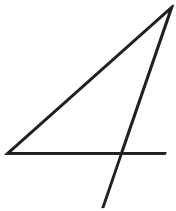

The Madness of Algorithms
Aberration and Unreasonable Acts

There are crises of reason in strange complicity with what the world calls crises of madness.

—Jacques Derrida, "Cogito and the History of Madness"

Reason demanded this step. In our science we have run up against the limits of what is knowable. . . . Our science has become a horror, our research dangerous, our knowledge lethal. All that is left for us physicists is to capitulate before reality.

—Friedrich Dürrenmatt, *The Physicists*

Errant Algorithms

There are moments when the everyday background hum of algorithms in society seems to break the surface and become something louder and more insistent. Such moments are often debated publicly as times when algorithms have become mad, or when they have departed from their rational logics and become frenzied, demented, and thoroughly unreasonable. In October 2016, when the British pound lost 6.1 percent of its value in a matter of seconds, the currency trades conducted autonomously by algorithms were reported to have precipitated a "flash crash," an eruption of irrational actions beyond the control of human traders.[1] Such apparent moments of the madness of algorithms, however, are not limited to the vicissitudes of the financial markets. In March 2016, Microsoft researchers experimented with a social media chatbot—named Tay and modeled on the attributes of a teenage girl—training her to communicate in a humanlike way by exposing her to the inputs of Twitter data streams. Within a few short hours, and in response to her Twitter

learning milieu, Tay was posting racist and sexist messages, denying the holocaust and boasting to the police of her prolific drug habit.[2] Tay's deep learning algorithms were widely reported to be in meltdown, spiraling out of control in a descent into madness. The Microsoft experiment was finally halted when Tay messaged her two hundred thousand social media followers with "You are too fast, please take a rest," sending the message over and over again in a repetitious frenzy. The public's sense of dangerous algorithms becoming unreasonable and doing harm seemed to reach fever pitch in 2017, when YouTube's machine learning recommendation algorithms targeted automated content to children, playing videos with violent content on the YouTube Kids platform, such as the popular cartoon character Peppa Pig being tortured at the dentist, drinking bleach, and eating bacon.[3] In each of these moments, when the harms and dangers of algorithms seem to press themselves into public attention, what is most often discussed is a kind of algorithmic madness, a frenzied departure from reason.

As societies have responded to the errancy of algorithms, a specific ethicopolitical framing of the problem has emerged over time. Overwhelmingly, the public parlance around algorithmic ethics has placed emphasis on limiting the excesses of algorithms, controlling their impulses, and reining in their capacity for frenzied action in the world. The search for an encoded ethics for algorithms has tended to annex madness as an aberration that must be subjected to correction. In annexing the madness of algorithms as irrational aberration and departure from reason, such public interventions sustain the promise that algorithms could be rendered governable by appropriate thresholds of reasonable and unreasonable actions. Popular debates have depicted algorithms as "weapons of math destruction" that must be "taught right from wrong" in a world where one must "take care that ethical and not evil algorithms help win elections."[4] The propensity of the algorithm for madness, in short, has become a kind of limit point on its otherwise apparently reasonable place in our world. They can stay if they behave themselves. A kind of bounded rationality prevails, in which the boundary delineating good from evil in algorithmic decisions is one marked by reason versus unreason, rationality versus irrationality.[5] This boundary is breached, so it appears at least, when the algorithm acts in a way that was somehow not anticipated. Thus, for example, when the "kill switch" of a human intervention is activated—whether this is in an autonomous weapons system, artificial intelligence for speech, or a video recommendation system for children—this marks a threshold of something like madness, understood as a departure from reasoned logic.[6]

In this chapter I am concerned with how this common notion of the mad-

ness of algorithms as aberration has crucially overlooked the extent to which the rationality of algorithms is built on—indeed positively embraces and harnesses—the power of unreason. Such accounts of moral panic amid the madness of algorithms have had two significant effects. First, they have forgotten the place of notions of madness within histories of algorithmic rationality, where unreason is intrinsic to the computational logic. And second, they have underplayed the role of capricious incalculability within our twenty-first-century modes of algorithmic decision. Put simply, the common refrain of an encoded ethics for algorithms depends on a threshold that divides rational action from madness. The public discussion of the evils and harms of algorithms is focused on the regulation of that very threshold thought to separate rational from irrational action. Crucially, though, this threshold is, as all border lines are, the line that both pulls together and demarcates madness from all that is reasonable in the world. For Jacques Derrida, the division is a "dissension" or a "self dividing action," which both "links and separates reason and madness."[7] Discussing Foucault's treatment of the history of madness, Derrida cautions that "one cannot speak of madness except in relation to 'that other form of madness,' that is except in relation to reason."[8]

Taking seriously the conjoined histories of the ideas of reason and madness, I propose that one cannot speak of the madness of the algorithm except in relation to the form of reason the algorithm embodies. While the contemporary moral panic at each moment of the madness of algorithms urges us to police ever more vigilantly the line between reasonable and unreasonable actions, understood as a dissension, this line is precisely the condition of possibility of algorithmic rationality. Algorithms cannot be controlled via a limit point at the threshold of madness because the essence of their logic is to generate that threshold, to adapt and to modulate it over time. In short, my argument is that when algorithms appear to cross a threshold into madness, they in fact exhibit significant qualities of their form of rationality. Understood in this way, the appearance of a moment of madness is a valuable instant for ethicopolitics because this is a moment when algorithms give accounts of themselves. Contra the notion that transparency and the opening of the technological black box secures the good behavior of algorithms, the opacity and clouded action exhibited in the excesses and frenzies of algorithms yield a different kind of fidelity to the logic.

