2 Transparency

There's a famous story about a German horse who could do arithmetic.[1] At least, that's how it seemed at first. His name was Hans—"Clever Hans" they called him—and his legend holds important lessons for us today. Hans was reputedly able to add, subtract, divide, multiply, and even read, spell, and tell the time.[2] If his owner asked him what day of the month the coming Friday would fall on, Hans would tap out the answer—eleven taps for the eleventh day of the month, five taps for the fifth day of the month, just like that. The astonishing thing is that his answers were very often right. He had something like an 89 percent hit rate.[3] Hans understandably caused a stir. He even featured in the *New York Times* under the headline, "Berlin's Wonderful Horse: He Can Do Almost Everything but Talk."[4]

Alas, the adulation wasn't to last. A psychologist investigating his case concluded that Hans wasn't actually performing arithmetic, telling the time or reading German. Rather less sensationally—but still quite remarkably—Hans had become adept at reading his *owner*. It was revealed that as Hans would approach the right number of taps in response to any of the questions put to him, his owner gave unconscious postural and facial cues that Hans was able to pick up on. The owner's tension would be greatest at the second-to-last tap before the correct tap, after which the owner's tension eased. Hans would stop tapping exactly one tap after his owner's tension peaked. So Hans didn't know the answers to the questions he was asked. He gave the right answers alright, but for the wrong reasons.

The lesson resonates with anyone that's guessed their way through a multiple-choice quiz and ended up with a high score, but it's also illustrated vividly today by AI. The risk of a machine learning tool yielding correct results for frankly spurious reasons is high—indeed, worryingly high. In the prologue, we already mentioned how an object classifier trained on pictures of wolves with snow in the background is likely to discriminate between a wolf and a husky based on that single fact alone rather than on features of the animal it's supposed to be recognizing, like the eyes and snout. But this isn't an isolated case. Some classifiers will detect a boat only if there's water around,[5] a train only if there's a railway nearby,[6] and a dumb-bell only if there's an arm lifting it.[7] The moral of the story is simple: we should never trust a technology to make important decisions about anything unless we've got some way of interrogating it. We have to be satisfied that its "reasons" for deciding one way or another actually check out. This makes transparency something of a holy grail in AI circles. Our autonomous systems must have a way for us to be able to scrutinize their operations so we can catch any Clever Hans-type tricks in good time. Otherwise a classifier that works well enough in standard contexts (when there's lots of water around, say, or lots of other gym equipment strewn across the floor) will fail abysmally in nonstandard contexts. And let's face it, the world of high-stakes decision-making nearly always takes place in nonstandard contexts. It's not so much whether a system can handle the standard open-and-shut case that we care about when the Home Office assesses a visa application—it's precisely the *outlier* case that we're most anxious about.

But now what exactly does it mean for a system to be transparent? Although almost everyone agrees that transparency is nonnegotiable, not everyone's clear about what exactly we're aiming for here. What sort of detail do we require? Is there such a thing as too much detail? At what point can we be satisfied that we understand how a system functions, and why it decided *this* way rather than *that*?

Questions like these are vitally important in the brave new world of big data, so we'd like to consider them carefully in this chapter. First, though, let's lay out the terrain a little. After all, transparency covers a *lot* of ground—and here we're only interested in a small patch of it.

The Many Meanings of "Transparency"

"Transparency" can mean a lot of different things. It has general meanings, as well as more specific meanings. It can be aspirational and whimsical but also definite and concrete.

At the broadest level it refers to *accountability* or *answerability*, that is, an agency's or person's responsiveness to requests for information or willingness to offer justification for actions taken or contemplated. This is the most overtly political sense of the term. Importantly, it's a *dynamic* sense—there's never a moment when, as an authority committed to this ideal, your commitment is fully discharged. To be committed to transparency means *being* transparent, not *having been* transparent. We expect our elected representatives to act in the public interest, and transparency stands for their ongoing obligation to meet this expectation. When government is open, answerable, and accountable to its citizens, it is less tempted to become insular, self-serving, and corrupt. Transparency in this broad sense is therefore a safeguard against the abuse of power. All democracies notionally value this sense of transparency, but it's obviously hazy and aspirational. From here, the notion branches out in at least three directions, each of which takes the concept onto much firmer ground.

In one direction, transparency may be associated with moral and legal responsibility (see chapter 4). This captures such familiar notions as blameworthiness and liability for harm. Here the sense of transparency is often static, i.e., "once-for-all" or "point-in-time" (e.g., "judgment for the plaintiff in the amount of $600). Unlike the broader notion we started with, this sense of transparency isn't necessarily dynamic. Rather than prospectively *preventing* wrongdoing, it's most often *corrective* (and retrospective). But, as we point out in chapter 4, responsibility isn't always backward looking and static in this way. Companies that undertake to manufacture goods in a form that will reach the end-user *as is*—that is, without further possibility of the goods being road-tested or otherwise treated—will have a forward-looking responsibility to make sure the goods are safe. This is what lawyers mean by a "duty of care," and it is one example where legal responsibility has some of the qualities we tend to associate with accountability and answerability.

