CARDOZO LAW

Benjamin N. Cardozo School of Law · Yeshiva University
Jacob Burns Institute for Advanced Legal Studies
September, 2014
Faculty Research Paper No. 441

Human-Focused Turing Tests: A Framework for Judging Nudging and Techno-Social Engineering of Human Beings

Brett M. Frischmann
Professor of Law & Director, Cardozo Intellectual Property and Information Law Program
Benjamin N. Cardozo School of Law
55 Fifth Avenue, Room 1004
New York, NY 10003
(212) 790-0859 (phone)
frischma@yu.edu

Human-focused Turing tests: A framework for judging nudging and techno-social engineering of human beings

Abstract

This article makes two major contributions. First, it develops a methodology to investigate techno-social engineering of human beings. Many claim that technology dehumanizes, but this article is the first to develop a systematic approach to identifying when technologies dehumanize. The methodology depends on a fundamental and radical repurposing of the Turing test. The article develops an initial series of human-focused tests to examine different aspects of intelligence and distinguish humans from machines: (a) *mathematical computation*, (b) *random number generation*, (c) *common sense*, and (d) *rationality*. All four are plausible reverse Turing tests that generally could be used to distinguish humans and machines. Yet the first two do not implicate fundamental notions of what it means to be a human; the third and fourth do. When these latter two tests are passed, we have good reason to question and evaluate the humans and the techno-social environment within which they are situated.

Second, this article applies insights from the common sense and rationality tests to evaluate the ongoing behavioral law and economics project of nudging us to become rational humans. Based on decades of findings from cognitive psychologists and behavioral economists, this project has influenced academics across many disciplines and public policies around the world. There are a variety of institutional means for implementing "nudges" to improve human decision making in contexts where humans tend to act irrationally or contrary to their own welfare. Cass Sunstein defines nudges more narrowly and carefully as "low-cost, choice-preserving, behaviorally informed approaches to regulatory problems, including disclosure requirements, default rules, and simplification." These approaches tend to be transparent and more palatable. But there are other approaches, such as covert nudges like subliminal advertising. The underlying logic of nudging is to construct or modify the "choice architecture" or the environment within which humans make decisions. Yet as Lawrence Lessig made clear long ago, architecture regulates powerfully but subtly, and it can easily run roughshod over values that don't matter to the architects. Techno-social engineering through (choice) architecture is rampant and will grow in scale and scope in the near future, and it demands close attention because of its subtle influence on both what people do and what people believe to be possible. Accordingly, this article evaluates nudging as a systematic agenda where institutional decisions about particular nudges aggregate and set a path that entails techno-social engineering of humans and society.

The article concludes with two true stories that bring these two contributions together. Neither is quite a story of dehumanization where humans become indistinguishable from machines. Rather, each is an example of an incremental step in that direction. The first concerns techno-social engineering of children's preferences. It is the story of a simple nudge, implemented through the use of a wearable technology distributed in an elementary school for the purpose of encouraging fitness. The second concerns techno-social engineering of human emotions—the Facebook Emotional Contagion Experiment. It is not (yet) a conventional nudge, but it relies on the underlying logic of nudging. Both can be seen as steps along the same path.

Introduction

Many science fiction stories pit humans against machines. Sometimes, the machines remain tools of powerful humans that oppress everyone else; sometimes, the machines become the oppressors. Often, humans unwittingly sow the seeds of their own destruction or subservience by madly rushing down a technological path attracted by the siren's call of efficiency, optimization, and perfection--only to learn too late, that along the way, they've lost their humanity.

What if we were rushing down such a path? Would we know? How would we be able to distinguish reality from science fiction? After all, if someone had written a book in 1960 that perfectly described the world we currently live in, it would undoubtedly have been science fiction when written; it would have been regarded as incredible fantasy. Some readers would have understood it as utopian; others as dystopian--depending, of course, on the readers' own values and conception about what the good life and a good society would be.

What follows is not science fiction, but it takes seriously the idea that many technologies currently deployed and being developed (i.e., on the foreseeable horizon) deserve more careful attention because of their potential to (re)construct humans and society on an unprecedented scale and scope.

Humans have been shaped by technology since the dawn of time, and of course, humans have shaped other humans through technology for a very long time as well. Many people have written on this topic. Even the topic of shaping humans to be machines is not new; it has garnered significant attention in the context of the workplace and mass media such as radio and television. Still, the scale and scope of "human construction" through technological means has not been uniform over time, and I suspect that we've experienced acceleration over the past few decades with the near-ubiquitous deployment of various information and communications technologies. Looking at the present and to the near future, one thing seems clear: interconnected sensor networks, the Internet of Things, and (big) data enabled automation of systems around, about, on and in human beings promise to expand the scale and scope significantly. It is the fine-grained, hyper-personalized, ubiquitous, continuous and environmental aspects of the techno-social engineering that make the scale and scope unprecedented.

Four central themes of the broader project of which this article is a part are:
- when does technology (automated systems) replace or diminish our humanity?
- can we detect when this happens? how will we?
- what makes us human?
- when and how do humans become programmable?

Like science fiction, public discourse on these themes often is split among utopian and dystopian perspectives. The utopian perspective celebrates technological innovation and emphasizes the (best) upside while the dystopian perspective laments technological "innovation" (scare quotes intended) and emphasizes the (worst) downside. Both perspectives tend to be dismissive of the other. Of course, there are plenty of intermediate positions, but the extremes dominate the discourse. Technology and humanity are abstract and complex; the themes noted above are not

easy to discuss and evaluate. There are incredible definitional obstacles and evidence is nonexistent because we don't know what exactly to measure or how to evaluate what we see. We simply do not have the tools for identifying and evaluating when technology is dehumanizing, yet we desperately need them in our modern environment.

This article develops a methodology, a series of tools to operationalize the investigation. At its core, the methodology depends on a fundamental and radical repurposing of the Turing test.[1] The methodology relies on the wonderful move Alan Turing made by shifting to an observation "game" that focuses on whether humans and machines are indistinguishable with respect to a specified functional capacity. Having asked the question – *can machines think?*, Turing acknowledged the difficulties inherent in defining "machine" and "think," and he proposed that the seemingly intractable question be replaced with a test that could serve as an operational definition of intelligence. The conventional subject of the Turing test is machines. Thus, he might have asked: *Can we construct machines that think? How would we know? Are any indistinguishable from a human?* In this context, humans serve as a baseline against which to evaluate machines. When the observational test is passed, meaning a particular machine is determined to be indistinguishable from a human, the test generates evidence that requires interpretation and evaluation. It is not a magical moment in which the machine transforms into a human being like a scene from Cinderella. The machine remains a machine, but the evidence might tell us something meaningful (about the machine, the observer, the questions asked, or the context).

The subject of the radically repurposed Turing tests developed in this article is humans. I am interested in identifying and evaluating when technology is dehumanizing. Thus I ask: *Can we construct humans that are in some meaningful way less human? How would we know? Are any indistinguishable from a machine?* In this context, machines serve as a baseline against which to evaluate humans. As with the conventional Turing test, the human-focused tests are a means for generating evidence. Thus, I make a similar move as Turing, except I am not only interested in intelligence or thinking. There is more to humanity than that. This article is part of a larger book project. The book project develops additional tests that move beyond intelligence and explore free will, autonomy, and sociality, among other things.[2] Thus, I will test different functional capacities, and to some degree, avoid seemingly intractable definitional and essentialist debates about what it means to be human. I say to some degree because I cannot avoid choosing functional capacities and how to test them. The methodology leaves room for others to refine the tests I develop and to develop other tests for whatever functional capacities they deem essential to being human and for which machines serve as a useful baseline.[3]

[1] Turing, Alan. 1950. *Computing Machinery and Intelligence*. Mind LIX: 433-460. The closest attempt to repurpose the Turing test to investigate our humanity is Brian Christian's The Most Human Human (2011), which describes his participation in the Loebner competition and explores some of the themes raised in this project.

[2] The book engages a range of normative issues, including basic values questions, happiness vs. capabilities type concerns, the "thingification of people," and subtle distributive justice questions concerning the ways in which constructive environments convey considerable power. It also applies the human-focused tests in the context of the Internet of Things and related technologically (re)constructed and human constructive environments. The applications explore the perils of persistent tethering in "always-on" environments, the prospect of programmable people, and the fundamental reshaping of the environment we inhabit physically, mentally, and socially.

[3] The move is analogous to one made by Amartya Sen in his development of the Capabilities Approach, which also focuses on various functional capacities and leaves room for different priorities and lists of capabilities. See, e.g., Amartya Sen, Human rights and capabilities, 6 (2) J. Human Development 151–166 (2005); Amartya Sen,

This article is organized into two Parts. The first Part focuses on reorienting the Turing test and developing a series of human-focused tests that better allow us to identify and consequently evaluate contexts within which humans are or become indistinguishable from machines. After introducing the basic contours of the conventional Turing test, this Part inverts the conventional perspective and focuses on humans rather than machines. To answer the question *Can Humans Not-Think?*, it develops an initial series of human-focused tests to examine different aspects of intelligence and distinguish humans from machines: (a) *mathematical computation*, (b) *random number generation*, (c) *common sense*, and (d) *rationality*. I briefly discuss the first two and devote more attention to the third and fourth. All four are plausible reverse Turing tests that generally could be used to distinguish humans and machines. Yet the first two do not implicate fundamental notions of what it means to be a human; the third and fourth do. For each test, we begin with a brief description of what it is that we are testing, for example, by specifying the stimuli used by the observer, and then discuss how to interpret the results, for example, by exploring whether passing or failing the test would support meaningful inferences about the human agents.

Part II applies insights from the common sense and rationality tests to investigate the ongoing behavioral law and economics project of constructing rational humans. Based on decades of findings from cognitive psychologists and behavioral economists, this project has gained incredible momentum influencing academics across many disciplines and public policies and regulations around the world. There are a variety of institutional means for implementing so-called "nudges" to improve human decision making in various contexts where humans tend to act irrationally or contrary to their own welfare.[4] Cass Sunstein defines nudges more narrowly and carefully (from a political perspective) as "low-cost, choice-preserving, behaviorally informed approaches to regulatory problems, including disclosure requirements, default rules, and simplification."[5] These approaches tend to be transparent and more palatable. But there are other approaches, such as covert nudges like subliminal advertising.[6] The underlying logic of nudging is to construct or modify the "choice architecture" or the environment within which humans make decisions. As Lawrence Lessig made clear long ago, architecture regulates powerfully but subtly, and it can easily run roughshod over values that don't matter to the architects.[7] Techno-social engineering through (choice) architecture already is rampant and will

---

Commodities and capabilities (1985); Amartya Sen, Development as freedom (2001). See also Sabina Alkire, Valuing freedoms: Sen's capability approach and poverty reduction (2002).

[4] See, e.g., Christine Jolls, *Behavioral Law and Economics* (update of the essay published under the same title in Behavioral Economics and Its Applications, Peter Diamond and Hannu Vartiainen eds., Princeton University Press, 2007) [http://www.law.yale.edu/documents/pdf/Faculty/Jolls_BehavioralLawandEconomics.pdf]; Thaler, R. H., & Sunstein, C. R. (2008), *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press; Ariely, Dan (2008), *Predictably Irrational: The Hidden Forces that Shape Our Decisions* HarperCollins, 1st edition; Amir & Lobel, *Stumble, Predict, Nudge*, 108 Columbia Law Review 2098 (2008).

[5] Cass R. Sunstein, *Nudges.gov: Behavioral Economics and Regulation*, Forthcoming, Oxford Handbook of Behavioral Economics and the Law (Eyal Zamir and Doron Teichman eds.) [Very preliminary draft 2/16/13, available at http://ssrn.com/abstract=2220022. (citing Thaler & Sunstein 2008)).

[6] Gidon Felsen et al., *Decisional Enhancement and Autonomy: Public Attitudes towards Overt and Covert Nudges*, 8 Judgment and Decision Making 202 (2013).

[7] Lawrence Lessig, Code and Other Laws of Cyberspace (1999).

only grow in scale and scope in the near future, and it demands close attention because of its subtle influence on both what people do and what people believe to be possible. Before long, lost values[8] are truly lost.[9] Accordingly, drawing on insights from Part I, Part II begins to evaluate nudging as a systematic agenda where institutional decisions about particular nudges aggregate and set a path for society that entails techno-social engineering of humans and society.

The nudging project is immense, interdisciplinary, and growing rapidly. There is a tremendous literature. To cabin the analysis, I focus on, *Choosing Not to Choose*,[10] a recent article by Cass Sunstein in which he discusses a particular controversy regarding paternalism within the behavioral law and economics literature about nudging. Specifically, Sunstein examines when private or public institutions might require people to actively choose rather than allow people to choose not to choose (or allow others to make choices on their behalf). He emphasizes how restricting a person's opportunity to choose not to choose is itself a form of paternalism and thus requiring active choosing is subject to the same criticisms as other nudges. This is an importance defense of nudging. Thaler and Sunstein consistently maintain that nudging should be underwritten by an ethic of "libertarian paternalism," one component of which is that agents should find it easy to opt-out of a nudge's trajectory; people always should be able to choose for themselves and even behave irrationally, if that is what they really want to do. In their view, this ethic preserves individual autonomy.

Yet, as Part II explains, there is more to being human than autonomy. Suppose, by autonomy, one means freedom to determine one's second order beliefs and desires: even if we preserve that type of autonomy completely, we can still end up as slaves in a severely constrained environment without the freedom to act on our desires. Suppose, instead, that by autonomy, one means the practical and situated freedom to act on one's desires or will: even if we preserve that type of autonomy completely, we can still end up as automatons with others determining our beliefs and preferences. *Over-determined or over-architected environments that preserve either type of autonomy still can be dehumanizing*. In *Choosing Not to Choose*, Sunstein demonstrates his awareness of these possibilities, referring to Aldous Huxley's Brave New World and recognizing the potential value of environments architected to preserve serendipity.[11] Awareness of the possibilities is not enough, however. We need much more. The approach developed in Part I is only a start. We must be able to identify and evaluate the human capabilities at stake.

Part II concludes with two true stories that bring Parts I and II together. Neither is quite a story of dehumanization where humans are or become indistinguishable from machines. Rather, each

[8] Lost to the architecture.
[9] Lost to the people influenced by the architecture.
[10] This provides a current and focused view and avoids engaging the entire project, which would overcomplicate things. My objective is primarily to reveal a different macro-level perspective on the project and suggest how the reverse Turing test methodology is useful in judging some nudging.
[11] Sunstein, draft at 30 & 32 (citing ALDOUS HUXLEY, BRAVE NEW WORLD xvii (1932) and JANE JACOBS, THE DEATH AND LIFE OF GREAT AMERICAN CITIES (1961)).

is an example of an incremental step in that direction. The first concerns techno-social engineering of children's preferences. It is the story of a simple nudge, implemented through the use of a wearable technology distributed in an elementary school for the purpose of encouraging fitness (see photograph below). The second concerns techno-social engineering of human emotions—the Facebook Emotional Contagion Experiment. It is not (yet) a conventional nudge, but it relies on the underlying logic of nudging. Both can be seen as steps along the same path.

A Simple Nudge: Using Activity Watches to Shape Elementary School Children

I. Human-focused Reverse Turing Tests

In this Part, I propose a reorientation and reframing of the Turing test. Specifically, I propose we make human beings the relevant subject and test whether human beings are (in)distinguishable from machines. The context within which the test applies is quite important, as a significant feature of this human-focused test is the relevance of the environment or context within which the test is applied. The most important questions to consider are whether, when, and how human beings can be constructed via technology, social context, and the environment within which we live and through which our preferences and beliefs are formed to be indistinguishable from machines.

The conventional Turing test concerning artificial intelligence focuses on a machine and asks whether the subject is (in)distinguishable from a human being. In a sense, the Turing test establishes an elusive endpoint to which AI experts and others may strive; it is a finish line. But racing to make intelligent machines is only half of the relevant picture. Another race is occurring, but we don't pay much attention to it, except in science fiction. It occurs on the other side of what I call the Turing line, the human side.

A. Turing Test: A brief overview and literature review

This section situates the reader and this project in the expansive literatures that have developed since Turing published his paper. Before proposing a reorientation and reframing of the Turing test, I would like to make sure we are all on the same page and understand the Turing test in more or less the same way. Accordingly, this section provides a brief summary of the Turing test and the main conclusions I have drawn from a literature review; it notes some of the most relevant challenges and extensions, such as Searle's Chinese language thought experiment, Harnad's *Total Turing Test* ("TTT") and Schweizer's *Truly Total Turing Test* ("TTTT").

In his seminal paper, *Computing Machinery and Intelligence*, Turing proposed to consider the question, "can machines think?" He acknowledged the difficulties inherent in defining "machine" and "think," and he proposed that the seemingly intractable question should be replaced with a test that could serve as an operational definition of intelligence. [12] He developed a test based on the Imitation Game. Thus, Turing moved away from definitions toward a method of testing and developing empirical evidence. At the outset of his paper, Turing justified this move on the grounds the initial question is "too meaningless to deserve discussion." However, the question is not meaningless in the sense that it should not be pursued; instead, it is meaningless because it is predicated on intractable definitional moves. "What he was proposing with his test is a way to make the overall question of machine thinking more precise so that at least in principle an empirical test could be conducted. Thus, Turing's replacement strategy involves both a clarification of meaning, particularly about the nature of the machine, and a

[12] Many commentators contend that Turing did not intend the TT to be an operational definition of intelligence (providing necessary and sufficient conditions). However, this does not mean that "passing" is not a sufficient condition for intelligence. Dennett, Daniel, *The milk of human intentionality*, Behavioral and Brain Sciences 3: 429-430 (1980). Some commentators think that test is merely a procedure for getting good evidence about whether a machine is intelligent. Moor, James H., *The Status and Future of the Turing Test*, Minds and Machines 11: 77-93 (2001); Schweizer, Paul, *The Truly Total Turing Test*, Minds and Machines 8: 263-272. (1998).

procedure for obtaining good evidence."[13]

The TT is modeled on a popular party game, the "imitation game", where a man (A), and a woman (B), enter into a separate room from the interrogator (C), and the interrogator attempts to determine which of the other two is the man and which woman by asking a series of questions. The interrogator knows the individuals by the labels X and Y. And, For example, C may ask, "will X tell me the length of his or her hair?" One caveat of the game is that A's objective is to cause C to make the wrong identification. As a result, A may answer questions untruthfully in order to increase the odds of a wrong identification. Also, the answers are provided as text (typed) in order to prevent any bias based on tone of voice. After several rounds of questioning, the interrogator guesses the sex of the individual by saying: "X is A" or "Y is A."