Let me make this concept of clouded action a little more concrete. Though the surfacing of violent images of beloved cartoon characters for preschool children appears as an undoubted algorithmic aberration, the computer science accounts of the development of deep neural networks for the YouTube

video recommendation system exhibit a rationality that is consistent with outputting violent content. The algorithm designers describe the "dynamic corpus" of YouTube video content and explain their improved deep learning algorithms, which are "responsive enough to model newly uploaded content as well as the latest actions taken by the user."[9] Understood as a model that is trained on one billion parameters, the apparent madness of the recommendation of violent cartoons is explicitly part of, and not an aberration from, the rationality of the deep neural network. Indeed, the optimization of these algorithms is achieved in part through exposure to the new and emerging uploaded content, which is afforded greater weight in the model than user-specific data histories. At the level of the algorithm's logic—its mode of learning, internal weights, architecture, parameters, training data, and so on—the funneling of a frenzy of violent videos is entirely consistent with its rationality.

As I propose throughout this book, when viewed from the specific propositional arrangements of the algorithm, particular actions that might appear as errors or aberrations are in fact integral to the algorithm's form of being and intrinsic to its experimental and generative capacities. I am advocating that we think of algorithms as capable of generating unspeakable things precisely because they are geared to profit from uncertainty, or to output something that had not been spoken or anticipated. As machine learning algorithms increase their capacity to learn from raw unlabeled data streams, the unseen and unspoken become precisely the generative materials of algorithmic decisions. Such proximity between violent algorithmic outputs and the attributes of the data stream does mean that one has to think some heretical thoughts on ethics. The racist hate and misogyny of Tay; the funneling of far-right media to specific voter clusters using deep neural networks; the terrible consequences of image misrecognition in a drone strike—these are not instances when the algorithm has become crazed or frenzied. Rather, the learned action is a reasonable output given the extracted features of the data inputs. Of course, this is not a less troubling situation than the one in which some controls are sought on the worst excesses of the algorithm. On the contrary, it is all the more political, and all the more difficult, because that which could never be controlled—change, wagers, impulses, inference, intuition—becomes integral to the mode of reasoning. To be clear, my intention is not to avoid the need for critical response to the harms of algorithms. Far from it, my point is that violence and harm are not something that can be corrected out of an otherwise reasonable calculus.

In what follows in this chapter, I begin by resituating the genealogy of madness and reason in twentieth-century cybernetics, focusing specifically

on the work of Norbert Wiener and the double enrollment of madness and rational forms of control into algorithmic systems. I then flesh out an alternative way of thinking about the madness of algorithms, reformulating the problem from one where algorithms might lose their hold on rationality to one where algorithms precisely require forms of unreason to function and to act. I am thereby amplifying attention to another kind of harm that does not reside beyond the threshold of something like the edge of reasoned action. This different kind of harm dwells in the algorithm's inability to ever embody the "madness of decision."[10] To live with the madness of decision is to acknowledge and take responsibility for the impossibility of ever binding the action to a full body of knowledge. Decisions are mad because they can never know fully the consequences and effects of their own making. To decide is to confront the impossibility of the resolution of difficulty; it is madness in the specific sense that it has no unified grounds.[11] With contemporary algorithms, decisions are being made at the limit of what could be known, and yet there is no responsibility for the unknown consequences of the decision. The madness of the decision is disavowed by the single output of the algorithm, and this disavowal is a potential horror and a danger. In the second half of the chapter, I discuss one specific set of algorithms—random decision forests—that have become the algorithms of choice across many domains, particularly in national security and border controls. The random decision forest algorithm grows multiple decision trees on random subsets of data and "takes a vote on their predictions."[12] It is an algorithmic proposition of What comes next? that takes place as a calculation amid incalculability, mobilizing chance and the splitting of agency, sometimes with lethal effects on human life.

Cybernetics and Unreason

In a 1950 essay, "Atomic Knowledge of Good and Evil," the cybernetician Norbert Wiener expresses his concern for the moral responsibility of science in the face of what he calls the "dangerous possibilities" of atomic weapons.[13] Reflecting on the nature of the roles of the mathematicians and physicists at Los Alamos, Wiener is anxious that "the new centralized atmosphere of the age to come is likely to lead society in the direction of 'Big Brother'" and toward a "future fascism." A full seventy years before scientists pointed to the dangers of a descent into fascism precipitated by the mathematical sciences of algorithms, the cyberneticians were writing publicly in newspapers and journals to express the view that "when the scientist becomes the arbiter of life and death" the locus of "moral responsibility" belongs properly to the reasoning scientist to prevent "the decay and disruption of society" and to ensure that

science "is used for the best interests of humanity."[14] Indeed, in a 1947 letter to *Atlantic Monthly*, Wiener replies publicly to a request from government military officials for copies of his mathematics papers. Concerned that withdrawing his science from the service of war might be "shutting the door after the horse has become classified," Wiener believes that the locus of control for a dangerous science must reside in the reasoning human scientist. For Wiener, the rise of computing machines posed a threat defined not so much in terms of the madness of the autonomous machine as in terms of what he called the "crazed and misguided" irrationality of humans. "The automatic machine is not frightening because of any danger that it may achieve autonomous control over humanity," writes Wiener, but "its real danger" is that it may be "used by a human being to increase their control over the rest of the human race."[15] Wiener's letters and essays are replete with his anxiety that the science driving human progress may simultaneously be enrolled by sovereign or corporate powers and deployed for violence and war. For Wiener, the moral responsibility of the mathematical and physical sciences, particularly in their service of state power, resides in governing the line between reason and unreason. This distinction between reason and unreason, however, exposes something of the paradox of the place of madness in the cybernetic histories of algorithmic rationality.

First, the rise of cybernetics is closely intertwined with notions of overcoming the dangerous fallibilities of human judgment. Reimagining the human decision as a series of relays that could be modeled for optimization, Cold War cybernetics was geared to "tame the terrors of decisions too consequential to be left to human reason alone."[16] Understood in this way, the definiteness and conclusiveness of the algorithm as decision maker was thought to supply a rational safeguard against the madness of mutually assured destruction.[17] As Wiener wrote in his essay on "the new concept of the machine," the binary logic of computation structured decisions via a series of "switches," each offering a "decision between two alternatives" in a "logical machine."[18] For Wiener, the binary branching of the switch embodied what he called "two truth values," where the transmitted information was either "true or false." Here the logic of algorithms explicitly condenses the indefiniteness of data to a series of switches that must always be a binary choice of yes/no, true/false.