In a second direction, transparency definitely retains its dynamic quality but relates more narrowly to the inspectability (or auditability) of institutions, practices and instruments. Here transparency is about mechanisms: How does this or that tool actually *work*? How do its component parts fit together to produce outcomes like those it is designed to produce? Algorithms can be "inspected" in two ways. First, we can enquire of their provenance. How were they developed, by whom, and for what purpose(s)? This extends to procurement practices. How were they acquired, who commissioned them, on what terms, and—the lawyer's question—*cui bono*, that is, who benefits? This might be called *process* transparency. Second, we can ask of any algorithm, how does it work, what data has it been trained on, and by what logic does it proceed? This might be called *technical* transparency, and centers on the notion of *explainability*. Before any particular decision is reached using an algorithm, we can seek *general* ("*ex ante*") explanations. For instance, in a machine learning case, we can ask whether we're dealing with a decision tree, a regression algorithm, or some mixture. Information about the kind of algorithm we're dealing with can tell us quite a lot about its general principles of operation, and whether algorithm A is better than algorithm B. In the wake of any *particular* algorithmic decision, however, the questions posed can be more specific. Why did the algorithm decide *this* matter in *this* particular way? What are its "reasons" for so deciding? This is to seek a specific, individualized ("*ex post*" or "*post hoc*") explanation. In both of these cases it's important to remember that just because a decision system is explainable doesn't mean that all interested parties will be in a position to understand the explanation. If you want to challenge an automated decision at law, at the very least your lawyers will need to understand something about how the underlying algorithm works. This makes *intelligibility* a further key feature that any explainable system ought to have, and by this, of course, we mean intelligibility *relative to some domain of expertise*. Obviously, intelligibility for legal purposes would be different from intelligibility for software engineering purposes; coders and lawyers will want to know and be able to understand different things about an automated system. A final property that it's desirable for an explainable system to have—especially when we're talking about explanations of automated decisions—is *justifiability*. We don't just want explanations, or even intelligible ones. We also want good, fair, and reasonable explanations based on plausible reasoning.

In a third direction, transparency denotes accessibility. Meaningful explanations of an algorithm may be possible, but they may not be *available*. Intellectual property rights might prevent the disclosure of proprietary code or preclude access to training data so that, even if it were possible to understand how an algorithm operates, a full reckoning may not be possible for economic, legal, or political reasons. Algorithms that are otherwise technically transparent may therefore be "opaque" for nontechnical reasons. Figure 2.1 depicts these various nested and interacting notions of transparency diagrammatically.

In the context of algorithms and machine learning, concerns have been raised about transparency in every one of these senses. The sense which has

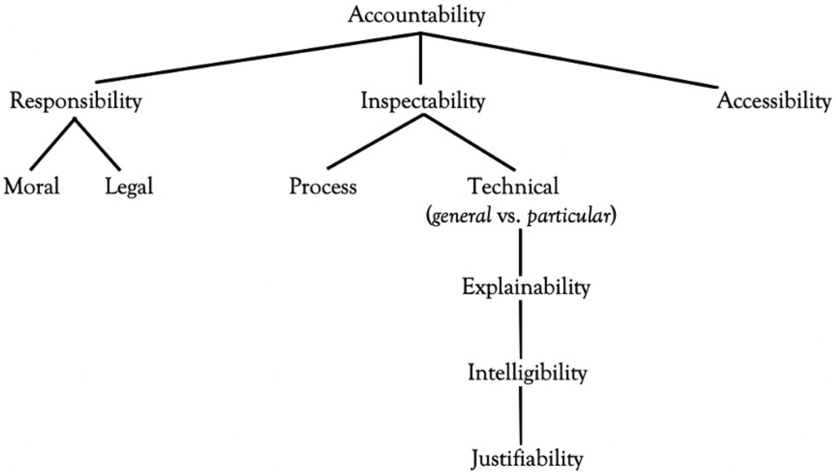

Figure 2.1
The various senses of transparency.

much exercised civil libertarians and a growing number of computer scientists, however, is technical transparency—that is, *explainability* and *intelligibility*. This is what we'll focus on in this chapter. It's the explainability of a system's decisions that helps us sort out the clever from the merely Clever Hans. (Process transparency is also a major concern, but we'll cover that in coming chapters.)

First up, we'll consider why algorithms have a "technical" transparency problem. Why are some algorithms' decisions so difficult to explain and interpret that we must consider them opaque? Second, we'll consider just how much of a problem this really is. We'll set automated decision-making against human decision-making to see how much of the problem is unique to automated systems and how much of it afflicts *all* decision makers, humans included. This is important because if human decision-making poses a similar transparency problem to the one posed by automated decision-making, perhaps we shouldn't be too condescending when discussing the opacity of machine learning systems, insisting on lofty standards of transparency for machines that not even we humans would be capable of meeting.

As you can probably tell, we think it's reasonable for the standards we apply to ourselves when assessing human decisions to set some sort of benchmark when assessing automated decisions. Artificial intelligence, after all, aspires to human-level intelligence in a variety of domains, and these not infrequently involve some form of routine decision-making.

