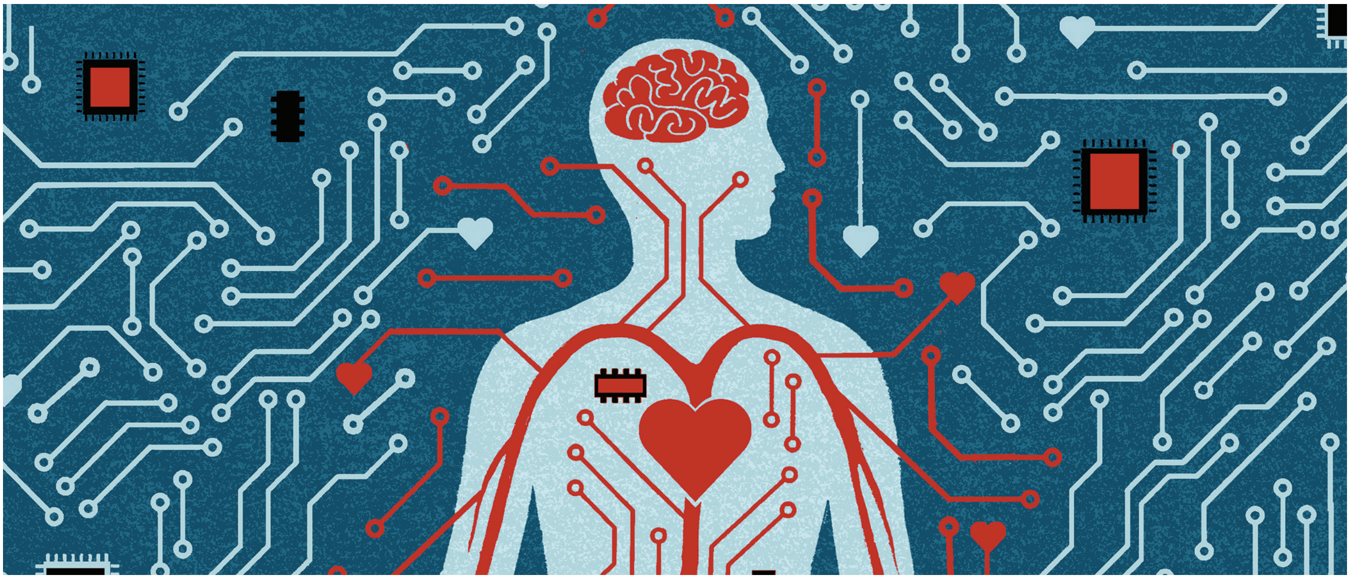
Deloitte
Review

ISSUE 26, JANUARY 2020

COMPLIMENTARY ARTICLE REPRINT

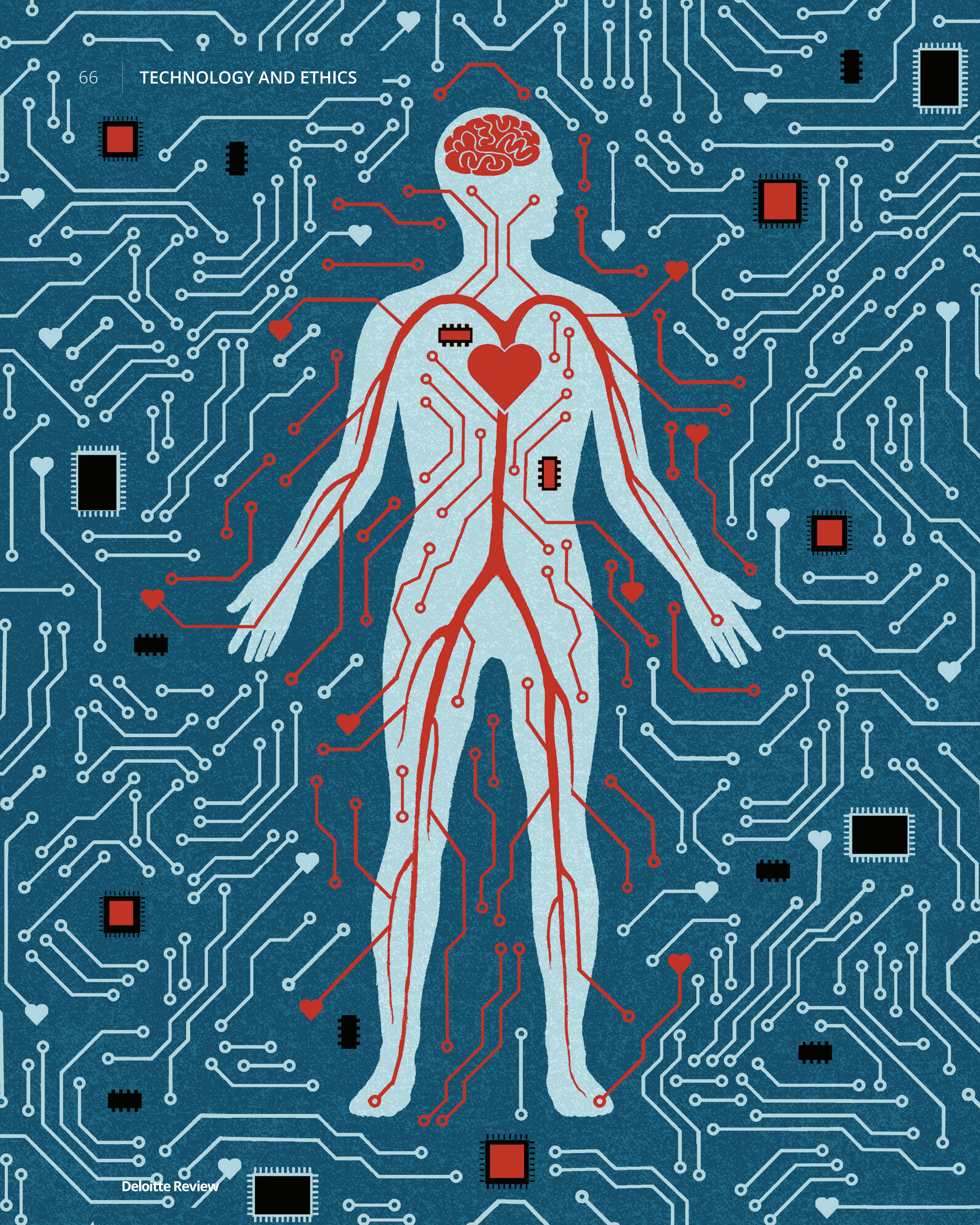


# Human values in the loop: Design principles for ethical AI

by James Guszcza, Michelle A. Lee, Beena Ammanath, and Dave Kuder

ILLUSTRATION BY JAMES STEINBERG

Deloitte.
Insights



# Human values in the loop

## Design principles for ethical AI

BY JAMES GUSZCZA, MICHELLE A. LEE,
BEENA AMMANATH, AND DAVE KUDER

*ILLUSTRATION BY JAMES STEINBERG*

**“This has to be a human system we live in.”**

— *Sandy Pentland*[1]

MORE THAN 60 years after the discipline’s birth,[2] artificial intelligence (AI) has emerged as a preeminent issue in business, public affairs, science, health, and education. Algorithms are being developed to help pilot cars, guide weapons, perform tedious or dangerous work, engage in conversations, recommend products, improve collaboration, and make consequential decisions in areas such as jurisprudence, lending, medicine, university admissions, and hiring. But while the technologies enabling AI have been rapidly advancing, the societal impacts are only beginning to be fathomed.

Until recently, it seemed fashionable to hold that societal values must conform to technology’s natural evolution—that technology should shape, rather than be shaped by, social norms and expectations. For example, Stewart Brand declared in 1984 that “information wants to be free.”[3] In 1999, a Silicon Valley executive told a group of reporters, “You have zero privacy ... get over it.”[4] In 2010, *Wired* magazine cofounder Kevin Kelly published a book entitled *What Technology Wants*.[5] “Move fast and break things” has been a common Silicon Valley mantra.[6]

But this orthodoxy has been undermined in the wake of an ever-expanding catalog of ethically fraught issues involving technology. While AI is not the only type of technology involved, it has tended to attract the lion’s share of discussion about the ethical implications.