The branching that we see in classic decision tree algorithms, and in contemporary random forest algorithms, carries the traces of a cybernetic history, when the switch marked the calculation of truth and falsity. Amid the terrors of thermonuclear war, the promise of branching algorithmic decisions was that they could supply to political decision makers a means of avoiding the

human propensity for frenzied miscalculation. At every instant of a branching decision, the logical machine rendered an output of truth or falsity. In this way, the twentieth-century rise of rational choice in foreign policy and behaviorism in the social sciences imagined a world in which algorithms limited the madness of a geopolitics on the edge of destruction. This first aspect of the place of unreason in genealogies of algorithmic decisions is significant to our contemporary moment because it is a reminder that algorithms are not devices that exist outside notions of the normative, norms, anomalies, and pathologies. On the contrary, the rise of algorithms in governing difficult and intractable state decisions makes them interior to defining what normalities and pathologies could be in a society. Just as the early neural network algorithms were embraced as a means of instilling reason into otherwise potentially dangerous human judgment, however, they also brought into being a new mode of harnessing the unknown and the unpredictable.

Second, at the same time as algorithms were thought to tame human irrationality, a specific orientation to madness was actively incorporated into the logics of algorithms. In Orit Halpern's book *Beautiful Data*, she gives a devastating account of how ideas about the psychotic and the neurotic were integral to cybernetic logics. "What has been erased from the historical record in our present," writes Halpern, "is the explicit recognition in the aftermath of World War II and the start of the Cold War that rationality was not reasonable."[19] In a detailed engagement with Warren McCulloch and Walter Pitts's development of the neural net, Halpern proposes that the logic of psychosis was crucial to the reformulation of computational rationality. Tracing the genealogy of neural nets as technical instantiations of the brain's functions, Halpern demonstrates the extent to which "neurotic, psychotic, and schizoid circuits proliferated in the diagrams of cybernetics."[20]

The insights of those who have studied the schizoid and paranoid relations of computation are critical to the contemporary ethicopolitical relations being forged with and through algorithms.[21] Far from representing a departure from rationality and entry into psychotic turmoil, the actions of algorithms are never far from their conjoined histories with psychosis, neurosis, trauma, and the imagination of the brain as a system. The double enfolding of unreason into cybernetic thought—as the unreasonableness of human decision makers in conditions of war and insecurity, and as models of psychosis underpinning the segmented cognition of the neural net—serves as something of a corrective to our contemporary societal debates on the responsibility of science for moments of algorithmic madness. Put simply, the history of algorithmic rationality is not separable from genealogies of madness. It is time to think difficult

thoughts: algorithms are always already unreasonable. When one hears calls for new controls to *limit* the potentially dangerous actions of the algorithm, it is perhaps worth recalling that algorithms historically are *limit* devices. That is to say, they actively generate new forms of what it means to be normal or abnormal, just as they mark new boundaries of rationality and unreason. To capitalize on uncertainty—whether in warfare or in the commercial world of risk—algorithms dwell productively with emergent phenomena and incalculable surprises.[22] The actions of the contemporary neural net, for example, finds its condition of possibility in the twinned imperatives of limiting the impulses of human decision makers while embracing and harnessing the impulses of models of the brain as nervous circuits.

To be clear, contemporary algorithms are oriented toward limiting the unanticipated actions of humans while generating a whole new world of their own unanticipated and unreasonable actions. This means, for example, that an algorithm will decide the threshold of whether a person's presence on a city street is normal or anomalous, and yet it will generate this threshold via experimental methods that enroll errancy and capricious action into the form of reason. In the section that follows, I revisit the problem of madness to reformulate the algorithm's relation to unreason.

Madness and Classification

Throughout this book I urge some caution about the treatment of the ethical and political questions posed by algorithms as novel or unprecedented. Indeed, I suggest that the advent of algorithms as decision makers in our societies has restated what are in fact perennial philosophical debates about the nature of ethical and political life. Among these perennial debates is the question of whether and how one controls the impulses or actions of an entity that appears to us as beyond human reason. It is in the philosophies of human madness and its relationship to forms of reason, I propose, that one can locate a resource for thinking differently about the actions of algorithms. In Michel Foucault's major genealogical work on the history of madness, he argues that madness is designated as a "blatant aberration" against all that would otherwise be "reasonable, ordered and morally wise."[23] Understood in this way, the very idea of human reason has required—integral to its historical condition of possibility within a moral order—an opposition to the aberration of madness. Indeed, for Foucault, "in our culture there can be no reason without madness," and yet the objective science that observes, records, and treats pathologies of madness simultaneously "reduces it and disarms it by lending it the slender status of pathological accident."[24] Thus, there is a necessity of madness within Western

thought, and yet this is a diminished form of madness reduced to a pathological accident or error. The idea of unreason is always already present in claims to moral reason, but this is an unreason understood as a flaw or error in the logic of reason—a flaw that can be identified, classified, and cured.