Explanations and Reasons for Decisions

When someone is charged with pronouncing on your personal affairs—your entitlements, your rights, your obligations—we take it for granted that they should be able to justify their decision, even if they don't, strictly speaking, *have* to. We naturally assume, for example, that if a bank declines to give you a loan, the bank didn't just roll a die or flip a coin. If a tenancy tribunal says your rent increase was justified, we expect there to be reasons—in this case laws—behind this determination. Explanations for decisions, whether by courts, public officials, or commercial entities, may not always be forthcoming. Private businesses generally have no obligation to explain themselves to you, and even public agencies aren't always required to furnish reasons for their decisions. But we—all of us—assume that explanations *could* be given, in principle. This is because decision-making shouldn't be arbitrary or capricious. As we noted earlier, it should be based on plausible reasoning.

Of course, explanations sometimes *are* mandatory in public decision-making. This is most clear in legal contexts, where courts of law are generally obliged to provide reasons for their decisions. The obvious rationale for this requirement is to facilitate appeal. Provided there *is* a right to appeal, you need reasons. How are you supposed to appeal a judgment unless you know how the court came to its decision? Generally, the moment a true right of appeal exists, a decision maker's duty to give reasons is implied. On the other hand, if you don't have a right of appeal, there's no strict need for a court to provide reasons. (See box 2.1.)

Of course, the duty to provide reasons may still be imposed by law for other reasons, for example, to further the aims of open justice. Often knowing the reasons why a particular decision has been taken, even if only in rough outline, can engender trust in the process that led to it and confidence that the people in charge of the process acted fairly and reasonably. Thus explanations can have both instrumental value as a means to overturn an adverse determination by appeal and intrinsic value as a democratic index of accountability and transparency in the broadest sense we considered above. In many legal systems, too, statements of reasons function as a source of law—we call it "precedent"—through which, over time,

Box 2.1

Administrative Law and the Difference between "Appeal" and "Review"

Administrative law is the area of law concerned with the actions of public officials (ministers, secretaries, attorneys-general, directors-general, commissioners, etc.). Traditionally, in common-law countries (UK, Canada, Australia, and New Zealand), there was no general duty for public officials to provide reasons for their decisions precisely because there was no true right of appeal from their decisions. Instead, there was a right to *review* (sometimes called "judicial review"), which can be understood as something like the right to appeal the *circumstances* of a decision, but not the decision itself. For example, if a public official is given authority (by legislation) to determine land rates within a municipality but not land *taxes*, one might challenge the official's decision to alter land taxes. Here you aren't said to be *appealing* the decision so much as having it *reviewed*. You don't need a published statement of reasons to review a decision in this sense because you aren't challenging the decision itself (i.e., how the official reasoned from the facts and the law to their conclusion). You're only challenging the *circumstances* of the decision (i.e., the fact that the official did not have power to fix land taxes in the first place). True appeals, in contrast, bring the decision itself into question, not just the circumstances of the decision. This sometimes gets called "merits review." It's the right that litigants frequently exercise in criminal and civil proceedings. A true right of appeal allows the appeal court to examine, in detail, the reasons for the decision maker's determination, including how they assessed the evidence, what conclusions they drew from it, how they interpreted the law, and how they applied that interpretation to the facts of the case. In other words, an appeal allows the appeal court to substitute its own decision for that of the original decision maker. For a true right of appeal to be exercised, clearly the aggrieved party needs a statement of reasons!

consistency in the law is achieved by treating like cases alike. And obviously the habit of providing reasons promotes better quality decision-making all round. A judge conscious of the many hundreds (or even thousands) of lawyers, students, journalists, and citizens poring over their judgment will be anxious to get it right—their reputation is on the line.

Someone's "right" to an explanation is, of course, someone else's "duty" to provide it. Understandably, jurisdictions differ on the existence of this right. Some jurisdictions (like New Zealand) have imposed this duty on public officials.[8] Others (like Australia) have declined to do so (except in the case of judges). Interestingly, the EU is potentially now the clearest case of a jurisdiction requiring explanations for automated decisions in both the public *and* private sectors (see below). Still, it's debatable whether explanation rights provide an adequate remedy in practice. Although they certainly have a place, their effectiveness can easily be overstated, as they do rely on individuals being aware of their rights, knowing how to enforce them, and (usually) having enough money to do so.[9]

In any case, it's hardly surprising that, as algorithmic decision-making technology has proliferated, civil rights groups have become increasingly concerned with the scope for challenging algorithmic decisions. This isn't just because AI is now regularly being recruited in the legal system. AI is also being used by banks to determine creditworthiness, by employment agencies to weed out job candidates, and by security companies to verify identities. Although we can't "appeal" the decisions of private businesses the way we can appeal court decisions, still there are laws that regulate how private businesses are to conduct themselves when making decisions affecting members of the public. Anti-discrimination provisions, for example, prevent the use of a person's ethnicity, sexuality, or religious beliefs from influencing the decision to hire them. A company that uses AI probably purchased the software from another company. How can we be sure the AI doesn't take these prohibited ("protected") characteristics into account when making a recommendation to hire a person or refuse a loan? Do we just *assume* the software developers were aware of the law and programmed their system to comply?