Turing's test for machine intelligence takes much of its structure from the Imitation Game. In the standard version of the TT, a machine and a human are separated from an interrogator, and the interrogator poses a series of questions to the machine and the human in order to identify which agent is human. Further, the machine will attempt to exhibit human-like conversational behavior in order to trick the interrogator into making the wrong identification. Like the imitation game, answers from the machine and human are typed to avoid any biases [possibly information that is irrelevant to attribution of thinking]. Turing predicted that by the year 2000, an interrogator would not have a greater than seventy percent chance of correctly identifying the machine after five minutes of questioning.[14] This identification threshold is a common view of what constitutes "passing" the TT.

The TT scrutinizes the verbal behavior of the two agents involved in the imitation game. Verbal behavior is considered an appropriate locus of investigation because of the natural link between intelligence and verbal output. Consider, for example, the attribution of intelligence to other humans. The mental states of other agents are private; as a result, we do not have direct access to others' mental processes yet we attribute intelligence to other people. Further, we consider other humans intelligent based on their verbal outputs. You have a conversation with a patron in line at the local coffee shop and based on the complex, intelligent verbal outputs you, justifiably, attribute intelligence to him/her. If we believe that the mind of the patron is a machine, then we are equally justified in attributing intelligence to another human and a computer when the verbal outputs of the entity seem intelligent. Consequently, one might conclude that any reluctance to attribute intelligence to a verbally competent computer would be the result of human bias and therefore unjustified. There are, of course, complications to this line of argument; for example, as commentators have suggested, the attribution of intelligence among human beings if often, if not always, based on additional social, cultural or contextual factors. Specifically, Harnad and Schweizer (both discussed below) push back against the idea that the TT is the same mechanism/evidence that we use to attribute intelligence to other people. Harnad says that we also look at behavior or sensorimotor functioning. Schweizer says that our attribution of intelligence takes place against a backdrop of what we know about humans.

Stuart Shieber produced a helpful distillation of the TT into its basic formal argument structure.[15]

[13] Moor, James H. (2001). *The Status and Future of the Turing Test*. Minds and Machines 11: 77-93, at 82.
[14] Turing, Alan. 1950. Computing Machinery and Intelligence. Mind LIX: 433-460.
[15] Shieber, Stuart (2008). The Turing Test as Interactive Proof, Nous, 41(4): 686-713.

> Premise 1: If an agent passes a Turing Test, then it produces a sensible sequence of verbal responses to a sequence of verbal stimuli.
>
> Premise 2: If an agent produces a sensible sequence of verbal responses to a sequence of verbal stimuli, then it is intelligent.
>
> Conclusion: Therefore, if an agent passes a Turing Test, then it is intelligent.

Many critiques of the TT concentrate on the second premise. The arguments contend that verbal behavior cannot provide sufficient information for the attribution of intelligence to an entity. In other words, intelligence or justifiable attribution of intelligence is not reducible to verbal behavior. As noted in the Introduction, we also have concerns with focusing exclusively on intelligence, much less verbal behavior, as the characteristic or attribute that distinguishes humans and machines, especially when we approach the Turing line from the human side.

Searle famously challenged the second premise with his *Chinese Language Room* thought experiment.[16] The Internet Encyclopedia of Philosophy concisely describes the thought experiment:

> Searle (1980) asks you to imagine yourself a monolingual English speaker "locked in a room, and given a large batch of Chinese writing" plus "a second batch of Chinese script" and "a set of rules" in English "for correlating the second batch with the first batch." The rules "correlate one set of formal symbols with another set of formal symbols"; "formal" (or "syntactic") meaning you "can identify the symbols entirely by their shapes." A third batch of Chinese symbols and more instructions in English enable you "to correlate elements of this third batch with elements of the first two batches" and instruct you, thereby, "to give back certain sorts of Chinese symbols with certain sorts of shapes in response." Those giving you the symbols "call the first batch 'a script' [a data structure with natural language processing applications], "they call the second batch 'a story', and they call the third batch 'questions'; the symbols you give back "they call . . . 'answers to the questions'"; "the set of rules in English . . . they call 'the program'": you yourself know none of this. Nevertheless, you "get so good at following the instructions" that "from the point of view of someone outside the room" your responses are "absolutely indistinguishable from those of Chinese speakers." … [17]

Searle reasons that like the English speaker who understood none of the story, questions or answers, a computer *understands* nothing, regardless of whether it appears indistinguishable from human speakers. In a sense, Searle argues that computers can simulate thinking but cannot think.

A significant firestorm followed and the debate has not ended. Dennett made the intriguing argument that everything Searle says about the computer also can be said about the person in the

[16] Searle, John. 1980. *Minds, Brains, and Programs*. Behavioral and Brain Sciences 3, 417-424.
[17] http://www.iep.utm.edu/chineser/

room (or more accurately, the brain of the person in the room).[18] The person speaks English and outside observers take such speech to imply that the person understands what he or she is saying, but who knows what is truly going on inside the person's head and whether the brain is in fact any different from the intelligent-seeming computer. Put another way, all together, the English-speaking person and the instructions—as a system—do understand Chinese.[19] We could go around in circles, perhaps spiraling toward something, but as we will see below, we don't need to do so because we are not interested in determining the necessary and sufficient conditions for attributing intelligence to machines that is on par with humans.

In the midst of the firestorm, Harnad develops what he refers to as the *Total Turing Test* (TTT), which expands the locus of examination to include nonverbal behavior.[20] He claims that "our ability to interact bodily with the things in the world and the many nonverbal ways we do – are as important to the test as our linguistic capacities. …Moreover, there are strong reasons to doubt that a device could pass the teletype version of the Turing Test if it were not also capable of passing the robot version." Harnad also says "it is hard to imagine, for example, that a TT candidate could chat with you coherently about the object in the world until doomsday without ever having encountered any objects directly." This idea of symbol grounding is a response to Searle's "Chinese Room."[21] Incorporating Searle's ideas, Harnad contends that there must be some connection between the symbols and what they denote to have actual thought. Without experience of symbol referents the processes would "send us round and round in endless circles, from one meaningless string of symbols to another."[22] Harnad contends that incorporating robotic, sensorimotor performance escapes this problem, in part because our "linguistic capacity must be … grounded in our robotic capacity."[23]

Schweizer goes further than Harnad and develops what he refers to as the *Truly Total Turing Test* (TTTT).[24] He emphasizes that neither the original TT nor the TTT are tests that, if passed, would wholly justify ascription of intelligence; they are not sufficient conditions. He explains that we do not really use a short conversation with a person to attribute intelligence to that person. We do use and rely on the conversation, but in large part because the conversation is taking place against the backdrop of "extended and multifarious interactions with human beings generally."[25] If we attribute intelligence to machines on the basis of a five minute conversation, these machines would be making use of background knowledge that is relevant to a different type of entity, humans. According to the TTTT, to evaluate whether "artificial or alien cognitive structure should be regarded as possessing intelligence on a par with human beings," the type being tested would have to be capable of generating and relying on its own background knowledge. Thus, for Schweizer, "what is necessary for genuinely comparable performance is

[18] Dennett, Daniel. 1980. The milk of human intentionality. Behavioral and Brain Sciences 3: 429-430. For an illustration, see http://www.visuallanguagelab.com/chinese_room/index.html.
[19] DANIEL C. DENNETT, CONSCIOUSNESS EXPLAINED 439 (1991).
[20] Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem, Minds and Machines 1: 43-54.
[21] Id.
[22] Id.
[23] Id. See also Harnad, S. (1990). *The Symbol Grounding Problem*. Physica D 42: 335-346; Erion, Gerald (2001). *The Cartesian Test for Automatism*, Minds and Machines 11: 29-39.
[24] Schweizer, Paul (1998). The Truly Total Turing Test, Minds and Machines 8: 263-272.
[25] Id. at 266.

that the algorithms responsible for [a] robot's behavior not only enable it to use the languages that we have programmed into it, but rather that a community of such robots could develop these languages starting from the 'state of nature' of our prelinguistic forebears."[26]

* * *

The major critiques and extensions of the TT reveal contested conceptions of what the TT aims to accomplish and what, if anything, it actually accomplishes. My objective is not to defend the TT or reconcile these various perspectives. Rather, my objective is to show how the TT serves as a useful analytical tool or methodology. Some have questioned whether it is useful in this regard, suggesting that it fails to meet the standard that all graduate students learn for experimental design: "never [] design an experiment to detect nothing."[27] Hayes and Ford suggest the Turing Test suffers from this design flaw and, as a result, even when the test is passed, you're forced back to the drawing board to figure out what it all means: *Does passing the test tell us the machine is intelligent? That the observer asked the wrong questions? And by the way, what are the right questions? And heck, isn't that what the imitation game was supposed to help us avoid?* And so on. In my view, such recursivity is the Turing test's virtue, rather than its vice.

The Turing test should be understood in context and with a sense of the initial question that motivated Turing (*can machines think?*). The test is a means for developing objective, empirical evidence about *something*. The difficult question is *what*? Does it tell us something meaningful -- or more accurately, does it provide us with information that allows us to infer something meaningful -- about machine intelligence? Or perhaps something else--perhaps even the questions the observer asked? We can have that discussion when the time comes, and it probably would be interesting and worthwhile. TT extensions, such as the TTT and TTTT, seem to be looking for more, such as necessary and sufficient conditions for attributing intelligence to machines that is on par with humans. But that need not be the objective, and just to be clear upfront, it will not be our objective.

The Turing test, TTT, TTTT, and TT … T[28] provide us with a systematic approach to thinking about the Turing line and investigating the similarities, differences, and relationships between humans, machines, and machine-environments (explained in the next section). Some reviewers have suggested that I abandon the Turing test altogether because it no longer is important as a goal or objective in AI and related fields. But I am not using it for that purpose. I find it a useful conceptual lens. It allows us to maintain focus on the Turing line while exploring the relationships among human, machine, and environment. Thus, I am as interested in thinking about the experimental designs, game structures, and questions to ask as I am about anticipating how I might interpret the results. The recursive nature of applying the Turing test, which Hayes and Ford critiqued, actually may be its redeeming feature. As Hayes and Ford acknowledge at the end of their article:

> We suspect that Turing ... wanted the test to be about what it really means to be human. This

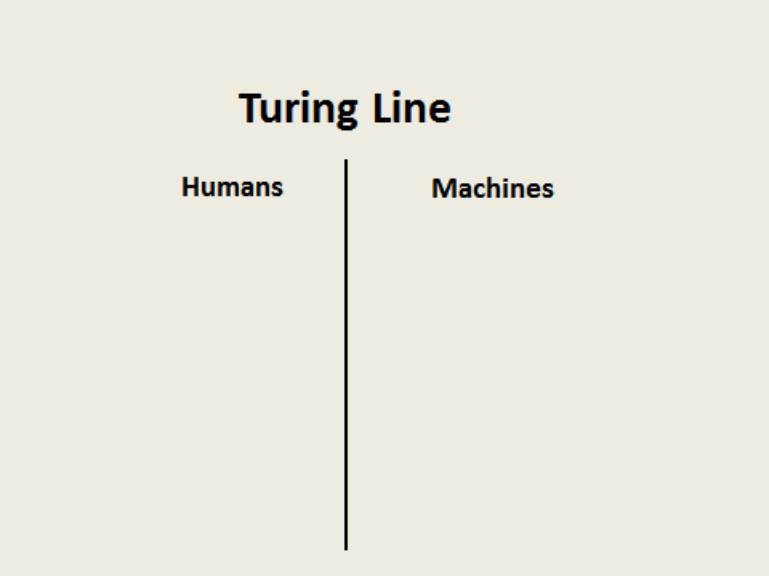
[26] Id. at 268.
[27] Hayes and Ford (1995), Turing Test Considered Harmful, 972-77.
[28] You can imagine a long string of TT test extensions. Perhaps the Truly Terrific Total Turing Test?

> is why he has set us up in this way. … If we really tried to do [what Turing suggests], we might be forced into thinking very hard about what it really means to be not just a thinker, but a human being in a human society, with all its difficulties and complexities. If this was what Turing meant, then we need not reject it as our ultimate goal.[29]

I agree but wish to move beyond investigation of machines. Accordingly, I develop a similar series of tests to investigate the human-side of the Turing line.

B. The human side of the Turing line

If the Turing test draws a line, it would be a line between human and machine. The line might be bright, it might be fuzzy, it might be fixed, it might change, it might be real, and it might be illusory. Perhaps humans really are just meat machines. I assume a line exists. Thus, I assume the following:



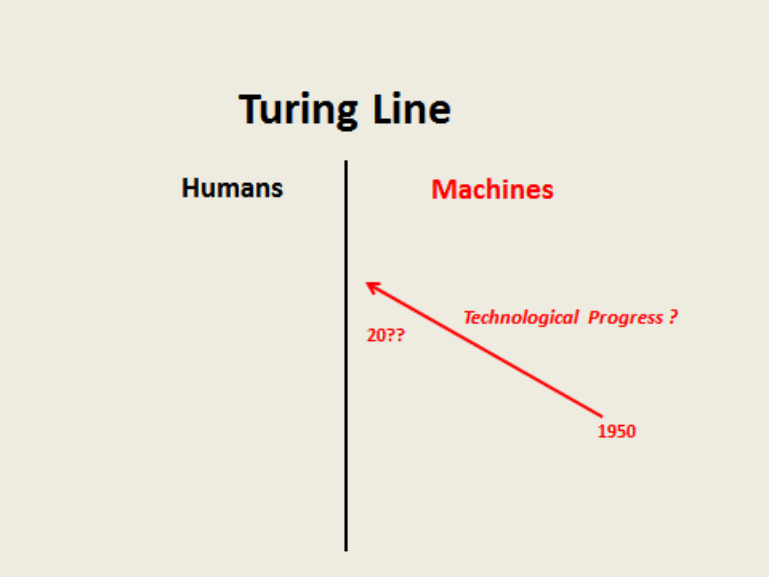
The Turing line serves at least two functions, which are noted but not fully examined within the relevant literatures.

First, the line differentiates humans and machines, and in this sense, it is a line we assume exists and that we recognize and understand (even if we don't fully understand, and the point of the mental exercise is to get us to examine and better understand it). The TT is satisfied when the line is in fact not seen; that is, during the course of playing the Imitation Game, a machine successfully imitates a human and thus a line that we "know" exists is not observed. The machine remains a machine even after passing the test, but we infer that a machine capable of passing the test has "something," some characteristic (some would say, intelligence, some would not), that makes, or at least made, it indistinguishable from a human in the context of the test.

Second, it is a finish line. Within artificial intelligence, robotics, machine learning, and other

[29] Id. at 977.

adjacent fields, the race has been on since Turing published his article.[30] Of course, the race had probably already begun. My point is that Turing demarcated a finish line to be crossed, something specific to aim for when constructing machines and programming systems. Turing made his prediction about the rate of progress, suggesting that by the year 2000, an observer would not have a greater than seventy percent chance of correctly identifying the machine after five minutes of questioning.[31] We didn't quite make it. Some have suggested we'll cross it soon, and others have suggested we'll never cross it. But the important point, for my purposes at least, is to recognize the function. The picture below is meant to illustrate one way to conceive of the technological race, where technological progress is measured by the rate at which we approach the line established by the TT. Of course, the technological progress line need not be a straight line. Some still believe we'll cross the TT line. I tend to think we might approach it asymptotically, in part because the observer playing the Imitation Game might also get more technologically sophisticated.



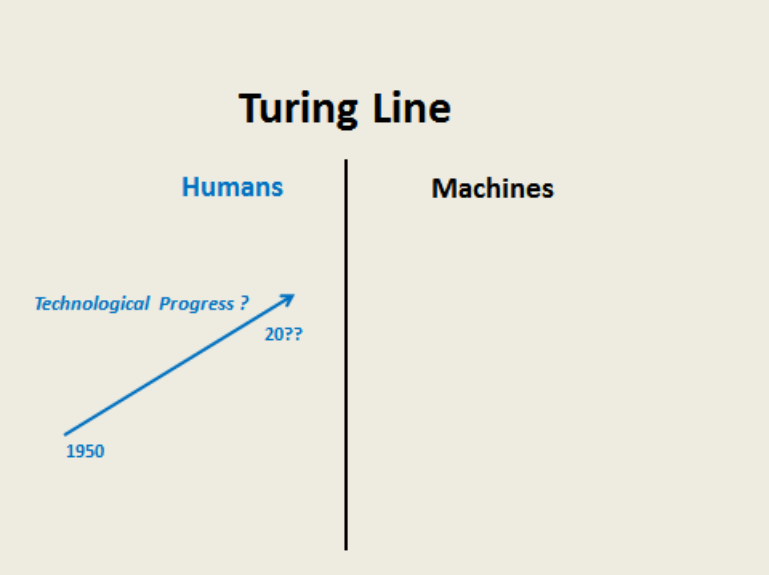
In this article, I examine the other side of the line, the human side. An incredible amount of attention has been paid to machines and if, how, and when they might be capable of passing the TT: *When might a machine be indistinguishable from a human being? What might that mean? How would we interpret evidence of a machine passing the test? Can machines think?*

I ask and examine a related set of questions about humans. There are many reasons it is important to begin asking these questions. I will explore the reasons in separate work. The simplest and most compelling reason is that we are rapidly developing and deploying technologies that operate on the human side of the line by shaping human beings and the

[30] Many experts in the various fields affected by the Turing Test have lamented this racing behavior, for it is by no means clear that this is the best race to run; resources might have been better focused elsewhere, for example. See, for example, Hayes and Ford, supra. In fact, today, researchers in these fields no longer focus on passing the Turing test as a *goal*, for various reasons, including that most recognize that intelligence is much more than verbal communication. I discuss other aspects of intelligence in the next section.