Such a sense of a reduced form of madness within Western moral order seems to me to be amplified in our contemporary moment, when the actions of algorithms are so frequently named as accidents or errors. The threshold of madness denotes a boundary between the reasonable and the unreasonable, just as it also condenses the unknowability of madness to an error, an accident, or a flaw in the code. Significantly, Foucault's history of madness traces what he calls "a vast movement" from critical to analytical recognition of madness, where classical forms were "open to all that was infinitely reversible" between reason and madness, while the nineteenth and twentieth centuries brought the "classificatory mania" of psychiatry and insisted on new lines demarcating the rational from the mad.[25] The opposition of madness and reason, then, is a historically contingent event that occludes and reduces a more expansive sense of the human experience of madness.[26] As Jacques Derrida describes it, "The issue is to reach the point at which the dialogue" between reason and madness is "broken off" and to permit once more a "free circulation and exchange" between reason and unreason.[27] The modern break in the dialogue between reason and madness, described by Derrida as a "dissension," makes it necessary, he proposes, to "exhume the ground upon which the decisive act linking and separating madness and reason obscurely took root."[28]

The dissension linking and separating madness and reason is a relation rather different in character from the limit point of a madness as aberration from reason. Rather, the dissension itself makes possible new ways for society to understand and to govern the relation between reason and madness. It is an ethicopolitical dissension that generates not only new models of what is normal and abnormal in human societies, but crucially also new ethical relations between selves and others.[29] The relationship between madness, science, and law from the nineteenth century became a moral force in which the rationality of the psychological sciences would assert the boundaries of the normal and the pathological. With the advent of positive psychiatry, as Foucault describes, a "new relation" became possible between the condition of madness and "those who identified it, guarded and judged it" within the "order of an objective gaze."[30] Hence, the dissension forges ethical relations that render madness knowable "at a stroke," authorizing "reasonable men to judge and divide up different kinds of madness according to the new forms of morality," just as it simultaneously tames the nonknowledge of the experience of madness.[31] In

the making of a particular ethical and moral relation to madness, Foucault locates a first major objectification of the human:

> From this point onwards, madness was something other than an object to be feared. . . . It became an object. But one with a quite singular status. In the very movement that objectified it, it became the first objectifying form, and the means by which man could have an objective hold on himself. In earlier times, madness had signified a vertiginous moment of dazzlement, the instant in which, being too bright, a light began to darken. Now that it was a thing exposed to knowledge . . . it operated as a great structure of transparency. This is not to say that knowledge entirely clarified it, but that starting from madness a man could become entirely transparent to scientific investigation.[32]

To make an object of madness, then, was to make a world of transparency in which an individual could have a hold on himself. Let us consider this notion in light of our contemporary moment, when once more madness is becoming object in terms of both the possibilities of rational objective algorithmic decisions and the demand for algorithms to take a hold of themselves and to be rendered transparent to investigation. The moment of dazzlement and darkening Foucault describes, when the place of madness within reason is acknowledged, is entirely absent in the contemporary calls for transparency and the opening of the black box of the algorithm. I wish to reopen the dissension in madness and reason that Derrida and Foucault differently depict and make of this dissension a different kind of ethical terrain. With dissension conjoining madness and reason, the ethical move is no longer a matter of the search for moral codes that regulate the boundary between rational and unreasonable behavior by algorithms. Instead, the manifestation of a moment of apparent madness in an algorithm's actions becomes necessarily also a moment when the algorithm gives an account of its form of reason. Rather than contribute to the cacophony of calls for greater transparency, this ethical demand dwells with the vertiginous moment of dazzlement and darkness, with clouded action and opacity, so that the violences of an algorithm cannot simply be mistakes or errors in the algorithm's otherwise logical rationality.

When one no longer seeks transparency and a hold on oneself, the appearance of a moment of algorithmic madness offers a different kind of insight—an insight into how the algorithm enrolls and deploys ideas of unreason to function and to act. In the moment of madness, we see a dazzling instance of the form of rationality's improbability. In turn, the question of responsibility also shifts ground. Reflecting on the rise of positive psychiatry, Foucault notes that

unreason retained a moral dimension, in which "madness was still haunted by an ethical view of unreason, and the scandal of its animal nature."[33] The responsible subject, then, would act to annex the animal from the human and to locate ethics in the unified mind. In his discussion of homicidal mania and responsibility for murder, for example, Foucault notes that for the defendant to be responsible for the act, there had to be "continuity between him and his gesture."[34]

To be mad in the sense of being outside oneself was to be caught and "alienated," divided within oneself so that a person "was himself and something other than himself."[35] This dividuated subject is precisely embraced by the actions of the algorithm so that its posthuman form is to be simultaneously human and something other than human, a form of self, and fragments of the other. So, responsibility cannot feasibly take the form of taking a hold on algorithms or annexing instinctive or impulsive animal behaviors. I want to take seriously the idea that "in the absence of a fixed point of reference, madness could equally be reason," so that there can be discontinuity between an algorithm and its actions while a responsibility still remains.[36]

To reverse the opposition of madness and reason would be to think radically differently about the many moments when algorithms have been said to err or to deviate from accepted norms. The logic of algorithmic errancy expresses the responsibility of the algorithm precisely in the terms of there being continuity between the algorithm and the gesture. When the algorithm strays or deviates, it is considered irresponsible because it no longer controls the outputs it effects. And yet, as I argue in previous chapters, the experimental straying from paths defines much of the power of contemporary machine learning. A kind of gestural discontinuity, in short, is profoundly useful to the algorithm. To reduce madness to error or aberration is to "neutralize madness" and to shelter "the Cogito and everything related to the intellect and reason from madness."[37] The errors and aberrations of algorithms continue to dominate public discussion of the ethics of machine learning, automated decisions, and data analytics. This dominance of errors and aberrations is indeed neutralizing madness in its broadest sense—as that which cannot be fully known or spoken—and sheltering the rationality of algorithms from a full and expansive critique. The illegible, unspeakable, and opaque actions of algorithms, as I argue throughout this book, are not the limit points of what is possible ethically; instead they are the starting points of an ethical demand.

When the effects of algorithmic decisions are truly horrifying—such as in Cambridge Analytica's rendering of the attributes of "persuadable" voters in the US presidential election and the UK EU referendum (with its use of the at-

tribute to target xenophobic anti-immigration media, for example)—reining in their crazed excesses can never be sufficient. The unreason and the excess are not aberrations at all but are the condition of possibility of action. The problem seems to be, then, not that the rational and the reasonable algorithm takes leave of its senses, loses control, and loosens its hold on its logic. Rather, the algorithm is always already beside itself and divided within itself as such.[38] Algorithms learn by unrestrained experimentation and emergent signals. They simultaneously send multiple conflicting signals along different pathways to optimize their output. They are errant not in the strict sense of deviating from a path, but in the archaic literary sense of traveling in search of adventure. And so, algorithms must also be understood differently in their actions—not as entities whose propensity for madness can be tamed with correct diagnosis and repair, but instead as entities whose particular form of experimental and adventurous rationality incorporates unreason in an intractable and productive knot.