AI's Transparency Problem

Traditional algorithms didn't have a transparency problem—at least not the same one that current deep learning networks pose. This is because traditional algorithms were defined "by hand," and there was nothing the system could do that wasn't already factored into the developer's design for how the system should operate given certain inputs.*

As we saw in chapter 1, however, deep learning is in a league of its own. The neural networks that implement deep learning algorithms mimic

*Of course, some traditional algorithms, like "expert systems," *could* also be inscrutable in virtue of their complexity.

the brain's own style of computation and learning. They take the form of large arrays of simple neuron-like units, densely interconnected by a very large number of plastic synapse-like links. During training, a deep learning system's synaptic weights are adjusted so as to improve its performance. But although this general *learning algorithm* is understood (e.g., the "error backpropagation" method we discussed in chapter 1), the actual *algorithm learned*—the unique mapping between inputs and outputs—is impenetrable. If trained on a decision task, a neural network basically derives its own method of decision-making. And there is the rub—it is simply not known in advance what methods will be used to handle unforeseen information. Importantly, neither the user of the system nor its developer will be any the wiser in this respect. *Ex ante* predictions and *ex post* assessments of the system's operations alike will be difficult to formulate precisely. This is the crux of the complaint about the lack of transparency in today's algorithms. If we can't ascertain exactly why a machine decides the way it does, upon what bases can its decisions be reviewed? Judges, administrators, and public agencies can all supply reasons for their determinations. What sorts of "reasons" can we expect from automated decision systems, and will such explanations be good enough? What we're going to suggest is that if human decision-making represents some sort of gold standard for transparency—on the footing that humans readily and routinely give reasons for their decisions—we think AI can in some respects already be said to meet it.

What Kinds of Explanations Have Been Demanded of Algorithmic Systems?

To date, calls for transparent AI have had a particular ring to them. Often there's talk of inspecting the "innards" or "internals" of an algorithmic decision tool,[10] a process also referred to as "decomposition" of the algorithm, which involves opening the black box to "understand how the *structures within*, such as the weights, neurons, decision trees, and architecture, can be used to shed light on the patterns that they encode. This requires access to the bulk of the model structure itself."[11] The IEEE raises the possibility of designing explainable AI systems "that can provide justifying reasons or other reliable 'explanatory' data *illuminating the cognitive processes* leading to … their conclusions."[12] Elsewhere it speaks of "internal" processes needing to be "traceable."[13]

It isn't just aspirational material that's been couched in this way. Aspects of the EU's General Data Protection Regulation (GDPR) have engendered similar talk.[14] For instance, the Article 29 Data Protection Working Party's draft guidance on the GDPR states that "a complex mathematical explanation about how algorithms or machine-learning work," though not generally relevant, "should also be provided if this is necessary to allow experts to further verify how the decision-making process works."* There's a sense in which this verges on truism. Obviously technical compliance teams, software developers and businesses deploying algorithmic systems may have their own motivations for wanting to know a little more about what's going on "under the hood" of their systems. These motivations may be perfectly legitimate. Still, it's a fair bet that such investigations won't be concerned with AI decisions *as* decisions: they'll be concerned with the technology as a piece of kit—an artifact to be assembled, disassembled and reassembled, perhaps with a view to it making better decisions in due course, bug-fixing, or enhancing human control. When, on the other hand, we want to know why a system has *decided* this way or that, and hence seek *justifying* explanations, we think that most often—although not in every case—the best explanations will avoid getting caught up in the messy, technical internal details of the system. The best explanations, in other words, will resemble human explanations of action.

Human Explanatory Standards

And just what sorts of explanations are human agents expected to provide? When judges and officials supply written reasons, for example, are these expected to yield the entrails of a decision? Do they "illuminate the cognitive processes leading to a conclusion"? Hardly.

It's true that human agents are able to furnish reasons for their decisions, but this isn't the same as illuminating cognitive processes. The cognitive processes underlying human choices, especially in areas in which a crucial element of intuition, personal impression, and unarticulated hunches are

*Strictly speaking, this "good practice" recommendation (Annex. 1) relates to Article 15, not Article 22, of the GDPR. Article 15(1)(h) requires the disclosure of "meaningful information about the logic involved" in certain kinds of fully automated decisions.

driving much of the deliberation, are in fact far from transparent. Arenas of decision-making requiring, for example, assessment of the likelihood of recidivism or the ability to repay a loan, more often than not involve significant reliance on what philosophers call "subdoxastic" factors—factors beneath the level of conscious belief. As one researcher explains, "a large part of human decision making is based on the first few seconds and how much [the decision makers] like the applicant. A well-dressed, well-groomed young individual has more chance than an unshaven, disheveled bloke of obtaining a loan from a human credit checker."[15] A large part of human-level opacity stems from the fact that human agents are also frequently *mistaken* about their real (internal) motivations and processing logic, a fact that is often obscured by the ability of human decision makers to invent *post hoc* rationalizations. Often, scholars of explainable AI treat human decision-making as privileged.[16] Earlier we noted that some learning systems may be so complex that their manipulations defy systematic comprehension and that this is most apparent in the case of deep learning systems. But the human brain, too, is largely a black box. As one scholar observes:

> We can observe its inputs (light, sound, etc.), its outputs (behavior), and some of its transfer characteristics (swinging a bat at someone's eyes often results in ducking or blocking behavior), but we don't know very much about how the brain works. We've begun to develop an algorithmic understanding of some of its functions (especially vision), but only barely.[17]

No one doubts that well-constructed, comprehensive, and thoughtful human reasons are extremely useful and generally sufficient for most decision-making purposes. But in this context, usefulness and truth aren't the same. Human reasons are pitched at the level of what philosophers call "practical reason"—the domain of reason that concerns the justification of action (as distinguished from "epistemic" or "theoretical reason," which concerns the justification of belief). Excessively detailed, lengthy, and technical reasons are usually not warranted, or even helpful, for most practical reasoning. This doesn't mean that the structure of typical human reasoning is ideal in all circumstances. It means only that for most purposes it will serve adequately.

Consider decisions made in the course of ordinary life. These are frequently made on the approach of significant milestones, such as attaining the age of majority, entering into a relationship, or starting a family, but they most often involve humdrum matters (should I eat in, or go out for dinner tonight?). Many of these decisions will be of the utmost importance to the person making them and may involve a protracted period of deliberation (e.g., what career to pursue, whether to marry, whether and when to have children, and what to pay for a costly asset—a home, a college education, etc.). But the rationales that may be expressed for them later on, perhaps after months of research or soul-searching, will not likely assume the form of more than a few sentences. Probably there will be one factor among three or four that reveals itself after careful reflection to be the most decisive, and the stated *ex post* reasons for the decision will amount to a statement identifying that particular factor together with a few lines in defense.

Actually, if you think about it, most "official" decision-making is like this too. It might concern whether to purchase new equipment, whether to authorize fluoridation of a town's water supply, whether to reinstate someone unfairly dismissed from a place of employment, whether to grant bail or parole, or whatever. But basically, the decision's formal structure is the same as that of any other decision, public, personal, commercial, or otherwise. True, the stakes may be higher or lower, depending on what the decision relates to and how many people will be affected by it. Also, the requirement to furnish reasons—as well as the duty to consider certain factors—may be mandated in the one case and not the other. But the primary difference isn't at the level of form. Both contexts involve practical reasoning of a more or less systematic character. And furnishing explanations that are more detailed, lengthy, or technical than necessary is likely to be detrimental to the aims of transparency, regardless of the public or private nature of the situation.

There are, of course, some real differences between public and private decision-making. For example, certain types of reasons are acceptable in personal but not public decision-making. It may be fine to say, "I'm not moving to Auckland because I don't like Auckland," but the same sort of reasoning would be prohibited in a public context. Furthermore, public decision-making often takes place in groups to mitigate the "noisiness" of individual reasoners, such as committees, juries, and appeal courts—although even here, many private, purely personal decisions (regarding, e.g., what to study, which career to pursue, whether to rent or purchase, etc.), are also frequently made in consultation with friends, family, mentors, career advisers, and so on. In any case, these differences don't detract from their fundamentally identical structure. Either way, whether there are more or fewer people involved in the decision-making process (such as jurors, focus groups, etc.) or whether there are rights of appeal, both decision procedures employ practical reasoning and take beliefs and desires as their inputs. Take judicial decision-making—perhaps the most constrained and regimented form of official reasoning that exists. Judicial reasoning is, in the first instance, supposed to appeal to ordinary litigants seeking the vindication of their rights or, in the event of a loss, an explanation for why such vindication won't be forthcoming. So it simply must adopt the template of practical reason, as it must address citizens in one capacity or another (e.g., as family members, corporate executives, shareholders, consumers, criminals, etc.). Even in addressing itself to lawyers, for example, when articulating legal rules and the moral principles underpinning them, it cannot escape or transcend the bounds of practical (and moral) reasoning.[18]

We're not claiming that these insights are in any sense original, but we do think they're important. As we've intimated, the decision tools co-opted in predictive analytics have been pressed into the service of practical reasoning. The aim of the GDPR, for instance, is to protect "natural persons" with regard to the processing of "personal" data (Article 1). Articles 15 and 22 concern a data subject's "right" not to be subject to a "decision" based solely on automated processing, including "profiling." The tools that have attained notoriety for their problematic biases, such as PredPol (for hot-spot policing) and COMPAS (predicting the likelihood of recidivism), likewise involve software intended to substitute or supplement practical human decision-making (for instance, by answering questions such as, how should we distribute police officers over a locality having X geographical characteristics? What is the likelihood that this prisoner will recidivate? Etc.). Explanations sought from such technologies should aim for levels that are appropriate to practical reasoning. Explanations that would be too detailed, lengthy, or technical to satisfy the requirements of practical reasoning shouldn't be seen as ideal in most circumstances.