Many concerns about AI-enabled technologies have been well-publicized. To cite a few: AI algorithms embedded in digital and social media technologies can reinforce societal biases, accelerate the spread of rumors and disinformation, amplify echo chambers of public opinion, hijack our attention, and impair mental well-being.[7] Experts warn of AI technologies being weaponized. Semiautonomous vehicles have been reported to fail in ways the owners did not expect.[8] And while fears of “smart” technologies stealing human jobs are often overstated, respected economists highlight growing inequality and lack of opportunity for certain workforce segments due to technology-induced workplace changes.[9]

**Until recently, it seemed fashionable to hold that societal values must conform to technology's natural evolution—that technology should shape, rather than be shaped by, social norms and expectations.**

Thanks in part to concerns like these, there have been increasing calls for AI to be designed and adopted in ways that reflect important cultural values. In a recent editorial, the investor Stephen Schwarzman urged companies to take the lead in addressing ethical concerns surrounding AI. He comments, “If we want to realize AI’s incredible potential, we must also advance AI in a way that increases the public’s confidence that AI benefits society. We must have a framework for addressing the impacts and the ethics.”[10]

And indeed, a large number of AI ethics frameworks have appeared in recent years. For example, a team at the Swiss university ETH Zurich recently analyzed no fewer than 84 AI ethics declarations from a variety of companies, government agencies, universities, nongovernmental organizations, and other organizations.[11] While the team identified some inconsistencies, there is also reassuring overlap in the broad principles articulated. In another such effort, the AI4People group led by Luciano Floridi analyzed six high-profile AI ethics declarations. They concluded that a set of four abiding, higher-level ethical principles—*beneficence, non-maleficence, justice,* and *autonomy*—captured much of these six declarations' essence.

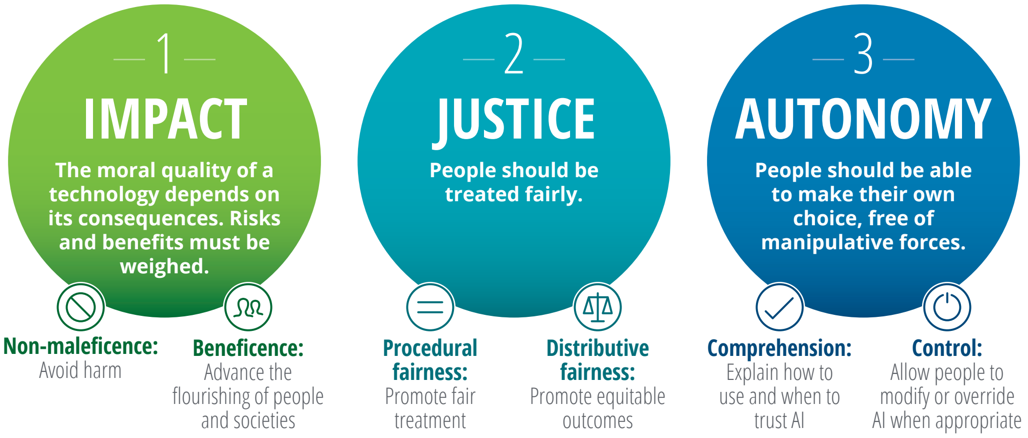
These four principles are rooted in major schools of ethical philosophy and, in fact, have been widely embraced in the field of bioethics for several decades.[12] It is perhaps unsurprising that they adapt well to the AI context. Writing for *Harvard Data Science Review*, Floridi and coauthor Josh Cowls note: "Of all areas of applied ethics, bioethics is the one that most closely resembles digital ethics in dealing ecologically with new forms of agents, patients, and environments."[13]