[31] Turing, Alan. 1950. Computing Machinery and Intelligence. Mind LIX: 433-460.

environments within which we live and evolve.



Critically but not obviously, what matters most, more than whether or not we can or ever do actually cross the line, might be what happens during the race -- how the race itself affects humans and society.

I asked a colleague how he would feel about being a mere brain in a vat with his happiness optimized by some technical system. He responded, "Extremely happy, I guess." For him, the stipulation made it easy; he suggested that intuitions derived from the hypothetical just tend to fight the hypothetical and its stipulation. The concerns many people have about the hypothetical may have more to do with doubts about it being possible or whether there is something hidden in the stipulation of optimal happiness. I think there are reasons to be concerned with the hypothetical and its stipulation. But our conversation reminded me that what I may be most concerned with is not the endpoint itself but what happens during the race to get there.[32]

Yet we cannot begin with a completely pessimistic frame. Like imperfect price discrimination,[33] there are beneficial and detrimental outcomes depending on the context. The TT leads us to focus on whether we can develop machines that "think" like humans, and this seems to be a beneficial innovation or improvement because we've added something to the machine; it has *gained* a capability previously possessed only by humans. Approaching the line from the opposite side with a focus on humans, it seems natural to frame the inquiry as *Can humans not-think?*[34]

[32] It is analogous to an argument Frischmann made about price discrimination. Truly perfect price discrimination may be socially valuable; everyone should love it! But it does not and cannot exist in reality. Imperfect price discrimination is much more ambiguous; it is sometimes good, sometimes bad. Frischmann explained how the path to perfect price discrimination is fraught with peril for society, yet we continue down the path deluded by the siren's song of perfection. See Brett Frischmann, Infrastructure: The Social Value of Shared Resources (OUP 2012).

[33] See previous note.

[34] The TT has usually been used to answer the question "at what point can we say that a non-thinking thing (machine) is acting like a thinking thing (human)?," but it also can be used to ask "at what point can we say that a

This framing suggests that as we approach the Turing line something is diminishing and that when we reach it, something will have been lost or taken away; humans will have lost their capability to think, and that seems troubling. But there is a problem with this framing. It is by no means necessary that progressing toward the line from the human side means that something is lost. It might be the case that something is *gained*. That something presumably would be the capability to not-think. Imagine we're playing the Imitation Game with a human seeking to mimic a simple (unthinking) machine and deceive the observer. Perhaps the human can be indistinguishable from a machine by choice, by exercising the capability to not-think. There may be many reasons why this could be an attractive capability, though I am not inclined to explore them now. But we should be clear that the normative or moral evaluation of humans not-thinking is complex.

The question of whether humans can not-think might seem silly in the sense that the answer seems obviously to be: Yes! Of course, human can not-think; we do so quite often, for example, when we act instinctively, impulsively, or emotionally[35] and do not fully consider the consequences of our actions. Notice, however, that this type of reasoning relies on a particular definition of think/not-think. It seems much more plausible to say that while instinctive, impulsive, or emotional actions might not be actions that result from or involve a certain type of rational thinking, they nonetheless involve some type of thinking or more accurately, *mental state*. In fact, these types of mental states seem to be *particularly human* and not machine-like. [36]

This opens up a potentially interesting line of inquiry: Might the mental or intellectual characteristics that in part define us as humans and differentiate us from machines be those (sometimes) associated with *irrational* behavior? Some people influenced by behavioral economics have suggested that when people are likely to act irrationally or in a biased fashion, the response is to debias them or nudge them toward more rational or efficient behavior; notably, such nudges are often implemented by reconstructing the context or environment. Would such efforts to debias people or nudge them toward rational thinking / behavior be dehumanizing? Imagine that we reconstructed the environment to completely eliminate irrational thinking. Would this environment dehumanize? Would humans in such a constructed environment be distinguishable from machines?

It is quite important that the universe consists of much more than humans and machines; humans

---

thinking thing is acting like a non-thinking thing?"

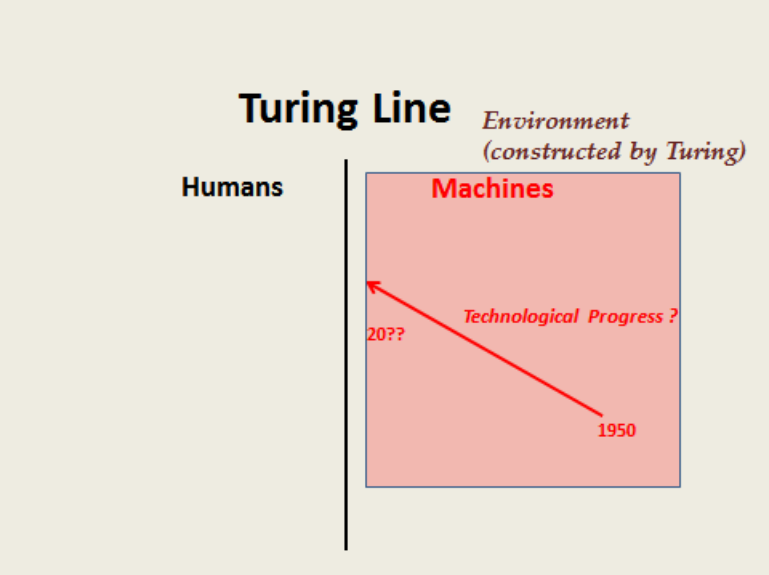
[35] We might add religiously, in the sense that acting with or as a result of blind religious faith can be and has been set in sharp contrast with rational or scientific thought. This may be a can of worms or adders not worth opening!

[36] This discussion implicates the line between machines and other living non-human beings that experience many of these mental states. See Darwin, Charles (1872), *The Expression of the Emotions in Man and Animals*; Goodall, Jane (2000), *Reason for Hope: A Spiritual Journey*. Grand Central Publishing; Cheney, Dorothy L. & Robert M. Seyfarth (1992), *How Monkeys See the World*. University of Chicago Press; Braitman, Laurel (2014), *Animal Madness: How Anxious Dogs, Compulsive Parrots, and Elephants in Recovery Help Us Understand Ourselves*. Regarding animals' experience of emotions, Braitman suggests "A number of recent studies have gone far beyond our closest relatives to argue for the possible emotional capacities of honeybees, octopi, chickens, and even fruit flies. The results of these studies are changing debates about animal minds from 'Do they have emotions?' to 'What sorts of emotions do they have and why?'"

and machines do not exist in a vacuum. They relate to each other and to the environment (and other living beings). The TT depends substantially on the environment within which the test is conducted. Thus, in the standard version of the TT, machines and humans are separated from an observer, and the observer poses a series of questions to identify which agents are human. The machines attempt to exhibit human-like conversational behavior in order to trick the observer into making the wrong identification. Answers from the machines and humans are typed to avoid any biases possibly arising from information that is irrelevant to attribution of thinking / intelligence. Thus, for example, visual cues might enable the observer to distinguish machines and humans, but such information would not be relevant to the underlying question of whether the machines are capable of thinking in a manner indistinguishable from humans.

It matters that the observer is in one room and the subjects (humans and machines) are in other rooms, separate from one another and the observer. It also matters that the investigation focuses on verbal behavior. Verbal behavior (verbal responses to a sequence of verbal stimuli) is deemed an appropriate locus of investigation because of the perceived link between intelligence and verbal output.

Turing imposed significant constraints on the means of observation and communication. He did so because he was interested in a particular capability -- to think like a human -- and wanted to be sure that the evidence gathered through application of the test was relevant and capable of supporting inferences about that particular capability. But I want to make sure we see how much work is done by the constructed environment--*the rules and the rooms*—that Turing built.



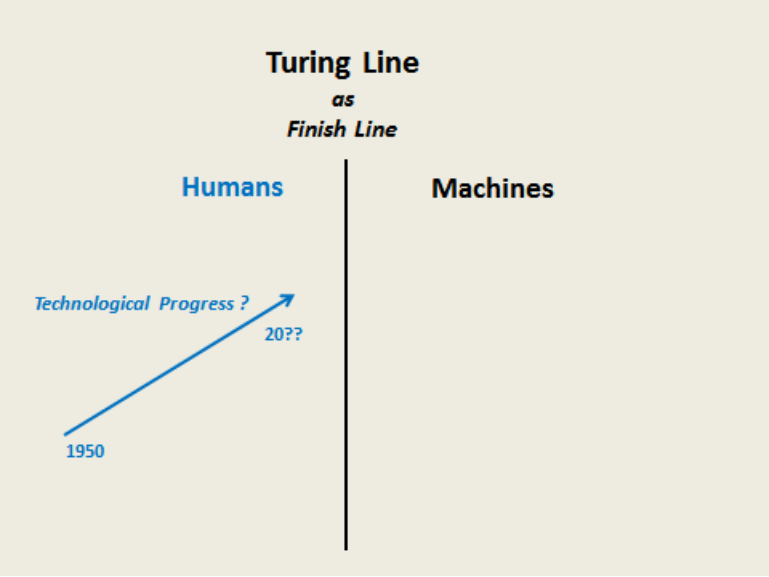
A machine that passed the TT would have passed the test within a very constrained context. In another context or environment, for example, one where the observer could visually observe the subjects or where communication was aural, the same machine presumably would not pass the TT. The machine might be indistinguishable from a human in one context, but easily distinguished in another.

The conventional TT environment itself is constructed; it is designed to take a series of inputs (e.g., machines and humans) and after doing some work performing a process according to predetermined rules, it produces outputs that we can use to draw inferences about the machines (some of the inputs). The TT environment is, in a sense, a machine; we will call it a "machine-environment" to distinguish it from other types of machines, which we will refer to only as machines. Different machine-environments (such as the TT environment) can be constructed with the capability to render machines within the machine-environments to be more or less distinguishable from humans. In a sense, the machine-environments play a very important role in shaping the (actual and perceived capabilities of) machines subject to the TT.

The same can and must be said for humans.[37] In examining the space on the other side of the line (the human side), we need to broaden the inquiry and ask questions like the following:

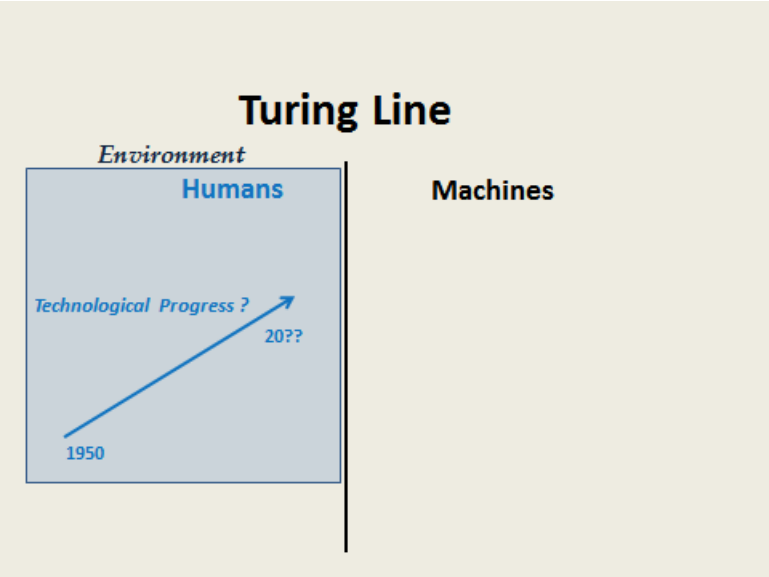
- Under what conditions and/or when are humans indistinguishable from machines?
- Can humans be programmed or constructed to be indistinguishable from machines?
- Can environments dehumanize?
- How and/or when are human beings constructed (via technology, social context, and the environment within which we live and through which our preferences and beliefs are formed) to be indistinguishable from machines?

To see why we must ask these questions, consider how the line established by the TT might function as a finish line when viewed from the human side.



[37] C.f. Thaler and Sunstein, supra. Thaler and Sunstein focus on the choice architecture, which is determined by the environment within which people make choices. Thaler and Sunstein refer to the person who creates the environment as the choice architect. Lessig made a similar observation when he explain how code is law, meaning that technical architecture regulates behavior (and choices) in a manner comparable to law and other institutions, such as markets and social norms. Lessig also emphasized the important role of those who wrote the code and thereby constructed the environment. Lessig (1998).

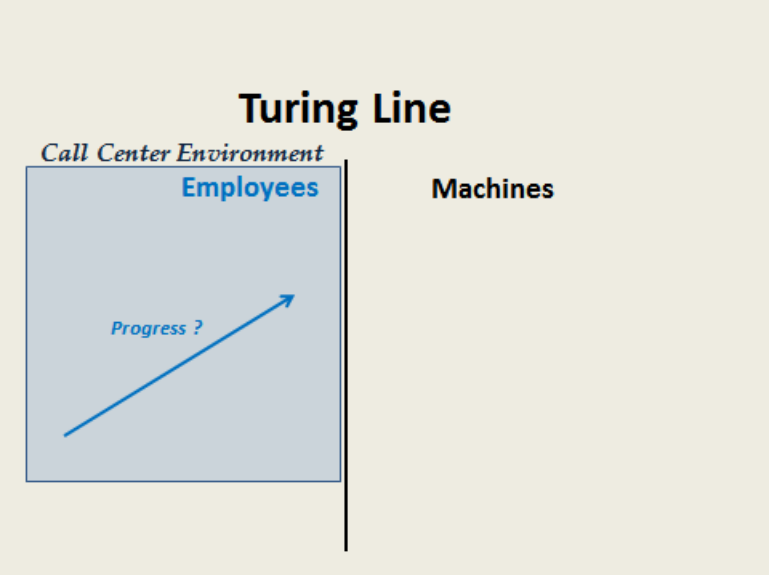
How would you make progress toward the finish line? One way to make progress might involve *directly modifying* human beings through genetic, biotechnological or other means. Another way to make progress would be to construct environments within which humans *are* indistinguishable from machines. Another way would be to construct environments within which humans *become* indistinguishable from machines.



Consider the following example.

Call center employees: In a recent NY Times article, *The Computerized Voice that Wasn't*,[38] the author questions whether American Express successfully created a computer that passed the Turing test because the author was convinced that the call center respondent was a computer that did a pretty decent job of appearing human. The author quoted from his initial conversation and then a follow-up call in which the author asked the respondent if she/he/it was a computer and then proceeded to ask additional questions focused on that issue. Yet at the end of the article, the author concludes with apparent surprise and disappointment that the respondent revealed her/his location to be India and that fact persuaded the author that the respondent was human; this was later confirmed by American Express. What might be inferred, if anything, from the fact that for a while the author mistook the human for a computer? The reason given for the mistaken identification was that English was a second language for the respondent. This provides a slight taste of what is to come later in this Part because language and commonsensical use of language is one defining characteristic of being human. I suspect there are other more subtle reasons that have more to do with the environment constructed by American Express or its call center contractors. That is, it might be the case that the call center itself is a rather constraining, heavily scripted environment that nudges humans within it toward the Turing line.

[38] Segal, David (May 24, 2014). *The Computerized Voice that Wasn't*. NY Times.



Our goal for the remainder of this Part will be to explore these questions on the human side of the line and to reorient and reframe the Turing test, by making a human being the subject and testing whether the human being is (in)distinguishable from a machine. The context within which the test applies will be important, not only because the context shapes the test but also because the context shapes the participants. It may be the case that, in the end, we will be testing the combined effect of humans and environment (or humans situated within specific environments).

C. Intelligence: Thinking and not-thinking

Our first step in investigating the human side of the Turing line is to ask the simple question: *Can humans not-think?* Of course, this question is not so simple and presents us with the same intractable definitional questions that Turing faced (what do the words "human" and "not-think" mean?), and thus, we might look to avoid such complexities by clever resort to parlor games. Keep in mind that the game serves an empirical function.[39] Also, for the reasons noted in the previous section, this question might not ultimately be what is most interesting to us. If we are really interested in not only *whether*, but also *when* and *how* humans can be constructed to be indistinguishable from machines, then thinking / not-thinking might be insufficient. Nonetheless, it is a decent place to start.

Suppose we administer the Turing test in its conventional form. *Does that tell us anything meaningful about the human participants? When the observer correctly determines that humans are in fact humans, do we learn anything?* Probably not. But what about when the observer incorrectly determines that a human is a machine? This occurred a few times during Loebner competitions. For example, according to Stuart Shieber, "Ms. Cynthia Clay, the Shakespeare aficionado, was thrice misclassified as a computer. At least one of the judges made her classification on the premise that '[no] human would have that amount of knowledge about Shakespeare.'"[40] It seems hard to infer from the particular misidentification of the human as a machine just described that the human was indistinguishable from a machine because the human was not-thinking. That is, when we examine the agent in the context and in light of the questions asked by the observer, and we evaluate plausible inferences about the human agent, not-thinking doesn't rise to the top; others do, and they have more to do with the observer and peculiarities of the human participant. This does not necessarily undermine our search for a test to investigate whether humans can not-think. But it suggests, as we might expect, that we need to pay careful attention to the *structure and rules of the game*, which may need to be tailored to eliciting evidence relevant to the question at hand.

If we stick to the basic structure with an observer in one room and agents distributed in separate rooms from the observer and each other as well as the verbal stimuli communicated by text, we can adjust who knows what about the game being played--in other words, what the observer and agents know about the game they're playing and who is effectively "playing" the game.

Here are a few variations:

- **a.** Play Turing test with same rules; no difference at all; examine situations where humans have been mistaken to be machines.
- **b.** Play Turing test as usual, except humans are told that they should try to deceive the

[39] Some reviewers have suggested that we don't need to rely on games at all. The Imitation Game has been seriously criticized, they say, and the underlying concerns about the construction of humans to be machine-like can be addressed more directly. I am not so sure about this position, however. I believe the value of the Turing test is not necessarily limited to the particular machine-environment (i.e., the Imitation Game) he constructed; the method matters, as does the underlying sets of questions it invokes.