It is a little odd, then, that many proposals for explainable AI assume (either explicitly or implicitly) that the innards of an information processing system constitute an acceptable and even ideal level at which to realize the aims of transparency. A 2018 report by the UK House of Lords Select Committee on Artificial Intelligence is a case in point. On the one hand, what the Committee refers to as "full technical transparency" is conceded to be "difficult, and possibly even impossible, for certain kinds of AI systems in use today, and would in any case not be appropriate or helpful in many cases."[19] On the other hand, something like full technical transparency is "imperative" in certain safety-critical domains, such as in the legal, medical, and financial sectors of the economy. Here, regulators "must have the power to mandate the use of more transparent forms of AI, even at the potential expense of power and accuracy."[20] The reasoning is presumably that whatever may be lost in terms of accuracy will be offset by the use of simpler systems whose innards can at least be properly inspected. So you see what's going on here. Transparency of an exceptionally high standard is being trumpeted for domains where human decision makers themselves are incapable of providing it. The effect is to perpetuate a double standard in which machine tools must be transparent to a degree that is in some cases unattainable in order to be considered transparent at all, while human decision-making can get by with reasons satisfying the comparatively undemanding standards of practical reason. On this approach, if simpler and more readily transparent systems are available, these should be preferred even if they produce decisions of inferior quality. And so the double standard threatens to prevent deep learning and other potentially novel AI techniques from being implemented in just those domains that could be revolutionized by them. As the committee notes:

> We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate *a full and satisfactory explanation for the decisions it will take*. In cases such as deep neural networks, where it is not yet possible to generate *thorough* explanations for the decisions that are made, this may mean delaying their deployment for particular uses until alternative solutions are found.[21]

This might be a sensible approach in some cases, but it's a dangerous starting position. As the committee itself noted, restricting our use of AI only to what we can fully understand limits what can be done with it.[22] There are various high-stakes domains, particularly in clinical medicine and psychopharmacology, where insisting on a thorough understanding of a technology's efficacy before adopting it could prove hazardous to human wellbeing.

Unconscious Biases and Opacity in Human Decision Making

It's a widely accepted fact that "humans are cognitively predisposed to harbor prejudice and stereotypes."[23] Not only that, but "contemporary forms of prejudice are often difficult to detect and may even be unknown to the prejudice holders."[24]

Recent research corroborates these observations. It seems that the tendency to be unaware of one's own biases is even present in those with regular experience of having to handle incriminating material in a sensitive and professional manner. In a recent review of psycho-legal literature comparing judicial and juror susceptibility to prejudicial publicity, the authors note that although "an overwhelming majority of judges and jurors do their utmost to bring an impartial mind to the matters before them … even the best of efforts may nonetheless be compromised."[25] They write that "even accepting the possibility that judges do reason differently to jurors, the psycho-legal research suggests that this does not have a significant effect on the fact-finding role of a judge,"[26] and that "in relation to prejudicial publicity, judges, and jurors are similarly affected."[27]

Findings like these should force us to reassess our attitudes to human reasoning and question the capabilities of even the most trusted reasoners. The practice of giving reasons for decisions may be simply insufficient to counter the influence of a host of factors, and the reasons offered for human decisions can well conceal motivations hardly known to the decision makers themselves. Even when the motivations *are* known, the stated reasons for a decision can serve to cloak the true reasons. In common law systems it's well known that if a judge has decided upon a fair outcome, and there's no precedent to support it, the judge might just grope around until *some* justification can be extracted from what limited precedents do exist.[28]

Sometimes in discussions about algorithmic transparency you hear people cite the possibility of *appealing* human decisions, as though this makes a real difference to the kind of transparency available from human decision makers. Although it's understandable to view courts and judges as paradigms of human decision-making, what's often forgotten is that legal rights of appeal are quite limited. They can rarely be exercised automatically. Often the rules of civil procedure will restrict the flow of appeals from lower courts by requiring appeal courts to "grant leave" first.[29]

But the situation is worse than this. Substantial parts of judicial reasoning are effectively immune from appeal (albeit for pragmatic reasons), even in the lowest courts, *and even when appeals are theoretically possible!* A degree of judicial discretion probably has to be exercised in every case, and yet discretion can often only be appealed within severely narrow limits.[30] Given how frequently judges are called on to exercise their discretion, this could be seen as contrary to the principles of open justice. Judges are also allowed considerable leeway in respect of their findings on witness credibility. Appeal courts are generally reluctant to overturn judicial determinations of credibility, because the position of trial judges in being able to assess the demeanor of a witness at first hand is seen to deserve particular respect. And let's not forget that jury deliberations are quintessential black boxes. No one (apart from the jurors themselves) can know why a jury decided the way it did, even when appeals from their verdicts are possible. How's that for transparency!