In his book *Bit by Bit*, the prominent computational social scientist Matthew Salganik has recently advocated the same core principles to help data scientists evaluate the ethical implications of working with human-generated behavioral data. Salganik comments: "In some cases, the principles-based approach leads to clear, actionable solutions. And when it does not, it clarifies the tradeoffs involved, which is critical for striking an appropriate balance. Further, the principles-based approach is sufficiently general that it will be helpful no matter where you work."[14]

This essay attempts to illustrate that ethical principles can serve as *design principles* for organizations seeking to deploy innovative AI technologies that are economically profitable as well as beneficial, fair, and autonomy-preserving for people and societies. Specifically, we propose *impact, justice,* and *autonomy* as three core principles that can usefully guide discussions around AI's ethical implications.

FIGURE 1

## Three core principles can help leaders think through AI's ethical implications


1
IMPACT
The moral quality of a technology depends on its consequences. Risks and benefits must be weighed.
Non-maleficence: Avoid harm
Beneficence: Advance the flourishing of people and societies
2
JUSTICE
People should be treated fairly.
Procedural fairness: Promote fair treatment
Distributive fairness: Promote equitable outcomes
3
AUTONOMY
People should be able to make their own choice, free of manipulative forces.
Comprehension: Explain how to use and when to trust AI
Control: Allow people to modify or override AI when appropriate

Source: Deloitte analysis.

looking back

### Deloitte on computers and human capabilities

1966

> While computers will be a major factor in our lives in the years ahead, they will not obsolete either modern man or modern management. Computers supplement of skills of man; expand the horizons of man's knowledge; endow him with new power to resolve problems, and to explore new ones.

**Robert M. Trueblood**

Chairman, Touche Ross,
Bailey & Smart

Achieving ethical, trustworthy, and profitable AI requires that ethics deliberations be grounded in a scientific understanding of the relative strengths and weaknesses of both machine intelligence and human cognition. In short, being "wise" about AI presupposes being "smart" about AI. For example, discussing ways to promote safe and reliable AI requires understanding why AI technologies—often created using forms of large-scale statistical analysis such as deep learning—succeed in some contexts but fail in others. Likewise, discussions of algorithmic fairness should be informed by both an appreciation of the biases and "noise" that affect unaided human decisions, and an understanding of the tradeoffs involved in different conceptions of algorithmic "fairness." In each case, ethical deliberations are more productive when informed by the relevant science.

At the same time, this essay does not prescribe *how to apply* the core principles. Organizations differ in their goals and operating contexts, and will therefore adopt different declarations, frameworks, rule sets, and checklists to help guide the responsible development of AI technologies. Furthermore, applying fundamental principles to specific problems often requires evaluating tradeoffs between alternatives whose perceived relative importance varies across individuals, organizations, and societies. We suggest that a grasp of core principles can help individuals and organizations more effectively create ethical frameworks and deliberate specific issues.

## Impact: Promoting acceptable outcomes

Two widely recognized ethical principles are *non-maleficence* ("do no harm") and *beneficence* ("do only good"). These principles are grounded in "consequentialist" ethical theory, whose proponents have included John Stuart Mill and Jeremy Bentham, and which holds that the moral quality of an action depends on its consequences.[15]

## NON-MALEFICENCE: AVOID HARM

### *Themes*

Safety, reliability, robustness, data provenance, privacy, cybersecurity, misuse

### *Examples*

- Refraining from causing intentional harm through phishing, cyber breaches, weaponized AI, or “fake news”
- Avoiding unintentional harm due to false positives, faulty data, poor model specification, or poor algorithm operationalization

## “FIRST, DO NO HARM”

Non-maleficence prescribes that AI should avoid causing both foreseeable and unintentional harm. Examples of the former could include weaponized AI,[16] the use of AI in cyberwarfare, malicious hacking, the creation or dissemination of phony news or images to disrupt elections, and scams involving phishing and fraud. But of course, the great majority of organizations building or deploying AI have no intention of causing needless harm. For them, avoiding unintended consequences is the paramount concern.

Avoiding harmful AI requires that one understand AI technologies’ scientific limitations in order to manage the attendant risks. For example, many AI algorithms are created by applying machine learning techniques, most prominently deep learning, to large bodies of “labeled” data.[17] The resulting algorithms can then be deployed to make predictions about future cases for which the true values are unknown. Such algorithms are used today to estimate the likelihood that a borrower will repay a loan, a student’s expected grade point average if admitted to a university, the odds that an X-ray image is a cancerous tumor, or the chances that the red object in front of a car is a stop sign.

The “artificial intelligence” moniker notwithstanding, however, these algorithms are not based on the sorts of conceptual understanding characteristic of human intelligence.[18] Rather, they are the product of statistical pattern-matching. Therefore, if automatic techniques or naïve statistical methodologies are used to train algorithms on data that contain inaccuracies or biases, those algorithms themselves might well reflect those inaccuracies or biases. This basic truth of machine learning has a key ethical implication: A machine learning algorithm is only safe and reliable to the extent that it is trained on (1) sufficient volumes of data that (2) are suitably representative of the scenarios in which the algorithm is to be deployed.