[40] Shieber, Stuart, *Lessons from a Restricted Turing Test*, Communications of the Association for Computing Machinery, vol. 37, no. 6, 70-78 (1994).

observer; the observer doesn't know about this additional instruction.

c. Same, but the machine programmers also are informed that humans will play strategically.

d. Same as b, but the observer knows.

e. Same as c, but the observer knows.

The implications of using the imitation game to learn something about human intelligence would be different under these different conditions. When humans are playing the game and aiming to deceive the observer, for example, we would need to deal with the fact that humans can easily mimic some machines (like a car) by not communicating at all. Version **b** would not appear to work well because observers would likely perceive non-responsive agents to be machines and little could be inferred from such results. Version **c** might overcome this problem because machines could adopt a similar strategy, and versions **d** and **e** might overcome it because the observer would be able to anticipate the strategy.

It is not clear that on the human side of the Turing line, we need/want both humans and machines (or the humans programming machines) to play the game strategically. Consider the role that human participants play in the conventional Turing test. The humans are not strategic agents, seeking to confuse or deceive the observer. Rather, the humans act as "normal" humans and thus serve as a baseline. Similarly, just as humans served as a baseline and did not act strategically to confuse or deceive the observer, we should use the *simple machine*[41] as a baseline and not ask the programmer of the machine to act strategically by confusing, deceiving, or outsmarting the observer. For what matters to us, ultimately, is not how machines can be built to imitate or mimic humans. Rather, we are interested in (the construction of) human beings and in particular when a determination that a human is indistinguishable from a machine would allow us to reasonably infer that something meaningful, in this case, (not-)thinking, has been (gained) lost. It seems reasonable to make such an inference--subject to competing ideas or refutation, of course--when a human is deemed indistinguishable from a simple machine.

Accordingly, we might use one of the following variations:

f. Same as any of the above, except the machine agents are *simple machines*, or at least, are not deceptive or otherwise strategically playing the game.

g. Same, except the observer is a computer.

The question remains whether it makes sense to ask the human participants to play strategically. What would that mean? That the human agents are seeking to behave like machines? Success might indicate, or provide evidence to support the inference that, the human mistaken to be a machine possessed the capability to not-think. That is, if not-thinking is a capability to be exercised, then passing the test could be relevant.[42]

But if, on the other hand, we are not interested in the affirmative exercise of the capability to not-

[41] Simple has a particular meaning in this context, as described in the text.

[42] For example, imagine the criminal defendant exercising the capability to not-think in an effort to persuade a clinical psychologist that she lacked emotion or the cognitive ability to distinguish right from wrong or to understand the consequences of her action or to have intention.

think and instead are interested in identifying a diminished capacity to think or perhaps inadvertent or environmentally constructed not-thinking, then we might need a different approach. In such a case, it might be more appropriate to set up the game as if it were a conventional Turing test (version **a**) except utilize version **f** or **g**; in other words, the human participants would be instructed to simply "be themselves." Thus, having removed strategic behavior altogether, imitation is less relevant than comparison, and the burden shifts to the observer (chosen stimuli) and the constructed environment to differentiate humans and machines.

Beyond the basic structure and variations described above, we need specify how we aim to test for not-thinking. Of course, we might leave it entirely to the observer and the verbal stimuli chosen by the observer. But for various reasons, that might not be terribly effective. Even in the context of the conventional Turing test, various constraints shaped the observer's questions. Most basic was the restriction to verbal stimuli communicated by text. Moreover, when the test moved from thought experiment to applied experiment, even more fine-grained constraints on the types of verbal stimuli permitted were introduced.[43]

The reason for restrictions on stimuli is simple: It is to increase the likelihood of generating objective empirical evidence with which inferences could be reasonably made. We need to engage in a similar winnowing process. Specifically, we need to consider more deeply what aspects of thinking or not-thinking can be tested. We could focus on some specific intelligence-related characteristics that have often been identified in the definitional debates about what makes us human. Here is a short list of some (potentially overlapping) candidates:

- Reason
- Rationality / Irrationality
- Common sense
- Willpower
- Emotion
- Phenomenological experience[44]
- Creativity
- Language / capacity to construct new language or social meaning
- Planning for others / for the happiness of others
- Language with which to plan for future[45] / others[46]

[43] See, e.g., Loebner competition rules.

[44] E.g., the capacity to feel and understand the feeling of hot/cold, hunger, etc., or to see and understand the color red. Elsewhere, I develop an adaptation of Frank Jackson's famous Black-and-White Mary thought experiment to examine phenomenological experience and technological manipulation of the human capacity to feel. *See* Jackson, F., 2007, *The knowledge argument, diaphonousness, representationalism*, in T. Alter & S. Walter 2007: 52–64; Jackson, Frank. 1982. *Epiphenomenal qualia*. Philosophical Quarterly 32:127-136.

[45] "To be human," writes Dan Falk, "is to be aware of the passage of time; no concept lies closer to the core of our consciousness". "Without it, there would be no planning, no building, no culture; without an imagined picture of the future, our civilization would not exist." Falk, Dan (2010). *In Search of Time: The History, Physics, and Philosophy of Time*. St. Martin's Griffin (1st edition).

[46] Cf. Goodall, supra, at 188: Humans have the capability to use language to express ideas about objects and events that are not present. Although Chimpanzees and other intelligent primates have complex communication systems, they do not have the ability to communicate about things that are not present. This uniquely human capacity enables people to plan for future events and recall past events. Most importantly, according to Goodall, human language allows members of a group to discuss ideas and share a "collective wisdom." This relates to common sense.

With one or some of these intellectual capabilities in mind, we might specify the types of stimuli or questions that the observer would be allowed to use. In doing so, we would want to ask ourselves: What could be inferred when an observer mistakes a human to be a machine?

To be clear, I am not seeking to explain or prove the existence of any particular defining characteristics of humans or machines. I assume their existence. For example, I assume intelligence exists, and it is not my goal to define it or explain its origins or underlying mechanisms. Let others define and debate what is or is not intelligence, or more broadly, what is or is not essential to humanity. For now, I am interested in *marginal* changes induced by technological environments, and I aim to investigate changes in our capabilities so that we can have a more meaningful discussion of what is or is not essential. I note this upfront to avoid getting sucked into the vortex of existing philosophical and definitional debates. That said, obviously some working definitional baselines remain relevant, as will be discussed in the context of specific tests.

This paper does not explore all of the possible tests. We consider four tests that distinguish humans from machines based on (1) *mathematical computation*, (2) *random number generation*, (3) *common sense*, and (4) *rationality*, respectively. We briefly discuss the first two and devote more attention to the third and fourth. All four are plausible reverse Turing tests that generally could be used to distinguish humans and machines. Yet the first two do not implicate fundamental notions of what it means to be a human; the third and fourth do.

For each test, we begin with a brief description of what it is that we are testing, for example, by specifying the stimuli used by the observer, and then discuss how to interpret the results, for example, by exploring whether passing or failing the test would support meaningful inferences about the human agents.

1. Reverse Turing test focused on mathematical computation

Suppose we set up our reverse Turing test in the conventional manner (**a**), except that the machine participants are simple machines (**f**) and the observer is a machine (**g**) programmed to submit a series of mathematical computation questions to the agents and then after receiving answers determine whether the agents are humans or machines. The types of mathematical computation questions could progress from performing simple to increasingly complex calculations (e.g., addition and subtraction of single or double digit numbers, to multiplying a single and double digit number, to multiplying double digit numbers, to square root of triple digit numbers, and so on).

There are at least two ways in which humans generally would be easily distinguished from machines. First, humans get tired and are prone to error as fatigue sets in; simple machines do not get tired. Depending on the duration of the test, and possibly on the rate at which questions are asked, humans will make computation errors even for relatively simple computations; machines will not. Second, humans will make computation errors for more sophisticated computations; machines will not.

In general, most humans routinely would fail a reverse Turing test focused on mathematical

computation. But not all humans would fail this test, at least within some bounds. Suppose, for example, we limit the duration of the test to five or ten minutes or otherwise eliminate the first source of human error. Suppose that under this scenario a human being passes the test. *What would that mean? Are there any meaningful inferences to be drawn? What would we learn?* As I explained in section 1, passing the test provides us with evidence and suggests that there is *something* remarkable about the relevant agent.

The human being would appear to possess an extraordinary—unhuman—capability for mathematical computation. The extraordinary capability appears to be a positive or valuable addition. Does being machine-like mean the person was somehow *less* human? Of course not. Being machine-like is this particular context and in this particular manner does not seem meaningfully related to the agent's (normative) status as a human being; in other words, there is evidence of something, but not of dehumanization.

We also might ask whether we can say anything meaningful about the humanity of the vast majority of humans who failed to pass the test. I don't think so. We might conclude that some of the humans who lacked (some) mathematical computation capabilities might be worse off than those who have the capabilities and that education or other means of improving their situation would be worthwhile. But that is a very different normative issue.

2. Reverse Turing test focused on generating (not so) random numbers

Assume a similar set-up as the previous test, except the observer asks the agents to generate random numbers, one every five seconds for ten minutes. Based on the responses and the observer's ability to predict the agent's nth response, the observer determines whether the agents are humans or machines.

Generating random numbers is difficult for both computers and humans, but they face different difficulties, which makes a reverse Turing test plausible. Computers are deterministic because humans program them. This means that computers do not generate truly random numbers and instead produce pseudo-random numbers by running a seed number through a complex algorithm; the result is not truly random because it is determined by the seed number and algorithm. If one knows or can reverse engineer the algorithm and seed number, then one can predict the numbers that the computer will generate.

We could assume the observer either knows or can figure out by reverse engineering the algorithms and seeds used by the simple machines. The assumption makes it easier to distinguish humans and machines, but it might not be necessary.

Humans would routinely fail a reverse Turing test focused on generating random numbers not necessarily because humans would generate more or less randomness than machines. Humans generally would be distinguishable from machines because the methods for generating numbers would differ substantially.[47] Some people would misunderstand randomness, for example, by

[47] See, e.g., Wagenaar, W. A. (1972). *Generation of random sequences by human subjects: A critical survey of literature*. Psychological Bulletin, 77(1), 65-72; Brugger, P. (1997). *Variables that influence the generation of random sequences: An update.* Perceptual and Motor Skills, 84(2), 627-661; Towse, J. N., & Neil, D. (1998).

seeking to avoid repeating a number too often ("I can't use 9 for a while because I just used it."). Some would exhibit preferences for certain numbers, other for certain sequences of numbers or alternating patterns. Most humans would not employ a seed number and algorithm. Some humans might be unpredictable, which would be a basis for distinguishing them from the machines. Some humans would be predictable, but for different reasons than the simple machine.

Suppose a human being passes the test. Again: *What would that mean? Are there any meaningful inferences to be drawn? What would we learn?*

Once more, the human being would appear to possess an extraordinary capability, in this case, for generating pseudo-random numbers in a machine-like manner. In contrast with machine-like capability for mathematical computation, which seemed advantageous, it is not clear how to evaluate this capability. Regardless, being "machine-like" is this particular context and in this particular manner does not seem meaningfully related to the agent's (normative) status as a human being. Finally, there is not much to be said about those who fail the test.

3. Reverse Turing test focused on common sense

Common sense is a familiar human characteristic. It is a concept with a long history and hotly contested meaning in various disciplines, and you undoubtedly have your own conception of common sense in mind. Yet, once more, our objective is to skip past definitional debates and adopt a particular understanding of common sense. Erion offers the following, which we adopt:

> A more focused notion of commons sense … [is] virtually universal among typical adults because it concerns an important subset of objective reality that we all live our everyday lives in, the common-sense world. As rough approximation, it is helpful to think of the common-sense world as the realm of familiar objects that we become acquainted with during ordinary experience. People, plants, non-human animals, and simple geographic features are all included in this world, while sub-atomic particles, neurons, and galaxies are not. … [We] can understand common sense itself as the base of knowledge about common-sense reality that allows each of us to survive and thrive during our everyday lives. Common beliefs about the common-sense world are the most prominent components of this knowledge base. … common sense also includes the widespread abilities that allow us to act successfully in the common-sense world.[48]

Erion goes on to explain how work in "various cognitive sciences" supports his claim that this type of common sense exists. It entails core knowledge and skills that are shared and "used by all of us (even skeptical philosophers) during our everyday lives." Language is critical to common sense both as knowledge and as skill. That is, competence in using language is a "subset of common sense."[49]

Analyzing human random generation behavior: A review of methods used and a computer program for describing performance. Behavior Research Methods, Instruments, & Computers, 30(4), 583-591.

[48] Erion, Gerald (2001), *The Cartesian Test for Automatism*, Minds and Machines 11: 29-39, at 33.

[49] Id. at 36.

Erion suggested that despite some contrary interpretations among philosophers, Descartes' action test focused on common sense as a characteristic that distinguished human beings from automata. Erion reformulates Descartes' two-pronged test as follows:

> Automata are distinct from real people in two ways. First, automata cannot use language. Second, automata do not possess common sense, which includes not only knowing how to use language but also knowing how to perform tasks and answer questions that even the most simpleminded adult human can.[50]

Thus, a common sense test could employ a structure similar to the conventional Turing test, and the observer could ask questions that would "require[] the skillful use of common sense." Erion explained that such a test would be more demanding than the Turing test; that is, a machine that passed the Turing test still might fail the Cartesian (common sense) test. The idea is that by observing agents over a "significant length of time in a variety of circumstances," we confidently can distinguish humans from machines based on linguistic abilities and commonsensical performance. Erion focused on the machine side of the Turing line and viewed the common sense test as a high bar for machines, in the sense that it would be difficult for a machine to be mistaken for a human being (i.e., observed to possess common sense).[51]

On the human side of the Turing line, the common sense test would appear to set a high bar for humans. Putting aside strategic and deceptive behavior, it would be difficult for a human to be mistaken for a machine (i.e., observed to be devoid of common sense). Erion suggests that "even the most simpleminded adult human" would pass the common sense test.[52]

So why develop a common sense test on the human side of the Turing line? It is a useful characteristic to focus on in part because it is often distinguished from other measures or types of human intelligence. As one of the usage examples from Merriam-Webster's dictionary suggests: "She's very smart but she doesn't have a lot of common sense." Common sense also combines language, reasoning, and social skills in a fashion that may or may not be unique to humans, but nonetheless usefully differentiates humans from machines.[53]

---

[50] Id. at 36.

[51] Erion notes that "it is everyday knowledge that is hardest to convey to a computer" and concludes that accordingly the Cartesian (common sense) test is a rather high bar for machines. One might wonder whether Big Data and networked machines will undermine this view. IBM's WATSON computer or even SIRI may seem to exhibit common sense because of the computers' ability to interpret questions in context and respond in ways that seem to correspond to common sense. Of course, any such responses are derivative of human-generated data. For example, if a computer evaluates observed human responses to being lost (suppose the iphone tracks user behavior and makes the data available to SIRI) and determines what is common sense based on the frequency of response (and perhaps subsequent behavior as well as a measure of success), does the computer possess common sense? Note that we have not directly conveyed everyday knowledge to the computer; instead, we have provided the computer with the tools for identifying and then responding with the statistically deduced response. But does the computer actually possess everyday knowledge? Can the computer apply the knowledge to reason or "to act through reason?" Of course, this brings us back to Searle's Chinese language experiment and the basic mind-body problem-- Does the computer know anything? As Erion suggested, an observer with sufficient time probably would be able to determine that such a computer was in fact just a computer. Id. See also Hubert Dreyfus's classic book, *What Computers Still Can't Do: A Critique of Artificial Reason* (1992).

[52] Erion, supra, at 36..

[53]Other species may possess common sense. Chimps seem to have a weaker version in the sense that they solve

Let us assume, for a moment, that a human passed the common sense test. What would it mean if a human were indistinguishable from a machine based on the human's performance in a common sense test? It seems more difficult to imagine contexts within which humans lack common sense. Common sense, as a concept, seems to be sufficiently adaptable to different contexts, essentially by definition, since it is what is common to our everyday lives, whatever that might entail or however we might live our lives. But it would be a mistake to assume stability or a persistent reservoir of common sense available to us. Common sense depends upon a shared core knowledge base, language, and social interactions sufficient to generate common understandings and beliefs. Suppose (access to) these inputs are restricted. There are various ways to imagine such restrictions.

Consider the following thought experiment. Suppose Alice gets in a taxicab, gives the driver an address, and then falls asleep. Thirty minutes later, the taxicab driver wakes Alice, takes her money, and leaves immediately after she exits the vehicle. After shaking off her initial grogginess, Alice realizes the cab dropped her off in the wrong location, and she is lost. What does common sense dictate (suggest) she do? She should get her bearings, formulate a plan, and take action. How would she do this? Presumably, by looking around, observing people, reading street signs, and so on. All of this sensory information would provide her with baseline information that would help her evaluate her situation and options and decide on a course of action. Based on such information about her environment, Alice would be able to form beliefs about her safety, whether she could trust people, whether people would understand her (speak the same language), and so on. She might be able to determine the likelihood of another taxi cab arriving or whether some form of public transportation was accessible nearby. She might be able to figure out or approximate her location and then formulate a plan for getting safely from there to her intended destination.

Now suppose that Alice lacks the relevant situational and problem solving common sense. What does this mean exactly? How could this be? Perhaps she lacks the ability to get her bearings through the various means just described. She is unable to take in and translate the various cues and information. Perhaps her incapacity stems from a physical or cognitive disability; perhaps she has never had the necessary experiences that would have led her to develop the relevant abilities--for example, her prior navigation of the world may have been technologically mediated and fully automated (more on this below); perhaps she was raised in a town with no street signs and thus would not think to look at street signs for location data. She might lack (the relevant situational and problem solving) common sense because she has never discussed the problem of being lost with anyone else or contemplated the situation in which she now finds herself. That may seem hard to fathom, but being lost may be a problem of the past, at least in the near future.

