, *Psychological Research*, 73(5): 674–84.

Zuboff, S. (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, New York, NY: Public Affairs.

Index