A case study discussed by the prominent machine learning researcher Michael I. Jordan illustrates how a failure to appreciate such risks can lead to physical harm. In this case, an AI device was designed to estimate the likelihood of a fetus having Down syndrome based on ultrasound images. At a certain point, the input data’s format, the resolution of the ultrasound images, changed: The AI began processing higher-resolution images to compute its estimates. This change resulted in a significant uptick in the machine’s Down syndrome diagnoses. This uptick was due not to previously unrecognized cases, but to the images’ higher resolution producing spurious statistical artifacts which the algorithm (trained on lower-resolution images) misinterpreted as Down syndrome indicators. It is likely that thousands of people opted for amniocentesis procedures, putting their babies at risk, based on these faulty diagnoses.[19]

Knowing that machine learning algorithms perform reliably only to the extent that the data used to train them suitably represents the scenarios in which they are deployed, an organization can take steps to identify and mitigate the risks arising from this limitation. Some tactics might include:[20]

- Assessing the training data's provenance—where the data arose, what inferences were drawn from the data, and how relevant those inferences are to the present situation—to assess an algorithm's applicability.

- Restricting algorithms' use to environments in which they are likely to be reliable. For example, autonomous vehicles could be restricted to special lanes that are off-limits to (unpredictable) human drivers, pedestrians, and animals.[21] Similarly, a chatbot could be designed to avoid collecting personally identifiable information (PII), or to ignore certain words in order to lessen the risk of being gamed.[22]

- Coupling humans, who are capable of common-sense reasoning and flexible decision-making, with algorithmic systems. For example, a semiautonomous vehicle could use AI not to replace the human operator, but to help him or her drive more safely.[23] Similarly, rather than replacing human experts (such as physicians, caseworkers, judges, claims adjusters, teachers grading student papers, or editors flagging unacceptable social media content), algorithms can be designed to help manage workloads and debias these experts' decisions by providing statistically derived indications. In high-stakes scenarios, a pragmatic default might be to assume the need for human-computer collaboration, and treat full machine autonomy as a limiting case.[24]

**Being "wise" about AI presupposes being "smart" about AI.**

### BENEFICENCE: ADVANCE THE FLOURISHING OF PEOPLE AND SOCIETIES

***Themes***

Human flourishing, well-being, dignity, common good, sustainability

***Examples***

- Using AI to improve medical care, deliver public benefits, create safer environments, or improve educational outcomes

## AI FOR GOOD

The principle of beneficence, reflected in many AI ethics declarations, holds that AI should be designed to help promote the well-being of people and the planet. In the book *Tools and Weapons*, Brad Smith used the term "inclusivity" to denote a similar idea, citing AI technologies created to help people overcome visual or hearing impairment. Beneficent AI applications can run the gamut of physical and emotional well-being, and operate at both individual and collective levels. Some examples are:

- An early application of affective computing (also called "emotional AI") that aimed to help autistic people, who characteristically have difficulty inferring other's emotional states, better navigate social situations.[29] Such deep learning-based systems can often infer emotional states from facial expressions better than many humans, and thereby function as "emotional hearing aids" to help decipher others' behavioral cues.

## AI'S COMMON-SENSE GAP

AI is best defined in functional terms as *any* kind of computer program capable of achieving a specific goal ordinarily associated with human intelligence.[25] At one end of the spectrum, AI encompasses such rule-based systems as robotic process automation (RPA). At the other end, many of today's headline-grabbing applications result essentially from large-scale statistical analysis: The application of supervised machine learning techniques to large data sets. One of these techniques in particular, known as "deep learning," underlies many familiar AI applications, such as chatbots and the image recognition systems used to help pilot cars or flag tumors in X-rays.

When researchers first introduced the term "artificial intelligence" in the 1950s, the aspiration was to build computer systems that manifest human-level general intelligence. Today, however, "AI" has largely come to denote more focused, narrow applications that do not possess the flexibility of human thought. The old idea that "general" AI would mimic human cognition has given way to today's multitude of practical, narrow AIs that operate very differently from the human mind.

Unlike human intelligence, AI algorithms do not possess common sense, conceptual understanding, notions of cause-and-effect, or intuitive physics. As an illustration, a human can use common sense and contextual awareness to learn a new bit of slang based on just a few encounters. A machine translation algorithm, in contrast, would need to be exposed to many pretranslated examples to hopefully get it right.[26]

Their lack of common sense, the inability to generalize or to consider context, makes AI algorithms "brittle," meaning that they cannot handle unexpected scenarios or unfamiliar situations. As Gary Marcus and Ernest Davis comment in their book *Rebooting AI*:

> Without a rich cognitive model, there can be no robustness. About all you have instead is a lot of data, accompanied by a hope that new things won't be too different from those that have come before. But that hope is often misplaced, and when new things *are* different enough from what happened before, the system breaks down.[27]

Some commentators have suggested that the auto industry's overly optimistic forecasts of the arrival of fully autonomous vehicles were likely influenced by the neglect of this fundamental point.[28]

- "Data for good" initiatives that use AI algorithms to identify high-poverty areas by analyzing satellite imagery, flag houses that pose a high risk of lead poisoning to their residents, recognize which high school students are at risk of not graduating on time, or identify police officers at greater risk of experiencing adverse events.[30]

- Chatbots that deliver cognitive behavioral therapy interventions to help ameliorate conditions such as low-level depression and anxiety.[31]

- Social robots that incorporate "growth mindset" interventions to help children stay focused in learning environments.[32]