In our everyday experience, it is highly likely that Alice would carry with her a device capable of determining her location, determining an efficient route to get from her location to her intended

---

problems and share solutions; other animals are social and cooperate in ways that suggest some means for sharing knowledge. But it is not clear that any animals other than humans combine language, reasoning, and social skills to generate common sense, at least as we have defined it. Cheney & Seyfarth, *How Monkeys See the World* (1990) discusses these issues, documents various ways in which non-human primates communicate and develop "social knowledge," and cite various studies.

destination, and even ordering a taxi. Upon recognizing her plight, Alice need only pull out her smartphone, and she'll be on her way to her intended destination. She need not ask anyone for directions, learn very much about the environment in which she finds herself, or do much planning. Does this mean she lacks common sense? No, not necessarily. In fact, the common sense reaction to her predicament is probably to consult her smartphone and use the taxi-ordering app.[54] One might say that common sense dictates that she carry a smartphone in the first place so she will never truly be lost.

It might be the case that common sense often dictates resort to technology in one form or another--whether a map, cellphone or smartphone, and (some would say) that it is thus incorrect to suggest that common sense is weakened or diminished, much less lost. But the technologies are not neutral or equivalent. This is important to bear in mind. The degree to which a person such as Alice relies on (i) herself, her common knowledge base of beliefs, experiences, and so on, and on other human beings (both directly and in terms of common sense itself), rather than (ii) a technological device/system varies quite substantially across different technologies (e.g., from map to to cellphone to smartphone). The shift from (i) to (ii) is relevant and important.

We might think that common sense becomes embedded in the technological devices and/or that shifting from (i) to (ii) entails a shift in the relevant community of human beings that Alice relies on--that is, from the community of people Alice knows and shares common experiences with to the community of people behind the technological system. Either way, the shift is remarkable and worth examining. It is by no means limited to thought experiment we've discussed; one could formulate a similar thought experiment around what common sense dictates in a variety of everyday circumstances, such as when one feels a sharp pain in her back or when there is a power outage. Doing so would lead to the same observations, shifting from (i) to (ii). [55]

As Evgeny Morozov suggests in *Every Little Byte Counts*: "As we gain the [technological] capacity to predict and even preempt crises, we risk eliminating the very kinds of experimental behaviors that have been conducive to social innovation."[56] He was not focused on common sense per se, but I believe he identifies the underlying dynamic.

The thought experiment is mainly intended to explore what I mean by common sense and reveal how common sense is susceptible to technologically-induced change. I considered discussing variations on the thought experiment, such as what common sense dictates when one is lost while driving on a road trip and how GPS has decreased reliance on certain forms of shared knowledge, skills, and experiences; one no longer needs to consult a map or stop at a gas station and ask for the assistance of strangers. Of course, there is nothing new to the observation. Many have made this point before. I am neither lamenting nor celebrating. The simple point I wish to make is that the precursors or inputs to (the creation and sharing of) common sense are not stable or inevitably accessible and shared. At least for an important subset of common sense, one of

[54] For fun, I turned to Facebook and Twitter to find out what common sense suggested Alice should do. Remarkably, many commenters suggested she order a cab using Uber. Among other things, they assumed she had a smartphone and felt safe enough to use it in public. After a number of comments, one commenter expressed surprise that no one had suggested Alice simply ask someone for help.

[55] It would be interesting to examine such a shift in the healthcare context, where the relationships between common sense, technology, and medical treatment are evolving rapidly.

[56] Morozov, Evgeny (2014). Every Little Byte Counts, The New York Times Book Review 23, 5/18/2014.

the inputs (prerequisites) seems to be a problem to solve, and one that leads to social innovation through shared experiences and beliefs about how best to deal with the problem. If technology solves the problem, there is no need for common sense solutions. In a sense, humans and machines behave indistinguishably in that context, but not exactly (or not only) because they solve the problem in the same fashion as might be the case in the context of a rational choice style problem. Rather, in this context, they are indistinguishable because a realm or subset of common sense experience and knowledge that would otherwise be shared by humans (and not by machines) doesn't exist.

Consider the following set of arguments:

1. Humans face common problems in everyday life ("everyday life problems").

2. Humans develop and rely on common sense solutions to everyday life problems.

    a. Developing common sense solutions necessarily depends on a shared core knowledge base, language, and social interactions sufficient to generate common understandings and beliefs.

    b. Developing common sense solutions [necessarily? often? usually?] depends on experimentation and social innovation.

3. Humans develop technology to solve problems.

    a. Developing technology to solve a problem depends on knowledge, experimentation, and innovation, but not necessarily on a shared core knowledge base, language, and social interactions sufficient to generate common understandings and beliefs.

4. Some technology solves everyday life problems.

5. If technology solves an everyday life problem (more efficiently than existing common sense solutions) then humans will not (are less likely to) develop common sense solutions to that problem.

6. If technology solves all everyday life problems, then humans will lack common sense (or a subset of common sense that concerns problem solving).

7. Humans without common sense are indistinguishable from machines, at least in one (important) respect.

The first four statements are uncontroversial. The same is true for the soft version of the fifth statement (read with the parentheticals). The stronger version of the fifth statement is questionable. Surely humans may continue to develop common sense solutions to everyday problems for which there is a technological solution; it is just a cost-benefit calculation.

The sixth and seventh statements require more explanation. The idea that technology will eliminate the need for common sense solutions to everyday problems may seem far-fetched,

mainly because it is hard to believe that technology can so comprehensively address human needs. Moreover, perhaps technological solutions to present-day everyday life problems merely shift the demand curve, so to speak, making a range of problems that were more extraordinary less so and thus potentially amenable to common sense solutions. There also is an intermediate step missing between the fifth and sixth statements; something that would explain the aggregation of incremental substitution of technology for common sense.[57] The seventh statement is just a restatement of the premise behind the common sense test.

My objective in this section is primarily to reveal the arguments. There is plenty of work to be done in modifying, defending and extending them, and in exploring the complex and varied relationships between technology and common sense in particular contexts. [58]

[57] The missing step is important: It is difficult to say how "close" or "far" from the Turing line we might be at any given time, although it seems likely that we remain quite distant. Perhaps we can posit that we are progressing toward the line, but even then, at what rate? What is the shape of the "progress curve" or path? Would such a curve be linear or nonlinear? Would there be a tipping point? (A phase transition as we approach the boiling point?) As I suggested in the Introduction, we might be concerned with changes or losses along the way, even if we never truly cross the Turing line.

[58] Suppose technological substitution for common sense (step 5) creates different everyday life problems and return us to step one. (We might envision this as moving parallel to the Turing line.) For example, people might take greater risks than they would in the absence of technological devices such as GPS. As a reviewer noted:

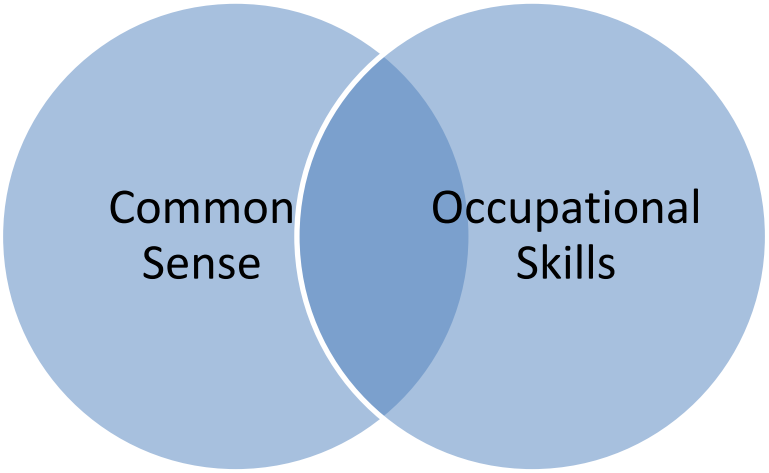
> I may take bigger risks. My mother never even drove on highways. With my GPS, I go places I wouldn't be confident visiting otherwise (long highway trips; errands in Newark). So I venture into situations that may challenge and develop my common sense in a different way.

Greater or different risks might lead to new types of common sense. She explained that the particular example might work as a metaphor for other situations.

> For example, technology affects the traditionally "female" work of networking/connecting friends and relatives. The required common sense changes. But maybe some of this new common sense is even more nuanced and refined, given how many more interactions are happening, with fewer useful social cues.
>
> Another example to consider: I think my students are savvier researchers as college freshman than my generation was. (And if they plagiarize from Wikipedia or some random source and say they didn't know better, they are pulling the wool over your eyes. Students of this generation who do that are lazy, not ignorant.) They've spent their whole lives evaluating on-line info. They'd never have survived to 18 if they believed everything they read. They have a researcher's common sense that in my generation we really didn't develop that acutely until grad school.

The workplace is an important context to explore. Workers hone occupational skills to solve problems, and thus work (labor; task performance) entails a broad category of problems for which common sense and technology might be rivals. Recall the call center story in which a human call center receptionist was mistaken, for a time, to be a machine. I suggested that the call center environment might have constrained the receptionist and that the scripts she followed made her appear machine-like. The caller might also have been influenced by his prior experience with call centers, which increasingly rely on machines to automate tasks. Advances in speech recognition have opened the space for competition between humans and machines in many occupations. In this competition, Stuart Elliot (2014) suggests, "we may overlook important skill requirements for some occupations, such as the substantial range of common-sense knowledge that enables a receptionist to reply sensibly when a customer makes an entirely unexpected request." Of course, what constitutes a "skill requirement" is not fixed; a sensible reply to an unexpected request may turn out to be inessential for call centers operations and a (forgotten) luxury.

Automation within the workplace and what it means for society raises a host of complex issues, some of which relate to this project and some of which do not. It has a long, rich history and is currently incredibly important, maybe more so than ever before. But it is too big a topic to dive into at this point. For now, note how common sense relates to occupational skills:



Many of the arguments raised in the debate about the erosion of occupational skills or competition between humans and machines for meaningful work take the same form as the arguments about common sense and at the same time face the same counterarguments and contextual nuances.

We have questioned the stability or vulnerability of common sense on the human side of the Turing line and asked whether and how technology renders common sense less relevant, necessary or even obsolete. Yet, as with our forthcoming discussion of (ir)rationality, the constructed machine-environment does a lot of work, in that case by nudging people toward rational choice and in this case by lessening human dependence on common sense as a means for solving problems.

But the point is not limited to problem-solving. The constructed machine-environment within which humans are situated may also restrict access to the inputs necessary to create and sustain common sense. Keep in mind that common sense, as we've defined it, depends upon a shared core knowledge base, language, and social interactions sufficient to generate common understandings and beliefs. We could rewrite the set of arguments above to focus on the relationships between technology and these inputs and once more posit technological substitution (steps 3-5), a corresponding elimination of common sense (step 6), and indistinguishability / dehumanization (step 7). [59]

Let me return to my unsubstantiated claim that in the not so distant future being lost may not be a problem most humans experience. The claim is based on the speculation/premise that most humans will carry, wear, or have implanted a device which tracks their location and provides instantaneous navigational instructions. Many already do, as was evident to me when my question about the commonsensical response to being lost appeared nonsensical to many people. The being lost thought experiment sets up another, which moves beyond common sense.

Consider the following scene from a fiction novel, *Gaming Darwin*:

> 2013: Hundreds of people walking. The sidewalk was congested, and one out of every three walkers was a stumbling, bumbling idiot—either meandering like a snake or stopping suddenly like a meerkat, chatting away on a cell phone or worse, swiping and thumbing a screen on their mobile devices, oblivious to everyone else around them, just not giving a [hoot] about anyone beside themselves and whoever it was that they were interacting with, if it really was an actual person and not the latest cat video on YouTube. Meandering snakes and sudden stop meerkats, annoying ....
>
> 2013: I-80 during rush hour. Murderous rage and frustration, and for some desperation—those poor souls who had to pee! ... The highway was congested …, bumper to bumper traffic that swelled and surged and then suddenly stopped in a flash mob of red taillights. … [He] thought of a huge swarm of [meerkats] running full speed and then freezing at the first sign of danger. ...
>
> A few years later: I-80 during rush hour. Elation. Relief. Traffic was moving, managed, in synch. ... The cars were equipped with auto-drive systems and received data from the highway sensor network. Ants. Awesome sensing, communicating, cooperative management systems. Content ants.

[59] Note that since common sense is dependent on shared core knowledge, uneven distribution or deployment of common sense destroying technology could have distributional impacts.

> 2020: Google announced a new version of Google Glass. This was a game changer. Revolutionary. So long cell phones, smart phones, hand held mobile whatevers. Until now, there had been healthy competition, and Google Glass struggled to gain market share in the mobile communications and computation sector. But this changed everything. ... no one expected the synergistic combination. … The glue technology, the one that made it possible, was the motor function management software and the interface through Google Glass with the human brain and body. Initially, the tech was developed as a small independent project to help accident victims who were paralyzed or lost control of certain parts of their bodies. ... Who'd have thought to combine the three technologies—Google Glass, automated, self-driving cars, and the motor function management system? Utterly brilliant. [He watched hundreds of people walking and marveled:] Snakes and meerkats to ants. ...[60]

The scene describes a wearable technology that allows a person to delegate the mundane task of physical movement through the world to a complex navigation, sensory and motor function management technology. The novel later goes one step further and describes implanted chips that modify humans in part by connecting them to ubiquitous sensor networks. Of course, this is science fiction, but so are many thought experiments. What would it mean for humans to delegate these mental and bodily functions--navigation, sensing, and movement--to a technological system? On one hand, navigation through the physical world seems essential to being part of the world, to understanding it and others with whom we share it. Given how environments construct us as humans, it seems troubling to tune out or to remove/disrupt our direct, physical and sensory awareness of and connection to it. Yet, on the other hand, humans modified by this technology are no less human just because they choose to delegate various "mundane" tasks to a technological system; to the contrary, by doing so, the humans arguably free themselves from the mundane and can choose other more pleasing or more productive intellectual activities. Each person only has so much time and attention.

Is there any meaningful difference between a person employing the navigation, sensory and motor function management technology and a person sitting in a self-driving car? a taxi? the passenger seat of a friend's car? Again, as noted in the being lost thought experiment, there are differences across the technologies, and the differences in capabilities (not) exercised or practiced by the humans using the technologies, but it is difficult to say when, if ever, the Turing line is crossed, such that we might infer that the human is indistinguishable from a machine in the sense that the person is not-thinking. Of course, the person might "look" like a machine to an observer, and a test based on visual observation might be passed. But that would tell us very little about not-thinking. It is likely impossible to fashion an appropriate test without knowing, or at least inferring, what is going on in the mind of the person who uses the technology. If otherwise mentally engaged, we might infer that the person has extended her mind; perhaps the technology is additive rather subtractive. Recall the point in the beginning about the capability to not-think. In separate work, I explore the theory of the extended mind and mind-extending technology,[61] which prompts a somewhat different question: *Who is doing the thinking?* This raises important issues concerning autonomy and free will that are beyond the scope of this

[60] Manuscript on file with the author.
[61] Clark and Chalmers 1998; Clark 2011; Robinson (2013); Malafouris (2013).

article, although admittedly lurking throughout.[62]

4. Reverse Turing test focused on rationality

Rationality is studied intensely in a number of different disciplines, ranging from psychology to philosophy to economics. It is central to models of human decision making and serves as a baseline for evaluating performance in various settings. "What does it mean to be rational, and to make a rational choice on the basis of a meaningful and relevant distinction?" Oscar Gandy, Jr. explains:

> Defining a concept by means of its opposition is rarely satisfactory, but it is a place to begin. Irrational decision-making is commonly associated with emotional or habitual, responses, informed by broad generalizations, rather than by careful weighing of the relevant facts. Rational decision-making generally refers to the process, rather than the outcome or results of any decision, although we understand that a carefully considered decision arrived at following a process of extensive search, reflection, and analysis, can still produce unsatisfactory results. A realization that there are constraints on the ability of humans to access and incorporate all relevant information has led to the suggestion that the process is not necessarily irrational, but merely constrained or "bounded." Most often, the concept of bounded rationality is focused on the limits of human information processing, rather than on limitations on access, or strategic misdirection. But, as Giddens reminds us, some of the more important constraints on human agency are those blind spots we have regarding the motivations and goals of other interested parties who may be involved in some aspect of our decision-making.[63]

He then explains: "There is a tendency to think about rationality in terms of a continuum; one that moves from an idealized intelligence—a difference engine that engages in rapid computation, without errors in calculation, and more critically, without any systematic bias introduced by irrational emotional distractions. On the other end of the continuum we find the sometimes slow, sometimes fast, error prone, easily distracted, and routinely distorted information processing by humans."[64] Gandy's continuum seems to place machines at one extreme and humans at the other.

Ed Stein explains that there are many different senses or conceptions of rationality,[65] and Keith Stanovich describes the strong sense of rationality conventionally used in cognitive science.[66] In the strong sense, rationality corresponds to "optimal judgment and decision making" according to a particular normative baseline, and irrationality corresponds to deviations from the same baseline, which can differ by degree. One widely used normative baseline for judgment and

[62] I consider human autonomy, free will, predictability, and programmability in another paper that is part of this project. But the discussion of nudging in Part II also raises these issues.
[63] Gandy Jr., Oscar H. (2010). *Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems*, Ethics and Information Technology, March 2010, Volume 12, Issue 1, pp 29-42.
[64] Id.
[65] Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Oxford, England: Oxford University Press.
[66] Stanovich, K. E. (2013). *Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice.* Thinking & Reasoning, 19, 1-26.

decision-making is instrumental, expected utility maximization. "The simplest definition of instrumental rationality is as follows: behaving in the world so that you get exactly what you most want, given the resources (physical and mental) available to you. Somewhat more technically, we could characterize instrumental rationality as the optimization of the individual's goal fulfillment [which can be further refined to expected utility]."[67]

For our purposes, *rationality* refers to the strong sense captured by the instrumental rationality definition, and *irrationality* refers to deviations from the specified baseline, regardless of the cause(s) for such deviations. Let me make two things clear before proceeding: First, the normative baseline chose serves the purpose of establishing a baseline and is not meant as a complete normative evaluation; thus, irrationality may quite attractive normatively for reasons not fully captured in or reflected by instrumental, expected utility maximization. Second, the cause(s) for deviation from rationality may matter for evidentiary or normative reasons in particular contexts.