But let's dig a little deeper into the cognitive underground. The purely neurophysiological aspects of human decision-making are not understood beyond general principles of interneural transmission, excitation, and inhibition. In multi-criterion decision cases in which a decision maker must juggle a number of factors and weigh the relevance of each in arriving at a final decision, one hypothesis suggests that the brain eliminates potential solutions such that a dominant one ends up inhibiting the others in a sort of "winner takes all" scenario.[31] Although this process is to some extent measurable, "it is essentially hidden in the stage where weights or relative importance are allocated to each criterion."[32] It serves as a salutary reminder that even when a sentencing judge provides reasons allocating weights to various statutory factors, the actual inner processing logic behind the allocation remains obscure.

More general work on the cognitive psychology of human decision-making is no less sobering. "Anchoring" and "framing" effects are well known to researchers in the field. One such effect, the "proximity" effect, results in more recent events having greater weight than older ones and bearing a greater influence on choices in the search for solutions.[33] The tendency to see false correlations where none exists is also well documented.[34] This bias is at its strongest when a human subject deals in small probabilities.[35] Finally, constraints imposed by short-term memory capacity mean we can't handle more than three or four relationships at a time.[36] Because it's in the nature of complex decisions to present multiple relationships among many issues, our inability to concurrently assess these factors constitutes a significant limitation on our capacity to process complexity.

The upshot is simple: let's not pretend we humans are paragons of transparency next to those unfathomable black-boxes we call deep networks.

Explainable AI 2.0

We've suggested that because the demands of practical reason require the justification of action to be pitched at the level of practical reason, decision tools that support or supplant practical reasoning generally shouldn't be expected to aim for a standard any higher than this. In practice, this means that the sorts of explanations for algorithmic decisions that are analogous to ordinary, run-of-the-mill, interpersonal explanations should be preferred over ones that aim at the architectural innards of a decision tool. The time has come to flesh this out. What exactly would these analogues of everyday explanations look like?

Modern predictive models operating in real-world domains tend to be complex things, regardless of which machine learning methods they use (see chapter 1). If we want to build a predictive system that can convey to a human user *why* a certain decision was reached, we have to add functionality that goes beyond what was needed to generate the decision in the first place. The development of "explanation tools" that add this functionality is a rapidly growing new area of AI.

The basic insight behind the new generation of explanation tools is that, to understand how one predictive model works, *we can train another predictive model to reproduce its performance*. Although the original model can be very complex and optimized to achieve the best predictive performance, the second model—the "model-of-a-model"—can be much simpler and optimized to offer maximally useful explanations.

Perhaps the most useful thing a decision subject wants to know is how different factors were weighed in coming to a final decision. It's common for human decision makers to disclose these allocations, even if, as we mentioned earlier, the inner processing logic leading to them remains obscure. Weights are classic exemplars of everyday logic, and one way for algorithmic decision tools to be held accountable in a manner consistent with human decision-making is by having them divulge their weights.[37] Some of the most promising systems in this area are ones that build a local model of the factors most relevant to any given decision of the system being explained.[38] Such "model-of-a-model" explanation systems also have the added benefit of being able to explain how a system arrived at a given decision without revealing any of its internal workings. This should please tech firms. By providing explanations of how their software "works," tech companies needn't worry that they'll necessarily be disclosing their patented "secret sauce" at the same time. This feature of model-of-a model systems shouldn't be surprising. Remember, they aren't actually telling you "*this* is how the algorithm decided X, *this* is how the algorithm works." Instead, as their very name implies, they're providing a *model*, and a model only has to give a high-level, simplified description of something. Rather like the London Tube map—its depiction of the London Underground is economical and compact, which is no doubt what makes it useful for catching the tube, but obviously no one thinks it provides a reliable guide to London's topography. It'd be fairly useless for navigating at street level, for example.

Still, you might wonder, can explanation systems go one better and give you explanations that reflect more faithfully how an algorithm actually decides while still being intelligible—indeed, while mimicking the structure of our own humanoid logic? The answer, it seems, is yes. Researchers at Duke University and MIT built an image classifier that not only provides comprehensible human-style explanations, but that actually proceeds in accordance with that very logic when classifying images.[39] Imagine for a moment you're presented with a picture of a bird, and your job is to recognize what species it belongs to—a pretty standard object classification task, you might say. It's not always easy, even for trained ornithologists, to correctly classify a bird just by looking at an image. There are simply too many species to contend with. But let's say you were to have an educated guess. How would you explain your answer? We considered this sort of problem when we discussed the huskies and wolves. Remember we noted there that if you were going to justify why you thought an image of a husky was actually a wolf, there'd be certain things you'd point out: the eyes, the ears, and the snout, most likely. In other words you'd *dissect* the image and point out that this or that part of the image is typical for this and that species of canine. The combined weight of all these "typical" features would point us in the direction of one species over another. In the prologue, we noted that many object classifiers would not reason this way, instead focusing on the absence or presence of snow in the image (which is understandable given the likely number of images of wolves in the training set with snow in the background but still absolutely ridiculous). Well, the researchers at Duke and MIT managed to build a bird species classifier that reasons more or less as we would—by dissecting the bird image and comparing selected parts against species-typical parts in a training set. Importantly, this isn't just how the classifier reasons to a conclusion—it's also how it *explains* its conclusions. As the team put it, "it has a transparent reasoning process that is *actually* used to make predictions."[40] Win-win.