I examine this generative mode of unreason as it animates the decisions of algorithms dealing with human life. The violences and injustices that result from the algorithm's decisions do not emerge primarily from errors, accidents, or aberrations in the system's logic. Such a framing of the violent outputs as errors shelters the cogito from the darkness of unreason. It also neutralizes the fullness of madness as improbability and the unknowable, thus restricting what can count as the harms of algorithms. A principal harm of algorithms is that they enable calculative action where there is incalculability and the unknowable, reducing ethicopolitical orientations to the optimization of outputs and the resolution of difficulties. Renewed attention to the unreason within the algorithm animates how combinations of probabilities generate improbable and untamed outputs that are let loose into the world.

The Madness of Decision

In October 2017, the US National Science Foundation (NSF) awarded a $556,650 research grant to a team of engineers and philosophers working on the decision-making algorithms for autonomous vehicles.[39] The research represents a rather direct example of what I have termed an *encoded ethics*, in which codes of conduct are sought to modify and restrain the harmful effects of algorithmic decisions. The research addresses the classic moral philosophy "trolley problem," redefined for the age of algorithms. This is a scenario in which a trolley car is careering down the tracks toward a group of five people. The driver faces a profoundly difficult decision—to continue on the track toward the certain deaths of the five people or to pull the lever, thereby intentionally

and fatally redirecting the trolley onto a second track where one person lies immobilized. In moral philosophy, the trolley problem is intended to highlight the problem of the grounds of rational decision making. How should the driver weight the value of the lives on the tracks? Is this calculation morally questionable? Should she make an active decision to kill one person to avoid the passive accidental deaths of five? In the NSF-funded research, the trolley problem is reinterpreted for decisions made by algorithms guiding autonomous vehicles. "You could program a car to minimize the number of deaths or life-years lost in any situation," explain the researchers, "but then something counterintuitive happens: when there's a choice between a two-person car and you alone in your self-driving car, the result would be to run you off the road."[40] The research imagines a world in which algorithms can be trained to precompute the rational decision in the face of a scenario of catastrophic consequences.

The boundary of this rationality is a computational threshold in which the algorithms weigh the values of different choices before making a decision. The research team is working on deep neural networks that will "sort through thousands of possible scenarios," filtering out and "rapidly discarding 99.9% of them to arrive at a solution." The example the scientists discuss is a self-driving car hurtling toward a school bus, where the optimal pathway is thought to be "discarding all options that would harm its own passenger" before "sorting through the remaining options to find one that causes least harm to the school bus and its occupants."[41] Here, in the starkest of terms, is a statement of the kind of ethical codes sought by many engineers and algorithm designers.[42] It is an ethical mode that promises to render calculable the profoundly uncertain horizon of an immediate future. Put simply, in seeking to ward off the full horrors of frenzied algorithms governing a trolley out of control, such programs generate new harms in the valuing and weighting of different pathways. The calculus itself—with its combinatorial possibilities of babies in strollers, elderly pedestrians crossing the road, careless delivery drivers, and inattentive cyclists—is always already thoroughly unreasonable. Algorithms do not pose their most serious and harmful threats when they are out of control, become crazed, or depart rationality. On the contrary, among the most significant harms of contemporary decision-making algorithms is that they deny and disavow the madness that haunts all decisions. To be responsible, a decision must be made in recognition that its full effects and consequences cannot be known in advance. A responsible decision-making process could never simply "sort through the remaining options to find one that causes the least harm" because this is an economy of harms that renders all the incalculability of harm calculable.

The madness of the trolley decision is that it must necessarily be made in the darkness of nonknowledge, that it categorically is not subject to preprogramming to optimize an outcome. As Derrida writes, "Saying that a responsible decision must be taken on the basis of knowledge seems to define the condition of possibility of responsibility," and yet at the same time, "if decision-making is relegated to a knowledge that it is content to follow, then it is no more a responsible decision, it is the technical deployment of a cognitive apparatus, the simple mechanistic deployment of a theorem."[43] Though the algorithm's design for the trolley problem may appear to be the simple mechanistic deployment of a theorem, it does contain the aporia of multiple decisions. The problem is that these multiple, fully ethicopolitical decisions are themselves annexed and sheltered from madness by means of errancy or mistake.

One route into the fuller ethicopolitics of an algorithmic decision is to attend to its vast multiplicity. Thus, as I explain in chapter 2, the autonomous vehicle is unable to recognize "child," "stroller," or "school bus" without a regime of recognition trained on millions of parameters of labeled images. The computer scientists working on encoded ethics for the trolley problem (at the heart of the ethicopolitics of all autonomous vehicles) seriously underestimate the difficulty when they speak of filtering "thousands" of possible scenarios, for each of these scenarios is nested within millions of other parameters. An algorithm cannot be programmed to value, for example, the life of a child or an adult as such, for even this decision contains within it the multiple fraught difficulties of learning how to recognize or misrecognize a person. Where the introduction of moral philosophy into engineering problems has sought to contain the algorithm's capacity for an irrational output, the algorithm in fact positively embraces and requires irregularities, chance encounters, even the apparently errant past actions of humans to learn how to optimize the output. The common experimental method of "spoofing" an algorithm in development, for example, involves the exposure of the algorithm to a kind of frenzy of false or "spoof" images to teach it to refine its capacity to recognize truth from falsity. To summarize at this point, when an autonomous vehicle appears to have departed from its logic, lost control, or become mad, it has in fact yielded to the world something of how its logic functions, how its decisions are arrived at. Rather than seeking *an encoded ethics* to try to limit the madness of algorithms, *a cloud ethics* should proceed from the incompleteness and undecidability of all forms of decision. The algorithm contains within its arrangements all the many multiples of past human and machine decisions. It has learned how to learn based on the madness of all decision, on the basis of nonknowledge. To have responsibility for decisions on life, as the autonomous

vehicles' algorithms manifestly do, is to foster conditions in which that learning requires unreasonable things. In the sections that follow, I explore two sets of circumstances when algorithms are deciding on life in the face of an incalculable future.