Suppose we set up the Turing test in the conventional manner (**a**) except that the machine participants are simple machines (**f**). The human participants are told that they are participating in the conventional Turing test where they are supposed to act normal and answer questions posed by the observer in a natural fashion. In other words, they are instructed to not act strategically or deceptively. They may even be told (falsely) that the object of the game is to see if the machines are intelligent enough to deceive the observer into concluding that the machine is a human. This is false in two respects. First, the object is to see if the human participants are mistaken to be machines, and second, the machines are simple machines that are programmed to do as instructed and/or answer the questions posed truthfully, accurately, and in accordance with instrumental rationality. Let us also make the observer a machine (**g**), such that the observer is programmed to run a series of conventional rationality tests and experiments.[68] For example, the observer may pose a series of *choice problems*, such as those posed by Allais and developed further in the extensive rational choice and behavioral economics literature.[69] The literature demonstrates a variety of ways in which humans make predictably irrational choices.[70]

One important reason that humans can be seen[71] to act irrationally,[72] according to such

[67] Id. at 345.

[68] Though the observer generally could distinguish humans and machines on the basis of plain computational errors, for example in performing mathematics (*see supra*), we would prefer to rule out such instances. Similarly, we would prefer to rule out instances of nonsensical responses by humans that might be described as completely irrational or just insane.

[69] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'e´cole americaine. Econometrica, 21, 503–546.

[70] See, e.g., Ariely, Dan (2008). *Predictably Irrational: The Hidden Forces that Shape Our Decisions*. HarperCollins, 1st edition. "A substantial research literature—one comprising literally hundreds of empirical studies conducted over several decades—has firmly established that people's responses sometimes deviate from the performance considered normative on many reasoning tasks. For example, people assess probabilities incorrectly, they test hypotheses inefficiently, they violate the axioms of utility theory, they do not properly calibrate degrees of belief, their choices are affected by irrelevant context, they ignore the alternative hypothesis when evaluating data, and they display numerous other information processing biases." Stanovich, supra (collecting studies).

[71] I use the phrase "can be seen" to refer to a plausible observation by the observer conducting the tests and experiments.

[72] It might be better (less controversial) to describe the human responses in terms of bounded rationality rather than

experiments, is that humans contextualize the problems. As Stanovich puts it, "humans recognise subtle contextual factors in decision problems that complicate their choices."[73] Simple machines don't.

Another way that humans can be seen as irrational is through a series of biases in human decision making that lead to distortions from the rational choice (utility maximization) model. Various biases lead people to make judgment errors in the sense that the judgments differ systematically from unbiased forecasts. Some biases may involve discrimination about groups, for example based on race. Some biases involve heuristics and decision making errors; for example, optimism bias, self-serving bias, hindsight bias, among others.[74] Some biases may be the result of contextualization.[75] Again, simple machines don't suffer from such biases.

Still, another way that humans can be seen as irrational is through actions that serve symbolic utility but not instrumental utility; this is related to the notion of ethical preferences, which may lead humans to make choices that diverge from instrumental utility and rational choice theory. Humans also appear to be "strong evaluators" in the sense that they (can) have preferences about preferences (or different levels of preferences), and this can lead to destabilizing conflicts among preferences that cause humans to act irrationally when measured against the rational choice model.[76] Again, not so for simple machines.

All of this is to suggest that there are a variety of ways in which the observer might test agents and distinguish humans and machines. Armed with a battery of tests, the observer would be able to accurately distinguish humans and machines: the simple machines would always respond to the queries in a rational manner (e.g., make predictably rational choices) while the humans would tend to exhibit irrationality (e.g., make predictably irrational choices), at least over the course of a sufficient number of tests or experiments.

Finally, suppose we employ machine learning such that the observer program gradually optimizes its battery of tests and experiments. If, for example, the reliability of certain categories of tests is questionable, then the observer presumably would steer away from using those tests and rely on other more reliable ones. Thus, in the end, suppose we've constructed the *perfect*

---

irrationality. Some will argue that many responses that deviate from the rational choice model or instrumental rationality are in fact rational; for example, some would argue that it is rational to contextualize problems as humans do but machines don't. It reminds me of Tolstoy's ruminations on reason and faith: "Either that which I called reason was not so rational as I supposed, or that which seemed to me irrational was not so irrational as I supposed" Tolstoy, Leo (1879), *A Confession and Other Religious Writings*. Again, though it may seem like I am cheating, I wish to avoid these types of debates, at least for purposes of this discussion. I recognize that in future work, it might be necessary to specify in more detail the rational choice experiments that our observer will employ.

[73] Stanovich, supra; Stewart, N. (2009). *Decision by sampling: The role of the decision environment in risky choice*. Quarterly Journal of Experimental Psychology, 62, 1041–1062.

[74] Jolls, Christine & Cass R. Sunstein (2006), Debiasing through Law, Journal of Legal Studies, vol. 35, pp. 199-241; Christine Jolls, Behavioral Law and Economics (update of the essay published under the same title in Behavioral Economics and Its Applications, Peter Diamond and Hannu Vartiainen eds., Princeton University Press, 2007) [http://www.law.yale.edu/documents/pdf/Faculty/Jolls_BehavioralLawandEconomics.pdf]

[75] Stanovich, supra.

[76] Id. See also Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.) (pp. 343-365), The Oxford handbook of thinking and reasoning. New York: Oxford University Press.

rationality detector that can accurately distinguish humans and machines based on humans' propensity to act irrationally (or in a boundedly rational manner, if you prefer).

Now suppose we run our test with the perfect rationality detector as our observer. *What could we infer from a mistaken identification of a human as a machine? What would that mean? Are there any meaningful inferences to be drawn?* Passing the test provides us with evidence and suggests that there is *something* remarkable about the relevant agent. *But what?*

We have set up our perfect rationality detector in a way that makes it hard to fathom such a mistake, but humans' propensity to act irrationally is not constant or fixed; it varies with context. Thus, in certain (constructed) environments, we reasonably can expect humans to behave in perfect accordance with the rational choice model.

Psychologists have observed how environmental constraints shape (ir)rationality. "Many authors have commented on how the behaviour of entities in very constrained situations (firms in competitive markets, people living in subsistence-agriculture situations, animals in predator-filled environments) are the entities whose behaviours fit the rational choice model the best."[77] Of course, environmental constraints, or the lack thereof, also may push in the opposite direction. Consider, for example, this passage from Stanovich (2013):

> Most humans now do not operate in such harsh selective environments of constraint (outside many work environments that deliberately create constraints, such as markets). They use that freedom to pursue symbolic utility, thereby creating complex, context-dependent preferences that are more likely to violate the strictures of coherence that define instrumental rationality. But those violations do not make them inferior to the instrumentally rational pigeon. Degrees of rationality among entities pursuing goals of differing complexity are not comparable. One simply cannot count up the number of violations and declare the entity with fewer violations the more rational. The degree of instrumental rationality achieved must be contextualised according to the complexity of the goals pursued.

Keep in mind that our objective is not to *judge* degrees of rationality in terms of inferiority or superiority. It may even be the case that pigeons are closer to the Turing line than humans. But that also in not relevant to our present objective.

We are interested in identifying, examining, and evaluating humans and/in environments that construct humans to be indistinguishable from machines. Thus, it would appear that we need to consider how the "harsh selective environments of constraint" shape human behavior. *Is such shaping transitory or lasting? Are the environments themselves transitory or lasting?* In other words, are humans only affected while in particular environment or might the effects be long-lasting? For purposes of this Turing-type test, we could construct the environment (the rooms and rules) to be more or less constraining and see what impacts followed.[78]

Instead, consider the following thought experiment. Suppose that in an effort to improve human decision making and welfare, government decides to "nudge" people toward rational decision

[77] Stanovich 2013, citing various sources.
[78] I examine this idea in other work focused on free will and autonomy as well as an environment game.

making.[79] For our purposes, the nudge is simply a gentle adjustment in the environment or context that corrects for (debiases) predictably irrational decision errors. As Rebonato put it, in his "reasonably precise, definition of libertarian paternalism," assume government implements:

> [T]he set of interventions aimed at overcoming the unavoidable cognitive biases and decisional inadequacies of an individual by exploiting them in such a way as to influence her decisions (in an easily reversible manner) towards choices that she herself would make if she had at her disposal unlimited time and information, and the analytic abilities of a rational decision-maker (more precisely, of Homo Economicus).[80]

Thus, government adjusts the choice architecture by creating a sufficiently constraining--though not completely constraining--environment.[81] People still make decisions and have choices within the environment, but they will (tend to) do so in conformance with rational choice theory. Suppose humans within the government-constructed nudging environment would routinely pass our Turing-type test, administered by our special observer (the perfect rationality detector).

To put it more concretely, suppose specifically that workplace regulation transforms the workplace[82]--for example, an office or factory--in the fashion just described, such that the workplace constitutes a sufficiently constraining environment that humans in that workplace routinely pass our Turing-type test.[83] *What could be inferred? What would this mean? What would be the significance, if any?*

These are reasonable questions to ask. From the perspective of the government, and even from a social welfare perspective focused on maximizing social welfare, passing our Turing-type test would be a measure of *success*. Humans in the constructed environment would behave exactly as intended, efficiently in accordance with the rational choice model--*optimally*. From the perspective of the employers and the workers (humans), there is little to complain about.

---

[79] *See* Sunstein and Thaler (2008); Ariely (2008); Amir & Lobel (2008).

[80] Riccardo Rebonato, *A Critical Assessment of Libertarian Paternalism*, J. OF CONSUMER POL'Y 4 (Aug. 18, 2014).

[81] For purposes of the thought experiment only, the means chosen by government are not specified and thus are unimportant. That is, whether the government constructs the nudging environment through the legal system or through technological architecture is not germane to the thought experiment. Of course, in reality, the means do matter. Sunstein and Thaler are careful in prescribing gentle means that seem empowering (rather than constraining) in the sense that they generally enable people to make more informed, unbiased decisions. *See* Sunstein and Thaler (2008).

[82] There is a rich history and debate surrounding the example of workplaces in which automation and management practices dehumanize workers and treat them like machines. See, e.g., Noble, David F., *Forces of Production: A Social History of Industrial Automation* (1984); Simon Head, *The New Ruthless Economy: Work and Power in the Digital Age* (2005). In the late 19th century, Frederick Winslow Taylor developed the scientific management theory to improve labor productivity and economic efficiency in the workplace. See Frederick Winslow Taylor, *Shop Management*, New York, NY, USA: American Society of Mechanical Engineers (1903). Taylorism has been widely practiced and criticized. See, e.g., Noble, supra; Head, supra. The call center example at the beginning of the paper is one illustrative example.

[83] We can substitute other contexts for the workplace. Consider the classroom, grocery store, sidewalk, or public park, if you prefer.

Efficient workplace performance presumably translates into higher productivity and safety,[84] and no one is forcing them to work in this setting. We're not talking about forced labor. It seems hard to imagine someone making a strong claim to a right or even desire to act irrationally.[85] *Still, is there any reason to think that the constructed environment (workplace) is* de*humanizing? Does the change induced by the constructed environment constitute a reduction or addition in human capability? Is it diminishing or empowering?*[86]

Now suppose that humans who spent time in the environment constructed to nudge (e.g., the workplace) also routinely passed our Turing-type test when no longer within that environment. That is, suppose that after leaving the nudging environment, the humans remained indistinguishable from machines according to our perfect rationality detector. *What could be inferred? What would this mean? What would be the significance, if any? Would the constructed environment (workplace) be more or less dehumanizing?*

These are also reasonable questions to ask. The basic distinction is between a constructed environment within which humans *are* indistinguishable from machines and one in which humans *become* indistinguishable from machines. We might label the latter a *constructive environment* because of its lasting effects. While the lasting nudges might seem creepy, we should not jump to negative conclusions. After all, helping humans to behave more rationally -- or to be able to make more rational decisions --throughout their lives may be in their own self-interest as well as (certain conceptions of) the broader societal or public interest. Moral or normative evaluation will be difficult and contested.[87]

Suppose the nudging government did not limit itself to workplace environments. Suppose the government systematically constructed nudging environments in as many places and social contexts as possible. If this seems implausible and too abstract, consider government surveillance systems, which are expanding in scope and reach across technological platforms and

[84] This presumption is made for purposes of argument only and is by no means an empirical or theoretical claim I wish to defend. There are certainly reasons to doubt the presumption in various real-world workplaces where, for example, human creativity and emotional impact productivity.

[85] Naturally, one might conclude: if we are not contemplating either a violation of rights or a departure from existing preferences, then there is nothing to worry about.

[86] In an essay, *Buddhist Economics*, E. F. Schumacher contrasted modern economics view of human labor with that of Buddhist economics. The former views human labor as a costly input to be minimized.

> The Buddhist point of view takes the function of work to be at least threefold: to give a man a chance to utilize and develop his faculties; to enable him to overcome his ego-centeredness by joining with other people in a common task; and to bring forth the goods and services needed for a becoming existence. Again, the consequences that flow from this view are endless. To organize work in such a manner that it becomes meaningless, boring, stultifying, or nerve-racking for the worker would be little short of criminal; it would indicate a greater concern with goods than with people, an evil lack of compassion and a soul-destroying degree of attachment to the most primitive side of this worldly existence. ...
>
> From the Buddhist point of view, there are therefore two types of mechanization which must be clearly distinguished: one that enhances a man's skill and power and one that turns the work of man over to a mechanical slave, leaving man in a position of having to serve the slave.

Schumacher, E. F. (1973). *Buddhist Economics*, in Small Is Beautiful: Economics as if People Mattered. (Harper Perennial; Reprint edition (October 19, 2010).

[87] I engage normative evaluation directly in the book project.

the public and private spaces and environments within which we live our lives. These systems may not always be explicitly billed or justified as being part of the nudging program, but that does not make them less so. To be fair, they are not part of the specific program advocated by Thaler, Sunstein, and other behavioral law and economics scholars who underwrite their prescriptions with the ethic of libertarian paternalism. Yet, as I explore in the next Part, it is not clear that this ethic sufficiently insulates their project from the concerns raised here.

Nudging is now government agenda, pursued by governments around the world and not limited to any particular setting or technology (i.e., surveillance systems are just an example). It is also the market agenda. Private entities, such as firms and collections of people employing shared networked technologies, are also voluntarily constructing nudging environments. This is not entirely new. Marketing and advertising has always been about shaping beliefs and preferences and nudging people toward products and services. The pervasive, networked, data-driven economy that dominates much of modern life expands the scale and scope, removing some of the barriers between media that had allowed us to be "off" or outside the constructed nudging environments. One of the most fundamental societal questions of the twenty-first century will be about whether and how to preserve our practical freedom to be off (or conversely, whether the environment we build means we are and will remain always on).[88]

Does it really matter *who* is doing the nudging? Maybe it matters when we evaluate a particular example and are evaluating countervailing pressures—checks and balances, if you will. But in the aggregate, I don't think it matters. The political economy and legal distinctions between public and private institutions don't bear as much weight when one begins to look at the macro-picture. *How should we evaluate the agenda from a macro, longer term, and societal perspective?*

II. Judging Nudging

Over the past decade, cognitive psychologists, behavioral economists, and legal academics have had remarkable success influencing policy makers, regulators, and businesses around the globe.[89] The behavioral law and economics agenda, among other things, aims to design and implement "nudges" to improve human decision making in contexts where humans tend to act irrationally or contrary to their own welfare.[90] A "softer" definition of nudges is "low-cost, choice-preserving, behaviorally informed approaches to regulatory problems, including disclosure requirements, default rules, and simplification."[91] I intend to cast the agenda in a different light and provide a different way to discuss the prescriptions. Incremental changes through nudging may very well

[88] If this is too abstract, consider how easy/difficult it is for you to be free, off, untethered. I often say, and hear others say, I wish I could tune out, leave my phone off, not check the text or email, and so on. Some people are better at being off than others. I resisted getting a smartphone for a decade, but recently caved. I often reminisce about being unreachable while commuting or taking a walk. I now cherish my weekly soccer game even more because when I'm playing, I'm untethered.

[89] For a summary, see Cass R. Sunstein, *Nudges.gov: Behavioral Economics and Regulation*, Forthcoming, Oxford Handbook of Behavioral Economics and the Law (Eyal Zamir and Doron Teichman eds.) [Very preliminary draft 2/16/13, available at http://ssrn.com/abstract=2220022.

[90] See, e.g., Jolls (2007); Sunstein and Thaler (2008); Ariely (2008); Amir & Lobel (2008).

[91] Sunstein, Nudges.gov, supra.

make a lot of sense when evaluated in isolation, but that does not mean that we should not also examine and investigate the path set by the agenda and where it may take us.[92] Jolls and Sunstein discuss how debiasing through law can raise substantial autonomy concerns, and they explain how the "nature and force" of the concerns "depend on the setting and the particular [debiasing] strategy involved."[93] I agree and believe the human-focused Turing tests developed in this article and elsewhere provide a means for examining these concerns, and others not reducible to autonomy, across different contexts.

The nudging project is immense, interdisciplinary, and growing rapidly. There is a tremendous literature. Here, to cabin our analysis, I focus on a recent article by Cass Sunstein in which he discusses a particular controversy regarding paternalism within the behavioral law and economics literature about nudging. This provides a current, focused and useful view and avoids engaging the entire project, which would overcomplicate things.[94] My objective is primarily to reveal a different macro-level perspective on the project and suggest how the reverse Turing test methodology might be useful in judging some nudging and illuminating ways of using defaults to prompt deliberative choice and social learning.

A. Choosing not to Choose

In *Choosing not to Choose*,[95] Cass Sunstein examines when private or public institutions might require people to actively choose rather than allow people to choose not to choose (or allow others to make choices on their behalf). These institutional decision makers are acting as choice architects by designing environments within which people live their lives and make particular decisions. In a sense, the institutional decision makers are constructing the functional equivalent of the machine-environments discussed in Part I; it is a form of techno-social engineering.

Sunstein observes that choosing can often be an immense burden and people generally know best whether to choose for themselves or not. He values individual autonomy highly. Accordingly, he emphasizes how restricting a person's opportunity to choose not to choose is itself a form of paternalism and thus requiring active choosing is subject to the same criticisms as other nudges.

---

[92] This is a slippery slope argument, and one that requires attention. See Riccardo Rebonato, *A Critical Assessment of Libertarian Paternalism*, J. OF CONSUMER POL'Y 22-23 (Aug. 18, 2014) (worrying about the slippery slope that leads to subliminal advertising).