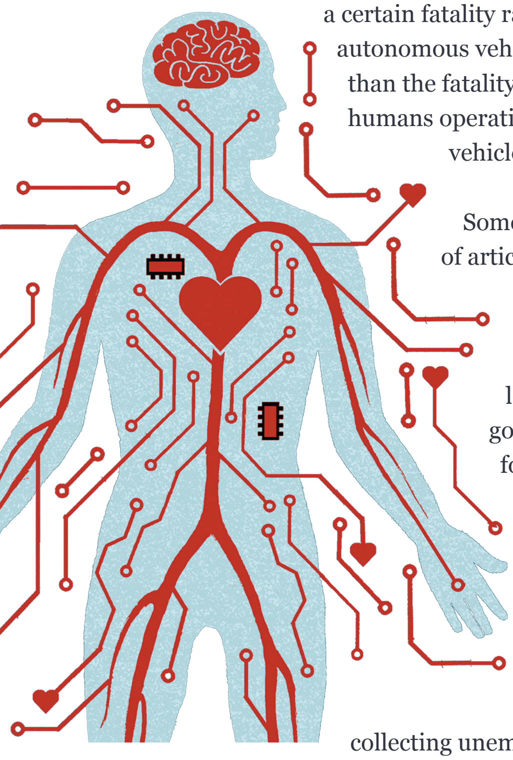
- Wearables paired with gamification or other behavioral "nudge" interventions designed to prompt healthier behaviors.[33]

- Data-rich apps, again infused with behavioral nudge design, designed to help gig workers save small amounts of money each day to better achieve their financial goals.[34]

Interestingly, while non-maleficence was the third-most common principle in the AI ethics declarations studied by the ETH Zurich team, the principle of beneficence appeared in less than half of them. It is possible that this disparity reflects a prevalence of alarmist discussions of AI that focus more on harm, but dwell less on AI's potential to help debias human decisions, extend human capabilities, and improve well-being.

## Managing tradeoffs

Ethical deliberations often involve managing tradeoffs between different principles that cannot be simultaneously satisfied. Tradeoffs between beneficence and non-maleficence are common. For example, the public might be willing to accept a certain fatality rate associated with autonomous vehicles if it is lower than the fatality rate resulting from humans operating more traditional vehicles.



Sometimes, the process of articulating an ethical tradeoff can spur innovations that render the tradeoff less fraught. One government agency, for instance, commissioned a machine learning algorithm to identify people at relatively high likelihood of improperly collecting unemployment insurance (UI) benefits. For unavoidable technical reasons, any such algorithm could have yielded a large number of false positives—mistakenly flagging legitimate claims as improper.[35] If the agency had simply used the algorithm to feed an automatic decision rule of the form "If the score exceeds *x*, deny benefits," the inevitable false positives would have led the agency to deny needed UI benefits to large numbers of deserving people.

For this reason, the data science team instead designed the AI system to function as a "nudge engine." Instead of denying benefits to high-scoring individuals, the agency delivered well-timed behavioral "nudge" pop-up messages—such as "nine out of 10 of your neighbors in [your county] report their earnings accurately"—to claimants the algorithm flagged as suspicious. These messages did no harm to individuals inaccurately flagged by the algorithm, but they had the desired effect among people who were in fact improperly claiming benefits. Randomized controlled trials of the system revealed that the machine learning-targeted nudge messages cut improper UI payments by approximately 50 percent.[36]

The broader point is that ethical AI requires organizations to consider not only *predictions*, but *interventions* as well.[37] Often, "classical economic" interventions such as setting prices, offering or withholding treatment, and delivering punishments or rewards are the only ones considered. The newer science of choice architecture expands the toolkit with "soft" interventions that can allow organizations to act ethically on ambiguous algorithmic indications.[38] In cases where nudge interventions aren't strong enough, ethical deliberation should help guide policy decisions about how machine-generated predictions are acted upon. For example, a certain predictive algorithm could be deployed either to deny benefits or provide proactive outreach to help at-risk cases.

A still broader point is that technological innovation, often involving multidisciplinary thinking, can also make it possible to mitigate difficult ethical tradeoffs. The increasingly popular tagline "human-centered AI" can perhaps be interpreted as a call to take human and societal needs into account when developing uses for AI technologies.[39]

## Justice: Treating people fairly

Justice is another core ethical principle that appears frequently in AI ethics declarations.[40] In the ETH Zurich analysis, it encompasses such related concepts as inclusion, equality, diversity, reversibility, redress, challenge, access and distribution, shared benefits, and shared prosperity.

Much of the conversation about justice as it relates to AI revolves around "algorithmic fairness"—the idea that AI algorithms should be fair, unbiased, and treat people equally. But what does it mean for an algorithm to be "fair"?

It is useful to distinguish between the concepts of procedural and distributive fairness. A policy (or an algorithm) is said to be *procedurally* fair if it is fair independently of the outcomes it produces. Procedural fairness is related to the legal concept of due process. A policy (or an algorithm) is said to be *distributively* fair if it produces fair outcomes. Most ethicists take a distributive view of justice, whereas a procedure's fairness rests largely on the outcomes it produces. On the other hand, studies by social psychologists and behavioral economists have shown that people often tend toward a more procedural view, in some cases caring more about being *treated* fairly than the outcomes they experience.[41]

While AI algorithms often attract criticism for being distributively unfair, many such discussions implicitly invoke procedural fairness as well. For example, some critics believe that giving female names and voices to digital assistants can reinforce societal biases.[42] Common examples point to cases where societal biases are reflected in the data sets used to train algorithms:[43] Searches for "CEO" may yield disproportionate images of white men,[44] and facial recognition systems have been shown to be less accurate when identifying individuals with darker skin.[45]

Clearly, the outputs of algorithms like these can be distributively unfair in that they could encourage biased outcomes: white males securing a

### PROCEDURAL FAIRNESS: PROMOTE FAIR TREATMENT

***Themes***

Algorithmic bias, equitable treatment, consistency

***Examples***

- Facial recognition software that recognizes dark-skinned faces just as reliably as light-skinned faces
- Internet searches that avoid amplifying implicit societal biases

### DISTRIBUTIVE FAIRNESS: PROMOTE EQUITABLE OUTCOMES

***Themes***

Shared benefits, shared prosperity, fair decision outcomes

***Examples***

- Addressing growing inequality due to technology-induced workplace changes
- Avoiding algorithmic biases leading to unfairness in hiring or parole decisions

disproportionate number of high-paying jobs, or higher autonomous vehicle accident rates among dark-skinned pedestrians due to the software's poorer performance in recognizing darker-skinned individuals.[46] But even setting outcomes aside, such algorithms may impact many people's sense of procedural fairness. For example, webcams that struggle to recognize dark-skinned faces, or internet searches for "CEO" that yield primarily male faces,[47] might be considered inherently objectionable regardless of the impacts on functionality or career progression.