For a few years now, companies have resisted providing explanation systems, taking refuge behind the excuse that it's either too difficult, that the explanations would be incomprehensible, or that disclosure risks compromising trade secrets. Slowly, the winds are changing, and even big players are beginning to see that explainability isn't just important—it's potentially commercializable. There's now "a growing industry of consultancy firms that claim to provide algorithmic auditing 'as a service' in order to assess existing algorithms for accuracy, bias, consistency, transparency, fairness, and timeliness."[41] Google and IBM are also getting in on the act. IBM has launched its very own explanation tool—its cloud-based open-source software will enable customers "to see, via a visual dashboard, how their algorithms are making decisions and which factors are being used in making the final recommendations. It will also track the model's record for accuracy, performance, and fairness over time."[42] And none other than Google itself has launched a "what-if" tool, "also designed to help users look at how their machine-learning models are working."[43]

Double Standards: Good or Bad?

A crucial premise of this chapter has been that the standards of transparency should be applied consistently, regardless of whether we're dealing with humans or machines. Partly this is because the field of AI to some extent makes human achievement a standard worth striving for and partly this is because (we have suggested) humans and machines are both opaque in their own ways. Differences will naturally arise when one system is organic and the other is synthetic, but these differences don't seem to justify adopting different standards of transparency. This isn't to say there are *no* circumstances in which different standards might be required—there certainly are. But they are exceptional (and probably a little too esoteric to get into here).[44] Instead of considering such cases, let's consider a few other factors which probably *don't* justify imposing different standards on humans and machines, although they might seem to do so at first.

One factor we can think of is the potential of AI to advance well beyond the level of human transparency. If algorithmic decision tools have a good chance of being significantly better than humans at explaining themselves, then regulations probably *should* be crafted with a view to bringing out the best that they can be—even if this means setting a regulatory standard that would be far stricter than one we'd ever apply to ourselves. However, we have our doubts about just how much less of a black box a multimillion-neuron deep network is likely to be than a human brain, in *practical* terms anyway. If an artificial intelligence is only *in principle* less opaque than a human intelligence, but not in practice, the two intelligences would be comparably opaque, and a double standard hard to justify.

Another argument for double standards might run as follows. The kind of decisions we're worried about when discussing algorithmic decision-making are decisions regarding policies that affect third parties. In these situations, procedures are in place to minimize individuals' biases, such as expert reports, committees, and appeal mechanisms. And such procedures might be thought to tip the scales in favor of human decision-making, justifying a more lenient standard of transparency.

Now we've already pointed out why appeal mechanisms are restrictive and limited in their potential to reduce bias. Regarding committees, we cited a recent paper demonstrating that both juries (a type of committee) and judges are vulnerable to prejudicial media publicity. So just having more people involved in a decision doesn't necessarily eliminate or reduce the potential for human bias to interfere with human reasoning. As for expertise, judges are a type of expert, and, as we said, even when their own motivations are known, the stated reasons for their decisions can serve to cloak the true reasons. But maybe there's more to this point about committees. Here the thought is that a high standard of transparency is naturally enforced by processes within a group, because members often need to justify and rationalize their points of view, which are typically challenged or queried in the ordinary course of discussion.

But actually, research in social psychology suggests that group-based mechanisms that ensure the *production* of justifications don't always guarantee their *quality*. In fact, participants in a group are often swayed by the mere presence of a justification, regardless of its quality. A classic study found that intrusions into a photocopier queue were more likely to be tolerated if a justification was provided, even if it was devoid of content.[45] "May I use the Xerox machine because I have to make copies?" was more effective than "May I use the Xerox machine?" Of course, the result speaks directly to the dynamics of an informal group setting, not a high-level public committee, but it has been taken seriously by legal theorists in discussions of legitimacy.[46] Thus group processes, which naturally elicit justifications, don't necessarily improve on solo decision-making. And anyway, even if it could be shown that a *single* machine's decisions were less transparent than those made by a group of people, this would seem less a shortcoming of algorithms than an asymmetry in the systems being compared. A decision made by *one* person would, for the same reason, be less transparent than a decision made by a *group* of people.

Summing Up

We've tried to expose an assumption behind many of the calls for more explainable, transparent AI systems. The assumption is that it's reasonable to impose a higher standard of transparency on AI systems than would ordinarily be imposed on human decision makers. Or perhaps the assumption is simply that human decisions are generally more transparent than algorithmic decisions because they can be inspected to a greater depth, with the hefty standards imposed on machines serving merely to level the playing field. We've suggested that both assumptions are false. At this stage, the sorts of explanations we can't obtain from AI we can't obtain from humans either. On a somewhat brighter note, the sorts of explanations we *can* (and *should*) expect from human beings may be increasingly possible to obtain from AI systems.