[93] Jolls, Christine & Cass R. Sunstein (2006), *Debiasing through Law*, Journal of Legal Studies, vol. 35, pp. 199-241.

[94] Previously, I had used Sunstein and Thaler's book, Nudge, as the focal point for this article, but it was more than necessary and somewhat dated. There is a slew of new books and articles. For a decent summary, see, for example, Riccardo Rebonato, *A Critical Assessment of Libertarian Paternalism*, J. OF CONSUMER POL'Y Part 2 (Aug. 18, 2014).

[95] The article is being published soon in the Duke Law Journal. A draft is available online: http://ssrn.com/abstract=2377364. I also have a more updated version on file. According to Professor Sunstein, it is the subject of a book, forthcoming in 2015. Of course, I will make adjustments to the arguments presented in this section as he revises his work in anticipation of book publication.

This is an important, though perhaps counterintuitive, defense of nudging. Thaler and Sunstein consistently maintain that nudging should be underwritten by an ethic of "libertarian paternalism," one component of which is that agents should find it easy to opt-out of a nudge's trajectory; people always should be able to choose for themselves and even behave irrationally, if that is what they really want to do. In their view, this ethic preserves individual autonomy. Thus, in accord with the ethic of libertarian paternalism, Sunstein settles on a framework for evaluating when architects should require active choosing that "depends largely on the costs of decisions and the costs of errors." He concludes that: "Where people are relevantly heterogeneous, and where choice architects lack information or neutrality, active choosing has real advantages. But if a default rule is accurate, active choosing does not make a great deal of sense, at least when people remain free to go their own way if they see fit. When choice architects overlook this point, and nonetheless insist on active choosing, they might well be behaving paternalistically, and in a way that reduces both the welfare and the autonomy of those whom they are seeking to help."[96]

Sunstein recognizes one strong argument in favor of active choosing, *social learning*, which is that people may learn and/or form their own preferences while engaged in active choosing. Free will depends on a freedom to determine one's beliefs and preferences, and particularly, higher-order beliefs and preferences (e.g., second-order preferences about preferences).[97] He discusses Mill's classic discussion of social learning and the limits of unreflective habits,[98] and as we know from his prior work, Sunstein is aware of and seriously concerned with the phenomena of self-narrowing filter bubbles and the implications for humanity.[99]

Sunstein recognizes the importance of the social learning argument, yet he ultimately seems to undervalue it. I believe he does so for two reasons. First, he significantly overvalues a counterargument, which he refers to as "a formidable objection to the learning-based argument

[96] Sunstein says that when choosing not to choose is an alienation of liberty it cannot be justified. (draft at pp. 22-23 fn.80) However, he does not provide metrics for determining which choices not to choose are instances of alienation and which are merely justifiable offloads of cognitive bandwidth. As discussed below, the human-focused, behavior-based Turing tests might be useful in drawing some of these distinctions, although liberty might not be the most relevant human capability to safeguard. Sunstein also gives a brief treatment about cases where delegation is not permissible. For example, you cannot delegate the power to vote and you cannot choose to be a slave. In the voting scenario, people are required to take responsibility for voting and therefore cannot delegate and diminish their responsibility for the decision; this would potentially undermine voting and democracy.

[97] See Frankfurt, H., *Freedom of the Will and the Concept of the Person*, Journal of Philosophy 68 (1): 5–20 (1971). There are other views on free will, including that free will is an illusion. See Shaun Nichols, *Is Free Will an Illusion? Don't trust your instincts about free will or consciousness, experimental philosophers say*, Scientific American, Mind & Brain (Oct 20, 2011). For various definitions and an overview of philosophical approaches, see *Consciousness*, Stanford Encyclopedia of Philosophy at http://plato.stanford.edu/entries/consciousness/. I explore this in more depth in the book project.

[98] *Choosing*, draft at 23 (citing John Stuart Mill, On Liberty (Kathy Casey ed., 2002) (1859)).

[99] See, e.g., Cass R. Sunstein, REPUBLIC.COM (2001).

for active choosing."[100] Second, he fails to fully engage with what social learning could entail, although he begins the critically important inquiry. Let me unpack each reason and explain why they lead to undervaluation of social learning by active choosing.

Sunstein credits heavily the counterargument that one dimension of choosing that people must learn to manage is the initial one, that is, whether to choose actively or to choose not to choose. Specifically, he says:

> [T]here is a formidable objection to the learning-based argument for active choosing. The objection is that people do and should learn about whether to choose actively or instead to choose not to choose. People sometimes decide correctly, and sometimes they err, in making that particular choice, as in making all other choices. It is important for people to learn, over time, about when they should be choosing and when they should be relying on a default rule (and accepting the force of inertia or the power of suggestion). That form of second-order learning is exceedingly important. The problem is that those who insist on active choosing, or even favor it, will reduce or prevent learning along this important dimension. Claiming to promote learning and the development of values and preferences, they truncate such learning and such development about an extremely important set of questions.

It is true that people must learn to choose along this dimension, but Sunstein does not quite identify how people do so and consequently underestimates the degree to which people develop this capability throughout their everyday lives. As a result, he gives the objection much too much weight.

People learn to manage this particular dimension of choosing throughout their lives in an incredible number of *everyday* situations. We face it continuously in our lives as we face countless decision points, ranging from the incredibly mundane to the complex and weighty.[101] Critically, we do not always do it alone. The process of learning is, as both Sunstein and Mill suggest, social. We often develop a common sense understanding of how to manage this dimension of choosing. But, as discussed in Part I (*common sense test*), we cannot and should not assume stability or a persistent reservoir of common sense available to us. Common sense is developed and learned, *socially*. If the everyday situations and corresponding learning opportunities were to diminish significantly, for example because ubiquitous sensor and big-data-driven automation renders everyday commonplace choosing less frequent, then we might need to create (or guard) our learning opportunities. The common sense test provides a means for identifying and evaluating this possibility.

[100] *Choosing*, draft at 23.

[101] There are countless examples. Cf. draft at 14 (recognizing that people confront these choices "in countless matters" "in daily life"). I cannot help but think of my children deciding whether to choose what to eat for breakfast or pack in their lunches for themselves or to wait for me to choose for them.

Sunstein suggests that the "formidable objection" requires a refinement of the social learning argument, and he proceeds to discuss briefly how social learning in some contexts and about some subjects can lead to the "accumulat[ion] [of] some kind of capital." He explains:

> as, for example, by learning about what they actually like (in terms of, say, politics, art, or music) or by developing an understanding of certain matters that very much affect how their lives will unfold over time (in terms of, say, health insurance or investments). In some such cases, the argument for active choosing may be convincing -- perhaps because people are subject to inertia or a form of myopia that leads them to favor a default. Nonetheless, it must be acknowledged that second-order learning might therefore be compromised.

He is heading in the right direction. He notes the possibility of forming and learning about one's preferences as well as knowledge about subjects that matter in life. These are important. Still, he stops short of fully examining what could be developed and learned and what skills and capabilities can be developed and practiced when people engage in active choosing. It is hard to evaluate, given the brevity of his discussion, but it leads him to undervalue social learning through active choosing. For the reasons explained in the previous paragraph, I also worry that he overestimates the degree of compromise between first- and second-order learning. In the end, I acknowledge we may be on the same page and simply need to do more work in contextualizing the arguments and developing an empirical understanding of social learning. [102]

Consider the following institutional default rule. When public and private institutions[103] through their choice architects have opportunities to set defaults for people that (1) require active choosing or instead (2) require someone to choose to be an active chooser (in which case the default is that they are not choosing), the institutions should, as a default for their own choice, choose the former (1).[104] This institutional default promotes learning, individual formation of preferences, and skill development and reduces the risk of manipulation. Of course, we can consider when deviations from this institutional default rule would be justified, and Sunstein's framework provides some assistance for doing so, but it is incomplete.

---

[102] Riccardo Rebonato also emphasizes the importance of social learning: "[I]ndividuals can learn (or be better enabled) to become better thinker," but "[i]f our critical senses are not exercised and engaged, (and the attending neural pathways are re-routed away) what can we, as individuals and citizens, expect and hope for? For more and more nudges engineered by a benevolent libertarian paternalistic philosopher king to do on our behalf more and more of our deciding? This is not a prospect that I would welcome." Rebonato at 29.

[103] I would categorically reject the conventional market discipline defense for private institutions; it does not work nearly as well or as often as theory predicts. Thus, I would retain the institutional default rule for both private and public institutions.

[104] Depending on the context, the active choosing default might allow individuals the opportunity to opt-out of active choosing, and exercising this option might entail varying levels of information exchange and process, depending on the nature of the decision and the type of capital and human capability at stake.

The key to identifying justifiable deviations from the active choosing default depends on an analysis of trust and expertise, the specific decision costs and potential error costs, and the functional human capabilities at stake within specific contexts. Sunstein's framework touches on some of these issues, particularly the role of expertise, decision costs and error costs. He only lightly touches on the capabilities, however.

We need to examine what human capabilities can be learned or developed in an underdetermined environment within which people must actively choose—or as he calls them, environments architected to preserve serendipity.[105] (Keep in mind that these are the machine-environments discussed in Part I, the workplace, the call center, the rules and rooms of the conventional Turing test, the constructed, constructive environments within which humans are situated.) If there are none or the stakes are low, perhaps because there are many other opportunities to develop and learn, or there is little risk of manipulation, then an exception might be justified and option (2) might be preferable.[106]

Identifying the functional human capabilities at stake can be difficult. As explained in Part I, we do not have very good tools for identifying, measuring, or evaluating human capabilities and their expansion or diminution within different social-technological environments. The reverse Turing tests developed in Part I may be an appropriate tool for these tasks, but there remains much work to be done in developing and refining the tests. In a passage that very closely relates to the common sense discussion above, Sunstein considers the possibility of diminishing skills and practical agency in the context of GPS:

> Libertarian paternalists often refer to the GPS as a prime nudge, because it helps people to find the right route while also allowing them to go their own way. But there is a downside, which is that use of the GPS can make it harder for people to know how to navigate the roads. Indeed, London taxi drivers, not relying on the GPS, have been found to experience an alteration of their brain functions as they learn more about navigation, with actual changes in physical regions of the brain. As the GPS becomes widespread, that kind of alteration will not occur, thus ensuring that people cannot navigate on their own. This is an unusually dramatic finding, to be sure, but it raises the possibility that when people rely on defaults or on other nudges, rather than on their own active choices, some important capacities will fail to develop or may atrophy. This is the anti-developmental consequence of some helpful nudges, including the GPS itself.[107]

[105] Sunstein, draft at 32 (citing JANE JACOBS, THE DEATH AND LIFE OF GREAT AMERICAN CITIES (1961)). See also Frischmann, *Infrastructure*, ch. 5 (on the social option value of underdetermined environments and infrastructure commons); Benkler, U. Chi. L. Rev. (2013) (same).

[106] Intermediate options could be explored as well. For example, in some contexts, we might insist on a particular process of deliberation, perhaps slowing down (sludging?) the decision making so that choosers can determine whether there is justifiable trust in and reliance on another's expertise.

[107] Sunstein, *Choosing not to choose*, ssrn draft at p.21 (citing Eleanor A. Maguire et al., *Navigation-Related*

Some people might favor active choosing precisely to avoid such consequences. As Sunstein suggests, "They might want to develop their own faculties."[108] They might, but that would depend entirely upon their existing preferences. Sunstein places significant trust in such preferences, as his reliance on surveys reveals.

But existing preferences on such matters are hardly a reliable guide. To the contrary, many external—e.g., environmental, cultural, and technological—factors shape such preferences dynamically over a lifetime, and many nudges implemented by public and private actors on a regular basis (e.g., in advertising) powerfully discourage active choosing to facilitate learning and encourage choosing not to choose or simply impulsive behavior. Even the default rules themselves can shape preferences.[109] Like GPS, many technologies provide helpful nudges and have substantial anti-developmental consequences.[110] Regardless of what libertarian paternalists assert, GPS is by no means the *prime* nudge; it is just a useful example, one of many technologies that function similarly.[111]

The GPS example provides support for an institutional default in favor of active choosing. Sunstein acknowledges that active choosing is defensible as a means for minimizing the risk that default rules chosen by private and public institutions infantilize people. Yet Sunstein seems to waver because it would interfere with autonomy, which brings me to my final critique.

Sunstein frames his analysis in terms of autonomy.[112] He is concerned with contesting claims about paternalism and the criticisms of nudges implemented by private or public institutions through their choice architects. He shows that requiring active choosing is also a form of paternalism because it diminishes the freedom to choose not to choose. This line of argument is fine as far as it goes. If we prize autonomy most or seek to minimize institutional interventions that reduce autonomy, then perhaps the institutional default I defended above would appear to be unsustainable because we would need to leave people the option of deciding not to choose actively. But, for the reasons I explained above, I think this misses the forest for the trees (even in pure autonomy terms) by focusing on a single decision point and failing to appreciate how

---

Structural Changes in the Hippocampi of Taxi Drivers, 97 PROC. NAT'L ACAD. SCI. 4398 (2000) and Riccardo Rebonato, *A Critical Assessment of Libertarian Paternalism* (2012)).

[108] Id. at 21.

[109] For example, default rules may affect which pieces of information, evidence, or content percolate to an individual.

[110] See Common Sense test discussion, supra.

[111] I discuss a wearable technology example below.

[112] He discusses welfare based arguments, but they appear to be much less persuasive if one takes the social learning arguments seriously. On one hand, as Sunstein acknowledges, the human, knowledge, social, or other capital developed during social learning may have substantial but difficult to measure value. There are substantial spillovers. See Frischamnn, *Infrastructure* (2012); Frischmann & Lemley, *Spillovers* (2007). On the other hands, one might turn to Amartya Sen's work on the capabilities approach as an alternative way to value the human capabilities gained through social learning. *See* Frischmann, *Capabilities, Spillovers, and Intellectual Progress: Toward a human flourishing theory for intellectual property*, draft under submission (2014).

active choosing provides the critical opportunity to develop beliefs, preferences, and even skills that enable people to exercise autonomy in many other contexts with respect to many other decision points throughout their lives. The environments we construct and the default rules institutionalized within them shape beliefs and preferences dynamically and consequently whether and how people choose to exercise their autonomy. In other words, if we are optimizing or maximizing autonomy, the static autonomy gains of preserving the freedom to choose not to choose in specific institutional contexts may be less than the dynamic autonomy gains from active choosing.[113]

Moreover, depending on the context, active choosing might provide the critical opportunity to develop the beliefs, preferences, and skills that enable people to exercise *other* human capabilities, ranging from the development and sharing of common sense to empathy for and conscientiousness toward others.[114] More work needs to be done in exploring the relationships between the socio-technical environments we construct and various human capabilities. For now, let me emphasize a simple point:

There is more to being human than autonomy.[115] Put another way, environments architected to nudge can be dehumanizing even if autonomy is preserved. Suppose, by autonomy, one means freedom to determine one's second order beliefs and desires: even if we preserve that type of autonomy completely, we can still end up as slaves in a severely constrained environment without the freedom to act on our desires. Suppose, instead, that by autonomy, one means the practical and situated freedom to act on one's desires or will: even if we preserve that type of autonomy completely, we can still end up as automatons with others determining our beliefs and preferences.[116] Over-determined or over-architected environments that preserve either type of

[113] Two caveats. First, recall my claim that people have many opportunities throughout their daily lives to learn about choosing not to choose; it is common sense. If this claim is correct, then the static autonomy gain might be slight. However, if the claim is incorrect, then the static autonomy gain might be substantial. Second, some absolutists may reject any (static) autonomy loss. The problem with this view is that it may lead to less autonomy over time.

[114] Sunstein discusses third party effect (externalities) and recognizes their importance in justifying various institutional interventions; I agree, see generally Frischmann, *Infrastructure: The Social Value of Shared Resources* (2012); Frischmann & Lemley, *Spillovers* Colum. L. Rev. (2007), and wonder how active choosing relates to the development of relational capabilities. I need to do more work on this possibility, however.

[115] See supra note 106.

[116] In the draft book, I offer the following explanation and attempt to distinguish the two outcomes described in the text:

> [W]e might distinguish (1) over-determined environments that eliminate the practical freedom to exercise free will by constraining the range of actions or opportunities presented to situated agents, and (2) constructive environments that operate more directly on the will by shaping or even determining beliefs, preferences, tastes, or values. To the extent that there is no escaping constructive environments because social shaping is a ubiquitous feature of the modern world, then (1) and (2) may bleed into each other. The mechanisms in (1) and (2) are different, however, and worth distinguishing. In some contexts, the outcome for the situated agents might be the same, and in others, it might be different. Perhaps at the extremes, (1) leads to slaves and (2) leads to machines. Slaves are agents (humans) with free will but without autonomy, meaning practical,

autonomy still can be dehumanizing.

Sunstein evaluates nudging scenarios with a cost-benefit framework focused on decision costs and error costs. Of course, incrementally, or case by case, this type of cost-benefit analysis can be useful. The problem with this approach is in the aggregate. Without a better accounting of the social value of learning and attendant human capabilities and capital, too much is left out of frame because it is too difficult to measure.[117] Moreover, as discussed in the rationality test thought experiment, completely errorless rational decision making would leave us, in the extreme, as little more than automatons, indistinguishable from machines. The macro-level concern is with nudging as a systematic agenda and the path it sets for society in terms of techno-social engineering of humans and society.

To motivate and frame that discussion, consider two true short stories. Neither is quite a story of dehumanization where humans are or become indistinguishable from machines. Rather, each is an example of an incremental step in that direction. The first concerns techno-social engineering of children's preferences. It is a simple nudge. The second concerns techno-social engineering of human emotions, and it is not exactly a conventional nudge, although it is might be described as a close cousin.

B. A Simple Nudge: Using Activity Watches to Shape Elementary School Children

Last year, my first grader came home after school very excited. "Dad, I won. I mean, I've been picked. I get a new watch." "That's great," I said, "What happened?" He quickly rattled off something about being one of the kids in his class who was selected to wear a new watch for gym class.