Addressing such issues typically requires that the statistical methodologies used to create algorithmic systems incorporate appropriate ethical deliberation. Recall the point made above that machine learning algorithms are reliable only to the extent that they are trained on suitable data sets. For many applications, it is desirable to train an algorithm on data that reflects the way the world *is*. To accurately forecast sales, for instance, an algorithm must work with data that is representative of the population of likely customers. But what if faithfully representing the world as it is means possibly perpetuating an unfair state of affairs? In such situations, the desire for fairness may motivate the construction of training samples that reflect judgments about the way the world *ought* to be—an ethically influenced choice.

Just as data science should incorporate ethical deliberation, so should ethical deliberation be informed by careful data science. A spate of important research was prompted by a 2016 *ProPublica* investigation that revealed that a widely used recidivism algorithm had a much higher false positive rate for Black people than white people.[48] Intuitively, this difference might seem blatantly unacceptable. But if (1) the overall recidivism base rate is higher for Black people than for white people and (2) the algorithm manifests "predictive parity" in the sense that a high score means approximately the same probability of reoffending, the higher misclassification rate for Black people is a mathematical inevitability.

This result is representative of a growing body of research pointing to mathematically inevitable tradeoffs in different conceptions of algorithmic "fairness."[49] An emergent theme is that, as with impact, assessing the "fairness" of an algorithm will often involve evaluating tradeoffs rather than making a binary determination.

**Discussions of algorithmic fairness should reflect not only the shortcomings of machine predictions, but the shortcomings of human decisions as well.**

A further point is that discussions of algorithmic fairness should reflect not only the shortcomings of machine predictions, but the shortcomings of human decisions as well. The behavioral economist Sendhil Mullainathan points out that the applications in which people worry most about algorithmic bias are also the very situations in which algorithms—if properly constructed and implemented—also have the greatest potential to reduce the effects of implicit human biases.[50]

For example, hiring is a realm notorious for its susceptibility to cognitive unconscious biases that may affect who eventually gets the job. A well-known field study in the United States, co-led by Mullainathan, demonstrated that simulated resumes with Black-sounding names attracted significantly fewer interviews than comparable resumes with white-sounding names.[51] In contrast, Michael Lewis's *Moneyball* illustrates that properly constructed algorithms can outperform unaided human intuitions in predicting who is most likely to succeed on the job.[52]

Naïvely training machine learning algorithms on "convenience samples" of data can quite possibly encode and reinforce human biases reflected in the data. At the same time, Mullainathan's point implies that simply avoiding algorithms altogether can *also* be ethically problematic. Unlike human decisions, machine predictions are consistent over time, and the statistical assumptions and ethical judgments used in algorithm design can be clearly documented. Machine predictions can therefore be systematically audited, debated, and improved in ways that human decisions cannot.[53]

## Autonomy: Respecting humanity and self-determination

Put simply, autonomy is the ability of people to make their own decisions. The Stanford Encyclopedia of Philosophy provides a somewhat more expansive definition:

> Autonomy is … the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external force.[54]

Many of the principles discussed in the various AI ethics declarations, such as transparency, explainability, privacy, and dignity, can be viewed as aspects of respect for autonomy.[55]

In bioethical contexts, the autonomy principle is often invoked in the context of people's freedom to choose whether to receive medical treatments or participate in medical studies. Its applicability to AI is perhaps equally obvious. When humans employ autonomous systems, they cede, at least provisionally, some of their own autonomy (decision-making power) to machines. However, autonomous systems can provide human users with clues about when it is appropriate to cede some of their autonomy, and also give the ability to override the system at appropriate points.[56]

Handing over some portion of one's autonomy to an intelligent machine need not pose an ethical problem. In fact, doing so can sometimes be the more ethical choice. For example, the use of diagnostic decision trees (a common type of statistically derived AI algorithm) in emergency rooms can improve the accuracy of triage decisions for patients suffering chest pain. The algorithm is good at a specific kind of task that humans are generally poor at: combining risk factors in consistent and unbiased ways. In one sense, a physician who uses the algorithm gives up part of his or her autonomy—but in a deeper sense, the algorithm can actually enhance the physician's autonomy, acting as a kind of cognitive prosthesis or assistant that can help the physician achieve the goal of better treating the patient.

### AUTONOMY DOES NOT REQUIRE EXPLAINABILITY

The medical decision tree is an example of what is increasingly called "explainable AI"—AI tools whose processes and indications are understand-able, in varying degrees, by human users. Though typically less accurate than more complex algorithms, decision tree models are sometimes preferred in medical contexts because of their relative transparency and intuitive nature.[57] Explainability can be viewed through the lens of promoting human autonomy: If a diagnostic algorithm is easy to understand, a physician can make an informed judgment about when it is appropriate to let the algorithm guide the decision. Greater *comprehension* allows for more informed decision-making and the ability to choose.