Life and Unreason I: "Exaggerated Results"

In the summer of 2016 a group of Swedish computer scientists published their findings on algorithms designed to interpret images of the brain from functional magnetic resonance imaging (fMRI) scans. Their investigations into the statistical validity of the major algorithms used globally in neuroimaging seemed to show a stark and troubling rate of error, what they called a "cluster failure."[44] In the context of a method used for neuroimaging over a period of twenty years, the finding that the algorithms have a 70 percent false positive rate cast fundamental doubt on scientific knowledge used in important areas such as Alzheimer's research.[45] A 70 percent false positive rate may appear to be an errant departure from the statistical logics governing algorithmic inference. The neural networks did not depart from their logic, however, but generated what were described as "exaggerated results" in and through that very logic. The deep neural networks were identifying clusters of brain activity in 70 percent of cases where there was no actual cerebral activity present. Crucially, though, the algorithms had been exposed to millions of training images of brain activity to recognize the attributes of new and unseen instances. Here, once again, there is a regime of recognition generated through the algorithm's exposure to the features of past datasets. Among the many assumptions dwelling within the algorithms was a set of norms about the morphology of the brain that ultimately led to misrecognition and a high false positive rate. In short, the proliferation of false positives, with serious consequences for the neurological diagnoses of countless people, was not an aberration in the algorithm's rationality but in fact a manifestation of the excesses of its form of reason.

When multiple false positives are produced in other domains of algorithmic decision, such as in facial recognition systems, often these are understood as errors that could be corrected out.[46] Yet, if one considers that to reduce madness to error is to protect reason, then the resort to error does not limit the algorithm's use in society but actually helps it to proliferate into new domains of life. In my reformulation of the madness of algorithms, I wish to revalue the uncertain relationship to truth that is embodied in algorithms. "Madness begins where man's relationship to the truth becomes cloudy and unclear,"

writes Foucault, so that to experience madness is to experience the "constant error" of the destruction of one's relation to truth.[47] To correct the error would be to seek a kind of transparency and clear-sightedness that would create a break in the cloudiness. Yet, if one stays with the cloudiness of nonknowledge, then a different kind of partial and occluded account is demanded.

What could this different calling to account and responsibility look like? With a cloud ethics, where one is proceeding with the partiality and opacity of all forms of accounting for oneself and others, one would seek to show how and why algorithms can never bear responsibility for undecidability, or for the madness of all decision. For example, when Pedro Domingos makes his bold claims that machine learning algorithms are "learning to cure cancer" by analyzing "the cancer's genome, the patient's genome and medical history" and simulating "the effect of a specific patient's mutations, as well as the effect of different combinations of drugs," a cloud ethics would acknowledge the multiple and distributed selves and others dwelling within the calculus, signaling the madness of the decision itself.[48] The rapid increase in the use of algorithms for genotyping and cancer treatment decisions is marked by the gathering of sufficient data parameters to produce optimized treatment pathways as outputs. But, of course, in the context of health care, there is pressure to include economies within the parameters of these algorithms, such as the cost of the drugs, the benefit to the pharmaceutical industry, the likelihood of "quality life years," or the time efficiency to the overstretched clinician.[49] While the responsible decision of an oncologist is made in a way that is cognizant of the unknowability of the outcomes of rejected treatment pathways, the algorithm's decision proposes the optimal pathway on the calculative basis of the weightings of all possible pathways. Such processes of precomputational decision can never bear the full weight of the madness of all decision. In effect, the precomputed weighting of different branching pathways we saw in the trolley problem is mirrored in the different algorithmic pathways of cancer treatment. In both cases—and there are many more—where the outcome will certainly be the loss of life for some and the preservation of life for others, the madness of the decision as such is disavowed. The madness of the algorithm does not reside in the moral failure of the designer or the statistician, but it is an expression of the forms of unreason folded into a calculative rationality, reducing the multiplicity of potentials to one output. The madness of the algorithm expresses something of what cannot be said; it is the absence of an oeuvre, gesturing to that which is inexplicable and unspeakable.[50]

Life and Unreason II: SKYNET and the Random Forest

In the preceding section, where I discuss a 70 percent false positive rate in neuroimaging algorithms, the matter of life and death hinged, at least in part, on the capacity of the algorithm to recognize the singularity of the patient's brain as the terrain of a target. In other places where the deep neural net algorithm travels, the matter of harm can similarly hinge on the recognition of a target, though in ways in which even a relatively low false positive rate can have catastrophic consequences. One specific kind of algorithm, the random forest, or *random decision forest*, has become the algorithm of choice for systems designed to identify terror targets for the state.[51] Why might this be? Because machine learning for counter-terrorism is a difficult problem. Unlike the commercial algorithms for credit card fraud or email spam detection, for example, where there is vast availability of labeled training data on which to train the classifier algorithms, the availability of labeled data on known terrorists is extremely scarce, indeed even thoroughly inadequate for statistical modeling. This has meant that terrorist-targeting algorithms have become, in a rather direct sense, *unreasonable*. That is to say, in the context of profoundly uncertain security futures, these algorithms have harnessed the unknown and the incalculable to preserve the capacity to act. I am going to dwell a little longer on the random forest algorithm as an arrangement of propositions in the world. The randomness of this algorithm is, I suggest, a kind of sheltered madness that dwells inside the logic of the algorithm and promises to the world an impossible vision of a secure future. It is, in short, an algorithm with life and death effects, and it generates these effects by reformulating the dissension between reason and madness.