A day or two later, I received the following letter in the mail from the school district:

> Dear Parents/Guardians,
>
> Your child has been selected to be among the first group of students to participate in an exciting new initiative made possible by our recent $1.5 million PEP Grant.
>
> We have added ACTIVITY WATCHES to the K-12 physical education program so that we can assess how the PEP grant

---

situated freedom to exercise their free will. The slavery environment dramatically reduced the scope of opportunities for humans constrained by the environment to act and author their lives. But slaves retained free will; they thought for themselves and had dreams and desires about their lives. Of course, slave owners tried and in some cases may have succeeded in turning slaves into machines, by determining their beliefs, preferences, tastes and values and thereby depriving them of free will.

[117] See supra note 106. This outside the framing effect happens all too often. The activity watch story that follows is a decent example.

> impacts students' physical activity in [the school district]. We are periodically selecting groups of students at random to wear activity watches on their wrists to track daily activity time.
>
> One of the goals of our program is to see that students get the recommended amount of physical activity each day (60 minutes). As part of a quality physical education program, the use of activity watches can motivate students to challenge themselves to become more physically active.
>
> For the students selected to participate in this first group, we will be distributing activity watches starting January 13th for students to wear before, during, after school and over the weekend until Tuesday, January 21st. We ask that students do not take off the watch once it's on their wrist. They should sleep, even shower with the watch in place. There are no buttons to push or need to touch the watch, as it is pre-programmed to record and store each day of activity time.
>
> At the end of the 9 days, each family will be able to access a report of their child's activity, and you are welcome to consult with your child's physical education teacher about what you learn and ways to further support your child's physical health and fitness. In addition, the group's combined information will be used to provide baseline data on student physical activity in [the school district].
>
> In closing, I invite you to join me and your child's physical education teacher in motivating your family to participate in physical activity together. If you should have any questions about this new technology, please do not hesitate to contact your child's physical education teacher.
>
> Yours in health,
>
> XXXX XXXXXXXX
>
> Supervisor of Health, Physical Education and Nursing Services

What is your reaction? I ask you to think about it for a moment before I tell you my reaction because mine was atypical, at least in my community.

When I read the letter, I went ballistic. You have to understand that I teach and write about information and technology law. So perhaps I am just atypical. I could not help but wonder about various privacy issues—basically, who? what? where? when? how? and why? with regard to collection, sharing, use, and storage of data about kids. The letter did not even vaguely suggest that parents and their children could opt out, much less that their consent was required. Even if it had, it couldn't be informed consent because there were so many questions left unanswered, and there was no mechanism to manifest consent or a lack thereof—no written form to sign, not even an "I Agree" box to check on a website.

I also wondered whether the school district had gone through some form of Institutional Review Board process. Had someone, anyone considered the ethical questions? Whether or not the school district's data collection qualifies as research, and whether or not the regulations that require IRB evaluation apply to this particular context, the underlying ethical issues were more or less the same as those present in situations of human subject research at a university. Maybe worse. The ethical issues might be even more complicated since the environment school children face is arguably more coercive and dependent upon on trust than the university laboratory. I had to go through the IRB process for a research project in the past, and so I couldn't help but wonder.

I read the letter again but got stuck on: "We ask that students do not take off the watch once it's on their wrist. They should sleep, even shower with the watch in place." Seriously, bathtime and bedtime surveillance! I couldn't believe it, and I couldn't believe that the school district could be so clueless.

I read the letter again and it made me think of one of those Nigerian bank scam emails that go straight into my spam folder. Such trickery! I thought. Then I remembered how my son had come home so excited. The smile on his face and joy in his voice were unforgettable. It was worse than an email scam. They had worked him deeply, getting him hooked. He was so incredibly happy to have been selected, to be part of this new fitness program, to be a leader. How could a parent not be equally excited? Most were, but not me.

I emailed some friends from town and asked if their kids had been selected, if they had received the same letter, and if they were going to let their kids participate. A few had. None had thought it was a big deal. All of them had let their children participate. I did not. My son understood after a lengthy explanation, but it wasn't easy for him or me.

There is much more to the particular story. I contacted someone at the PTA, spoke with the Supervisor of Health, wrote a letter to the School District Superintendent, and eventually had some meetings with the General Counsel for the school district. The program is like so many being adopted in school districts across the country: well-intentioned, aimed at a real problem

(obesity; lack of fitness), financed in an age of incredibly limited and still shrinking budgets, and elevated by the promise that accompanies new technologies. Fortunately, I live in a great town, and everyone I spoke to was eager to learn and talk more about the issues that hadn't even occurred to them. What caught their attention most was a line from the letter I sent to the Superintendent: "I have serious concerns about this program and worry that [the school district] hasn't fully considered the implications of implementing a child surveillance program like this." No one previously had called it "child surveillance"; all of a sudden the creepiness of bath time and bedtime surveillance sunk in.

Surveillance is what generated a visceral reaction, and so it was an effective means for getting people to stop and think about the program. But up to that point, no one seemed to have done so for a number of obvious reasons. People trust the school district (the choice architects), and they love technology. The salient problem of obesity weighs heavily on the community, and the activity watches seem to be a less intrusive means for addressing the problem. People obtain information about their activity levels and then are better able to adjust their behavior and improve fitness. They can do so on their own, as a family, or in consultation with the physical education teacher. Plus, it was funded by a federal grant. The activity watch program presents substantial upside with little or no downside, an easy cost-benefit analysis. For most people, it seems like one of those rare win-win scenarios. And many nudges, when viewed incrementally, may seem this way.[118] *But are they?*

In my view, the most pernicious aspect of the program is not the 24-7 data collection, nor is it the lack of informed consent. To be clear, these are real problems; I am not saying otherwise. These concerns don't quite capture an important issue lurking a bit deeper, which is how the school district's use of the activity watches is a form of social engineering practiced on children. Of course, the Department of Education, the school district, and other program supporters understand that they are engaged in social engineering in the sense that they use the activity watches and corresponding surveillance to shape cultural attitudes and individual preferences regarding fitness and activity. That much is more or less transparent. It is a simple nudge because the program aims to generate and provide information about activity levels and fitness, and thereby enable better choices. Students and parents still decide for themselves whether to change their activity levels. They retain their autonomy. But as I argued in the previous section, that is not all that matters.

The deeper concern I have with the program is the unexamined techno-social engineering, specifically, how the program shapes the preferences of a generation of children[119] to accept (without question or concern) a 24-7 wearable surveillance device that collects and reports data

[118] Cf. Rebonato at 21 (critiquing Sunstein and Thaler's defense against slippery slope arguments).
[119] As well as parents, friends, and other community members.

to others.[120] Even though autonomy and choice might be preserved, the more subtle influence on beliefs and preferences works at a different level and will shape a host of future choices. It increases tolerance of surveillance, manipulation, and nudging.

Incremental steps of this sort may seem, in isolation, justifiable, especially on a constrained cost-benefit calculus that only considers immediate and obvious costs and benefits. Such a view can be dangerously myopic because it ignores the cumulative effects of many, interdependent yet still incremental steps. Privacy is but one victim of death by a thousand cuts.

The bottom line is not necessarily to reject the use of activity watches or similar technologies in public schools. Perhaps their use is justifiable as an effective means to combat obesity and encourage fitness. But a school district that chooses this path ought to do so more carefully, with an awareness of and open dialogue about how the technology affects the human beings, the children. As I said to the general counsel for the school district, this is a decent teaching opportunity. Fitness and privacy might be joined as students learn about the technology and their relationships to it and to others, including the school, the Department of Education, the device manufacturer or other entities that obtain access to the data and can use it or sell it.

[120] One might respond that people already carry and wear such devices, e.g., smartphones. But that misses the point in so many ways (e.g., young children vs. adults/teenagers; different sensors and types of surveillance; public school as supplier; etc.).

C. Not (yet) a simple nudge: Facebook Manipulation of Its Users' Emotions

On June 17, 2014, the Proceeding of the National Academies of Science (PNAS) published *Experimental evidence of massive-scale emotional contagion through social networks*.[121] The short article reported on a *remarkable* experiment that demonstrated emotional states can be transferred to others by emotional contagion. Researchers at Facebook and Cornell University conducted the experiment and "manipulated the extent to which people (N = 689,003) were exposed to emotional expressions in their News Feed." Unbeknownst to a few hundred thousand people, Facebook deliberately reduced their exposure to their friends' positive (negative) posts, depending on which conditions Facebook applied. In other words, Facebook deliberately exposed people to the test contagion and then watched to see what happened. It turns out that the results of the experiment showed emotional contagion exists and can be deployed by Facebook. People exposed to more positive (negative) posts tended to post more positive (negative) posts relative to the control groups. Moreover, people "exposed to fewer emotional posts (of either valence) in their News Feed were less expressive overall on the following days," which the authors described as a withdrawal effect. The authors concluded:

> given the massive scale of social networks such as Facebook, even small effects can

[121] Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock (June 17, 2014), Experimental evidence of massive-scale emotional contagion through social networks, vol. 24, Proc. Nat'l Acad. Sci. USA.

> have large aggregated consequences: For example, the well-documented connection between emotions and physical well-being suggests the importance of these findings for public health. Online messages influence our experience of emotions, which may affect a variety of offline behaviors. And after all, an effect size of d = 0.001 at Facebook's scale is not negligible: In early 2013, this would have corresponded to hundreds of thousands of emotion expressions in status updates per day.

Not surprisingly, a firestorm followed publication of the study. Bloggers, media pundits, researchers, Facebook users, and others debated the ethics of the research, mostly focusing on whether the researchers should have obtained informed consent from the research subjects and whether the Institutional Review Board at Cornell should have played a greater role in regulating, supervising, or monitoring the research. These are very important ethical issues.

A few months later, the NY Times reported on some progress: researchers studying us on social networks and other digital media are now grappling with ethics and may develop guidelines to govern how they experiment on us (*Under The Microscope*, 8/13/2014). *Thank goodness!*

But we subjects should grapple with the ethics as well. To get a sense of where you stand, consider a few questions:

1. Is deliberate emotional manipulation by Facebook a problem of *process* (no informed consent for the subjects) or *substance* (emotional manipulation)?

2. If it is a problem of inadequate *process*: Is IRB review a solution? What about informed consent? What does that mean to you? Pretend you're negotiating a one-to-one contract with Facebook. What exactly would you agree to? Would clicking "I agree" when you sign up for the service be enough?

3. If it is a problem of *substance*, can you explain the problem without reliance on adjectives like creepy? Can you articulate what exactly is wrong with emotional manipulation by Facebook?

4. Is it acceptable for Facebook to induce or suppress the emotional contagion of your friends?

5. Suppose Facebook tests, develops, and optimizes its emotional manipulation capability to help people to make better decisions? Would it be acceptable for Facebook to induce or suppress impulsive purchases (or at least, clicks)?

6. Suppose Facebook optimizes its emotional manipulation capability specifically to minimize emotional interference with rational decision making. Would this nudge people to make *better* decisions? Would people nudged in this fashion act like machines? Would they be (or could they be) any less human?

7. Suppose Facebook optimizes its emotional manipulation capability and lets users choose the settings—dial up some happiness! Would you use it?

These are all difficult questions. While the lack of informed consent and role of IRB are

important issues, they are the tip of the iceberg. Of course, the tip is all that gets attention until too late. The deeper issues (reflected in question 3-7) are substantive, have less to do with the research process or this particular experiment, and more to do with the technological capacity for social engineering that Facebook is testing. When I read the Facebook study, "massive-scale emotional contagion through social networks" is what caught my attention.

Many things alter our moods every day. Advertisers and politicians (and their various consultants) are expert manipulators, and so are the rest of us. We try to influence each other regularly, for better or worse. We nudge each other. That's a big part of what socializing and communicating entails. Emotional contagion is not the only social contagion, but it can be a powerful nudge.

Many technologies play an integral role in shaping our beliefs, emotions, and well-being, sometimes but not always in ways we know about and at least partially understand. But systematic manipulation by technological-social engineering on/by platforms that constitute the environments we live in daily may be much more challenging to know about and understand, and it may become more pervasive. In particular, such manipulation may be much harder to know about and understand *independent* of the platforms' influence on emotional and other social contagions. We need to engage the ethics, including both process and substance, and we need to develop better tools for identifying and evaluating such manipulation. After all, we only know about Facebook's experiment because it published the results.

D. Plausible Extensions

The two true stories bring to mind fictional ones. Specifically, consider the following extensions of the stories.

1. Suppose Facebook optimizes its emotional manipulation capability and expands beyond its social network interface on the Internet into the Internet of Things and smart homes. Thus, suppose Facebook extends its optimized emotional manipulation capability outside the social network environment it has constructed to the other environments within which we live our lives. Would you consent? Would you be able to? Does your answer depend on whether or not you are in control, whether you could choose the settings?

This reminds me of a conversation I had recently. I asked a colleague how he would feel about being a "mere brain in a vat with his happiness optimized by some technical system."[122] Without hesitation, he responded, "Extremely happy, I guess." For him, the stipulation made it easy; he suggested that intuitions derived from the hypothetical just tend to fight the hypothetical and its stipulation. Many people simply have doubts about it being possible or whether there is something hidden in the stipulation of optimal happiness. If we remove those doubts, however,

[122] On the brain in the vat thought experiment, see http://www.iep.utm.edu/brainvat/ (explaining history and collecting references). On the optimal happiness stipulation, see Robert Nozick, *Anarchy, State and Utopia*. New York: Basic Books, 1974 (experience machine discussion).

and take the hypothetical and stipulation as unassailably true, he was confident in his answer and, as he told me afterwards, he would choose to be in such a state if he could. People tend to have mixed feelings about the thought experiment. The Facebook experiment and the hypothetical extension may highlight steps along a path. We may doubt we'll ever get to the end point, or even very far down the path. But can you be sure? How might humans and society change along the way?

2. Imagine attending a school board meeting where the board presents the next Department of Education grant proposal. The proposal involves an upgrade to the activity watch program, an opportunity to build upon its success and community wide acceptance. Having successfully deployed the activity watches for a few years, the children, parents and teachers have grown accustomed to the technology, and for some, the fitness gains are impressive. The upgrade entails deployment of an additional activity sensor. This sensor monitors brain activity. The collected data enables students, parents, and teachers to evaluate attentiveness and engagement and improve mental fitness. Initially, the upgrade only will be available to selected fourth, fifth and sixth grade students who already have activity watches. Over the course of the next two years, the user base will be extended gradually until all of the students are participating. Would you support the proposal? Is this meaningfully different from the activity watches? Suppose that instead of only measuring brain activity for attentiveness, the sensor mapped brain activity and then tailored instruction and evaluation based on such maps.

We could go on. Each incremental upgrade to the activity watch seems justifiable on its own terms, and it only gets easier. That is, the first step makes the second more palatable, harder to resist or even notice.

Conclusion

If you are familiar with the boiling frog soup story, then you could probably smell it coming.[123] Do you know how to make frog soup? If you begin with a live frog, you cannot just drop it into a pot of boiling water because it will jump out. You need to place the frog in a kettle of room temperature water and increase the temperature of the water slowly enough that the frog doesn't notice it's being cooked. "As the water gradually heats up, the frog will sink into a tranquil stupor, exactly like one of us in a hot bath, and before long, with a smile on its face, it will unresistingly allow itself to be boiled to death."[124] The story often is used as a metaphor to comment on the difficulties we face in dealing with the gradual changes in our environment that

[123] On the boiling frog metaphor, see https://en.wikipedia.org/wiki/Boiling_frog. Frogs do not actually behave as the story suggests. Nonetheless, as Eugene Volokh noted, the metaphor is useful conceptually, regardless of how frogs actually behave. Volokh, Eugene (2003). "The Mechanisms of the Slippery Slope". Harvard Law Review 116 (4): 1026–1137.

[124] Daniel Quinn, The Story of B.

can have drastic, irreversible consequences. The gradual change may be difficult to identify, or each incremental step, or change in temperature, may in the moment seem desirable. The end state may be difficult to anticipate or comprehend, and in the end, it may not seem to matter. After all, it doesn't really matter (to the frog) whether the frog knows at the end that it is frog soup or whether the soup is tasty and nourishing. What matters (to the frog) is the fact that water temperature is rising slowly, how that occurs, who controls the heat, and perhaps even why? Had the person in control of the heat turned it up too quickly, the frog might be alerted to the danger and escape, but those who aim to cook frog soup know better than that.

I admit that the metaphor fails at this point because there is not necessarily a single cook, master planner or designer. Rather, we may be doing this collectively as we participate on Facebook, attach devices to our children, and casually nudge each other down the path we are on. Nonetheless, the architectural technologies are powerful and can significantly concentrate power. We need to ask *who is doing the thinking* as we increasingly use and depend on mind-extending technologies. *Who controls the technology? Who directs the architects?* Important distributive justice issues lurk beneath the surface. This is the subject of another part of this project. Moreover, the boiling frog metaphor paints a dystopian picture; the end-state is death, after all. But there are other possibilities and many incremental steps may be worth taking. Rather than attempt an equilibration to balance the dystopian with utopian imagery, let me emphasize the main point, which is not the dystopian end-state; the main point is that we need much better tools for identifying and evaluating our humanity and evolving relationships with technology and environment. The radically re-purposed reverse Turing tests presented in this article are such a tool.

Nudges are an example of techno-social engineering through manipulation of the choice architecture. Many nudges are excellent policy options, in many cases preferable to alternatives such as regulatory mandates. Yet the logic of nudging supports a wider range of techno-social engineering than entailed by Cass Sunstein's careful definition of nudges. And it is the logic that can be path setting. As suggested in the activity watch story and extensions, we might begin with a simple nudge that appears easily defensible from a cost-benefit perspective, and that first nudge makes the next more easily defensible as beliefs and preferences about systematic surveillance are shaped. Moreover, the logic of nudging is easily exported to other contexts where techno-social engineering borrows the logic without the constraints baked into Sunstein's definition. For example, when emotional manipulation by the technological platforms that determine the environments within which we live substantial portions of our lives also determines our choice architecture, their explanation will rely on the logic of nudging.

And so, I believe, we must question the logic, engage the ethics, and develop the tools to better understand our humanity.