### COMPREHENSION: EXPLAIN HOW TO USE AND WHEN TO TRUST AI

*Themes*

Intelligibility, transparency, trustworthiness, accountability

*Examples*

- Explainable AI algorithms helping judges or hiring managers make better decisions
- A vehicle operator understanding when to trust autopilot technology
- An AI-based tool informing decision-makers when they are being "nudged"
- A chatbot not masquerading as a real human

Unfortunately, many forms of AI, such as medical decision algorithms derived from deep learning or the algorithms used to pilot semiautonomous vehicles, do not afford similar transparency or interpretability. Providing "why" explanations to accompany black-box predictive algorithms is an ongoing area of research.[58] But today's state of the art is such that full explainability is not always a realistic goal. Yet this need not raise an ethical red flag: Explainability might not always be necessary or even desirable. In such low-stakes scenarios as product recommendations, there may be little demand for the explanations behind specific algorithmic outputs. And in high-stakes scenarios, the additional accuracy provided by complex algorithms might trump the desire for transparency and explainability. This is yet another example of an ethical tradeoff that should be deliberated.

In scenarios involving highly complex algorithms, the concept of *trustworthiness* might be a more useful organizing principle than explainability.[59] For example, few drivers or airline pilots fully understand the inner workings of their semiautonomous vehicles. But through a combination of training, assurances provided by safety regulation, the manufacturer's reputation for safety, and tacit knowledge acquired from using their vehicles, the user develops a working sense of the conditions under which the vehicles can be trusted to help them achieve their goal of safely getting from point A to point B. It is notable that recent examples of semiautonomous vehicle crashes have resulted from unwarranted levels of trust placed in driver assistance systems.[60] To reduce the risk of accidents, what is needed is not full explainability but rather a working sense of the conditions under which the algorithmic system should and should not be trusted.

## NUDGING IN THE SERVICE OF AUTONOMY

Most discussions of AI's impact on human autonomy focus on the type of *deliberative* decision-making that the cognitive scientist Daniel Kahneman calls "System 2" or "thinking slow":[61] diagnosing a patient, hiring a worker, releasing a defendant on bail. But AI technologies can also affect more *reflexive* "System 1" or "thinking fast" decision-making. For example, people are disproportionately likely to choose the default option, the option described in the most intuitive language, the option that comes up first in the search engine, or the option they believe similar people tend to make. Because of such innate tendencies, what Richard Thaler and Cass Sunstein call "choice architecture"—or "nudging"—can significantly influence people's decision-making.

For example, recall the state agency that, in using an AI algorithm to flag potentially fraudulent UI claims, chose to selectively "nudge" claimants toward honest behavior rather than selectively cut off benefits based on the algorithm's output. This use of behavioral nudges allowed the agency to avoid the unintentional maleficence of denying needed benefits to legitimate claimants. But nudging can also have implications for autonomy. For example, nudge interventions shouldn't mislead with false information or otherwise manipulate people to act in ways that go against their self-interest.[62] Recall the ethical imperative to avoid "manipulative or distorting external forces."

Commentators increasingly warn of the autonomy-threatening potential of AI technologies infused with behavioral design. In a recent *Scientific American* editorial, a distinguished group of scientists commented:

> Some software platforms are moving towards "persuasive computing." In the future, using sophisticated manipulation technologies, these platforms will be able to steer us through entire courses of action, be it for the execution of complex work processes or to generate free content for internet platforms, from which corporations earn billions. The trend goes from programming computers to programming people … The magic phrase is "big nudging," which is the combination of big data with nudging.[63]

The overarching—and legitimate—fear is that AI technologies can be combined with behavioral interventions to manipulate people in ways designed to promote *others'* goals. Examples include, behavioral algorithms coupled with persuasive messaging designed to prompt individuals to choose products, political candidates, privacy settings, data-sharing agreements, or gig work offers that they might not choose if they had better self-control or access to better information.

However, the flip side is that nudging ethically carried out can often *enhance* rather than diminish human autonomy. For example, AI algorithms can customize and target behavioral interventions that, when embedded in data-rich, digital environments, can make it easier for people to save more for retirement, engage in healthier behaviors, drive more safely, and more effectively manage time and collaborate on the job.[64] Just as a medical decision tree enhances physicians' autonomy by enabling better deliberative decisions, so too can effective choice architecture enable boundedly rational individuals to better achieve their goals through improved reflexive or habitual decisions. In each case, the AI is autonomy-enhancing.

Furthermore, the choice architecture pioneer Cass Sunstein points out that in many situations, *denying* people the benefits of smart choice architecture can in fact *undermine* their autonomy. For example, when tasked with navigating a complex set of health or employee benefit choices, an algorithm might be used to highlight an appropriate default choice (with the full menu of choices a click away). Avoiding such choice architecture might force the individual to spend a great deal more time researching and deliberating this decision, potentially impairing his or her ability to pursue other goals that he or she deems more important. In such a case, Sunstein would say that people should be given the option to "choose not to choose."[65]

**Cass Sunstein points out that in many situations, *denying* people the benefits of smart choice architecture can in fact *undermine* their autonomy.**

The central issue in these considerations appears to be *control*. Respecting individual autonomy requires that people have the freedom to make their own choices—including, paradoxically, the freedom to choose to be "nudged" or guided in ways that they believe enhance their well-being.

### CONTROL: ALLOW PEOPLE TO MODIFY OR OVERRIDE AI WHEN APPROPRIATE

*Themes*

Consent, choice, enhancing human agency and self-determination, reversibility of machine autonomy

*Examples*

- Decision-makers (such as a vehicle operator) can override an algorithm that is clearly going astray
- Choice architecture enables access to the full menu of choices if the algorithmically generated default isn't acceptable

Once again, the concept of *trustworthiness* is paramount. When presenting people with a deliberately designed choice architecture, it is incumbent upon the architect to communicate to users that they are, in fact, being nudged, to give them the ability to opt out (in this case, see the full menu of choices); and, most importantly, cultivate their trust that the choice architect has designed the choice environment in ways that help them achieve their goals.