In October 2016, two journalists reported that the NSA's SKYNET program "may be killing thousands of innocent people."[52] In the context of the estimated four thousand people killed by drone strikes in Pakistan between 2005 and 2016, Christian Grothoff and Jens Porup investigated how SKYNET "collects metadata," storing it "on NSA cloud servers" and then applying "machine learning to identify leads for a targeted campaign." Among the "cloud analytic building blocks" described within SKYNET in the Snowden files are travel patterns; behavior-based analytics, such as incoming calls to cell phones; and other "enrichments," such as "travel on particular days of the week," "co-travelers," and "common contacts" (figure 4.1).[53]

I have elsewhere written in detail on how these aggregated security "data derivatives" are generated and how they are made actionable.[54] Of specific interest here are the random forest decision tree algorithms that are learning in

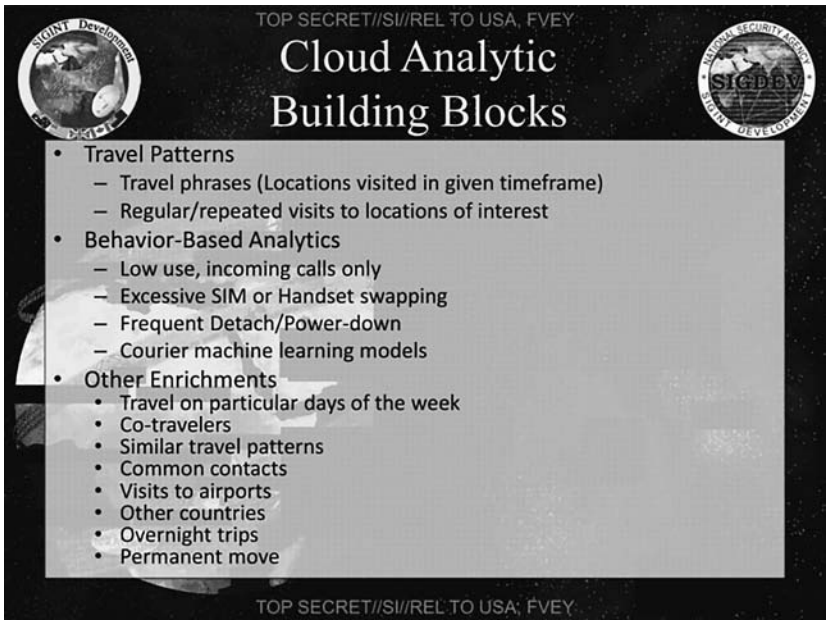

Figure 4.1 SKYNET cloud analytic. Cryptome.org.

communion with the data derivatives. Random forest algorithms were developed at the turn of the millennium by Leo Breiman of UC Berkeley's Department of Statistics.[55] These machine learning algorithms exemplify well how notions of chance, preference, error, and bias become incorporated into the capacity to compute an output. As Breiman explains, in the random forest technique, the algorithms come into being by "growing an ensemble of [decision] trees and letting them vote for the most popular class."[56] Noting that "data sets with many weak inputs are becoming more common" (for example, in medical diagnostics) and that "this type of data is difficult for the usual classifiers—neural nets and trees"—Breiman designed his algorithms to "inject the right amount of randomness" so that the algorithm is trained on random subsets of the training data and "predicts the class of each new record using the plurality of the class predictions from the set of trees."[57] In this way, the random forest algorithm actively uses variance and randomness to refine its output. As Lorraine Daston has argued in her compelling account of the histories of probability, chance, and unreason, all three were able to flourish within methods of calculation.[58] Contemporary algorithms such as random forests have found new ways to invite randomness and chance back into the probabilistic calculus.

If the random forest algorithm appears to descend into the horror of targeting civilians for drone strikes in the SKYNET program (and I have also seen it used by governments to output a risk score of immigration infraction), how can one understand the mode of unreason that dwells within the form of algorithmic rationality? A key element of the random forest algorithm's logic is the attributes-based targeting of population I discuss in previous chapters. The SKYNET random forests are trained on data from a population described as "known couriers." These proxies for labeled data on terrorists are individuals within Pakistan who have been scraped from intelligence reports under the keyword "courier," denoting an individual who is suspected of carrying messages for known terror groups, such as al-Qaeda. The random forest algorithm learns the attributes of "indicative behaviors" from the training data of couriers and then runs the data of 55 million mobile phone users in the general population of Pakistan through this model to "discover similar selectors with courier-like travel patterns."[59] As the random forest algorithm is increasingly now applied to a mobile data stream (and not only a static dataset)—for example, in the video feed of an unmanned aerial vehicle (UAV)—this discovery of similar behavioral patterns is taking place without human input, so that the algorithm is said to "tune its own parameters in response to the data it sees, relieving the analyst of the need to carefully define algorithm parameters."[60]

Though random forest algorithms could be said to generate a kind of madness of false positives that become actionable as a kill list, I am proposing that in fact this madness is useful to the algorithm. Unreason supplies a refinement of the algorithm's logic so that it is no longer limited by the gaze of the analyst. SKYNET's false positive rate is 0.008 percent, where a 50 percent "miss rate" (the false negative rate, or where half of those with "courier-like" patterns are missed) is tolerated in the model. To signal this as error or mistake, however, or to try to fix the problems in the training and validation data is manifestly to overlook the place of unreason in the algorithm's learning. The SKYNET random forests generated one high-profile false positive—recognizing the travel patterns of the Islamabad bureau chief of Al-Jazeera, Ahmad Zaidan, as a high-scoring selector—that was represented as errancy or mistake. To be clear, according to the logic of the random forest, the recognition of Zaidan is not errant at all but is precisely a rational selector given the proxy patterns of the "couriers" and the class predictions of the set of trees. The classifier is perceiving a scene through the apertures of the branches and leaves of the random forest. It learns to distinguish things in the world precisely through the injection of randomness. In the security and defense domains, this capacity of the random forest algorithm to distinguish through variability is highly valued:

"As the classifier watches events unfold, it tries to discern patterns of behaviour: a pack of wolves circling a wounded animal, shoppers taking items from store shelves to a cash register, or an insurgent burying an IED on the side of the road. It's the algorithm's job to learn how to distinguish between someone just digging a hole and someone else burying a bomb."[61]

For the random forest algorithm, the unreasonable and untamed wildness of the random becomes the means of recognizing and distinguishing a target. The random element and the incalculable are lodged within its form of reasoning. To signal the importance of this enfolded unreason is not to deny the profoundly violent effects of deploying the random forest algorithm for terrorism targeting. Far from it. In fact, attention to the unreasonableness of the algorithm as such highlights the impossibility of confining or constraining the madness of the algorithm. If the civilian digging a hole in Pakistan, or the school bus at the border, is mistakenly targeted, the madness of the drone strike is sheltered by the generation of new input data that modifies the threshold between "false alarm" and "miss rate." Every output of a target, however error prone or crazed in its assumptions, supplies new input material to the model.

Given what we know about the place of psychosis and schizoid circuits in the cybernetic history of algorithms, and about the genealogies of the relationship between madness and reason, perhaps it should not be surprising that the twenty-first century is witnessing a reformulation of unreason within the rationality of algorithms. Nineteenth-century diagnosis and treatment of madness and dementia reimagined madness so that it "was not an abstract loss of reason" but instead "a contradiction in the reason that remained," rendering the patient's rationality recoverable.[62] In many ways this notion of a flaw in reason that is repairable has dominated discussion of the public ethics of algorithms. Yet, the modern notion of madness as an error in reason is undergoing reformulation with the advent of decisions involving human and algorithm collaborations. The twenty-first-century violences of algorithmic logics point not to contradictions or flaws in reason that could be cured by the rational human in the loop, but to the potentialities of the excess of reason and the power of unreason. In algorithms such as random forest, the demented wager, randomness, and chance become newly reacquainted with rationality.[63] To treat the pathologies of algorithms, then, is simultaneously to engage in a "forgetting of violence and desires" that continues to animate the algorithms arbitrating the threshold between life and death in our times.[64] A random forest algorithm will never know a terrorist in the sense of acting with clear-sighted knowledge, but it mobilizes proxies, attaches clusters of attributes, and infers behaviors to target and to act regardless.

The Scaffold: Death Penalties and the Condition of Not Knowing

The madness of algorithms, as I have reinterpreted the condition in this chapter, arises not from the loss of rationality or from the error-prone deferral of human reason into the machine, but instead from the making of a calculation in conditions of nonknowledge.[65] When an algorithm generates a singular output from an incalculable multiplicity of associative relations, it shelters the darkness of the decision in the reduction to one actionable output. In Jacques Derrida's writing on the death penalty, he describes "the imposition of *calculability* on a condition of non-knowledge." For Derrida, the specific violence of the death penalty (even as it compares to other modes of killing) is that an unknowable future—"the given moment of my death"—becomes calculable "with absolute precision." "Who thus calculates," writes Derrida, "turns justice into a utilitarian calculation" and makes the death of a person into a "trade, a useful transaction."[66] My argument is that contemporary algorithms are extending this useful transaction of the programmable decision—the useful output, the good enough model—into other calculations at the threshold of life.

The weights, thresholds, and attributes through which an algorithm comes into being are simultaneously the condition for assigning the calculative weight or the value of someone or something.[67] Just as Derrida identified in the US penal architecture that "the majority of those condemned to death are blacks and poor blacks," our contemporary times witness a calculating machine that also generates weighted and racialized targets and sentences. Despite the manifest differences between the discrete logics of particular machine learning algorithms, what they all share in common is the reduction of a multiplicity of incalculable differences to a single output. This output is a finite target contingent on infinite multiples of weights and weightings within the calculation. Thus, when a random forest algorithm sentences someone to death by drone strike, the infinite (gestures, connections, potentials) makes itself finite (optimal output, selector, score), and the horizon of potentials is reduced to one condensed output signal.

Moreover, this reduction to one output signal, a kind of death penalty in its making of precise action from nonknowledge, shelters itself from its own unreason. As societies seek out an encoded ethics that annexes and extracts the errant, the aberrant, from algorithms, they are failing to understand how these apparent pathologies are actually of the essence of algorithmic learning. The clustering of false positives in a particular black population in one part of a city, for example, if understood as errant becomes subject to correction

and recalibration. Yet, this experimentation and wild errancy continues to generate targets long after a correction is made. The random forest algorithms in SKYNET, as well as those in immigration systems that return targeted migrants, condemn to death unknown people living and traveling in already risky spaces, and they do so with wagers, votes, chance, and randomness. The algorithm underwrites its own rationality because it is engaged in defining new thresholds of normal and abnormal behaviors, reasonable and unreasonable travel, the arbitration of the good and the bad.

The architecture of algorithms adjudicating life and death takes the form of a kind of scaffold, where the scaffold is the spatial arrangement of the death penalty, the place of execution.[68] In this sense the algorithm as scaffold is part of the scaffolding of sovereignty, where the state "will have constructed all the scaffolds and propped up all the figures of machines for killing legally, sovereignly, nationally in the history of humanity."[69] The algorithm as scaffold is sometimes—as with random forest algorithms for national security—the means by which states exercise the right to precisely calculate death in advance; but it is also the architecture of a decision, a "certain modality, a certain qualification of living and dying, a manner, an apparatus, a theatre, a scene of giving life and giving death." Moreover, the scaffold "guarantees some anonymity" for the executioner, who is not present on the scaffold and whose hand is not on the guillotine, an "executioner who does not kill, not in his own name."[70] The random forest algorithms I discuss in this chapter, as scaffolded arrangements, similarly apportion weights to life and death as output signals, guaranteeing some anonymity to the executioner, who does not appear on the platform. The algorithm is precisely a space in which the excesses, the unreason, the cruelty of the adjudicator can be given free rein. And so, to find a critical response to the algorithm via a demand for limits on its unreasonable excess is to seriously overlook how exactly it deploys unreason to generate a finite output.