## Ethical AI by design

Ethics is often viewed as a constraint on organizations' abilities to maximize shareholder returns. But we suggest a different perspective: that ethical principles can serve as design criteria for developing innovative uses of AI that can improve well-being, reduce inequities, and help individuals better achieve their goals. In this sense, the principles of impact, justice, and autonomy can help shape AI technologies in ways that achieve what marketing, management, and design professionals, respectively, call customer-centricity, employee-centricity, and human-centricity. Developing trustworthy AI technologies that safely and fairly help advance these goals is a distinctly 21st-century way for organizations to do well by doing good.

**JAMES GUSZCZA,** Deloitte Services LP, is Deloitte's US chief data scientist. He lives in Santa Monica, CA.

**MICHELLE A. LEE,** a manager with Deloitte LLP in the United Kingdom, leads the organization's Risk Analytics offerings with a focus on advising clients on managing new risks and ethical considerations introduced to their business by artificial intelligence. She is based in London.

**BEENA AMMANATH,** a managing director with Deloitte Consulting LLP, brings her extensive global experience in artificial intelligence and digital transformation to help organizations navigate ethical issues related to AI. She is based in the San Francisco Bay area.

**DAVE KUDER,** a principal with Deloitte Consulting LLP, leads Deloitte's US Cognitive Insights & Engagement offering. He is based in Atlanta.

## Human values in the loop

*page 66*

1. Jim Guszcza, Harvey Lewis, and Peter Evans-Greenwood, *Cognitive collaboration: Why humans and computers think better together*, Deloitte Insights, January 23, 2017.

2. The field of artificial intelligence is commonly agreed upon to have originated at a 1956 conference held at Dartmouth University. The conference was convened by John McCarthy, who coined the term “artificial intelligence” and defined it as the science of creating machines “with the ability to achieve goals in the world.” Other participants included Marvin Minsky, Alan Newell, Claude Shannon, and Herbert Simon. Their goal was to create artificial *general* intelligence in the sense of simulating “every aspect of learning or any other feature of intelligence.” In contrast, the AI applications discussed here are all forms of *narrow* artificial intelligence: the ability to achieve specific goals commonly associated with human intelligence. For example, an AI capable of diagnosing a patient will not be capable of making a product recommendation. For further discussion and references, see: Jim Guszcza, “Smarter together: Why artificial intelligence needs human-centered design,” *Deloitte Review* 22, January 22, 2018.

3. Steven Levy, “‘Hackers’ and ‘information wants to be free,’” Medium, November 21, 2014.

4. Polly Sprenger, “Sun on privacy: ‘Get over it,’” *Wired*, January 26, 1999.

5. Kevin Kelly, *What Technology Wants* (New York: Penguin Random House, 2011).

6. Hemant Taneja, “The era of ‘Move fast and break things’ is over,” *Harvard Business Review*, January 22, 2019.

7. Matthew Hutson, “Even artificial intelligence can acquire biases against race and gender,” *Science*, April 13, 2017; *Economist*, “Fake news: You ain’t seen nothing yet,” July 1, 2017; Paul Lewis, “‘Our minds can be hijacked’: The tech insiders who fear a smartphone dystopia,” *Guardian*, October 6, 2017; Holly B. Shakya and Nicholas A. Christakis, “A new, more rigorous study confirms: The more you use Facebook, the worse you feel,” *Harvard Business Review*, April 10, 2017.

8. Brady McCombs, “Utah driver sues Tesla after crashing in autopilot mode,” Associated Press, September 6, 2018.

9. Bill Snyder, “Our misplaced fear of job-stealing robots,” Insights by Stanford Business, March 7, 2019.

10. Stephen A. Schwarzman, “Can we make artificial intelligence ethical?,” *Washington Post*, January 23, 2019.

11. Anna Jobin, Marcello Ienca, and Effy Vayena, *Artificial intelligence: The global landscape of ethics guidelines*, *Nature Machine Intelligence* 1, no.9 (2019): pp.389–99.

12. Luciano Floridi et al., “AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations,” *Minds and Machines* 28, no. 4 (2018): pp. 689–707. The four-principles approach was outlined in Thomas Beauchamp and James Childress’s *Principles of Biomedical Ethics*, first published in 1985. The book’s goal was to identify health care’s “common

morality." Tom L.  Beauchamp and James F. Childress, *Principles of Biomedical Ethics, Seventh Edition* (Oxford University Press, 2012).

13. Luciano Floridi and Josh Cowls, "A unified framework of five principles for AI in society," *Harvard Data Science Review* 1.1, June 23, 2019. One prominent bioethicist has stated that the four principles can be thought of as "the four moral nucleotides that constitute moral DNA—capable, alone or in combination, of explaining and justifying all the substantive and universalizable moral norms of health care ethics and, I suspect, of ethics generally." The four bioethical principles are sometimes called "mid-level" principles, lying between fundamental philosophical theories and particular, context-specific rules. One might wonder about the principles' applicability across a variety of cultural contexts. While it is impossible to do justice to this topic here, we point to the study by D. F. Tsai which concluded that the four principles "are clearly identifiable" in ancient Chinese medical ethics. R. Gillon, "Ethics needs principles—four can encompass the rest—and respect for autonomy should be 'first among equals,'" *Journal of Medical Ethics* 29, no. 5 (2003): pp. 307–12; Steven S. Coughlin, "How many principles for public health ethics?," *Open Public Health Journal* 1, no. 1 (2008): pp. 8–16; D. F. Tsai, "Ancient Chinese medical ethics and the four principles of biomedical ethics," *Journal of Medical Ethics* 25, no. 4 (1999): pp. 315–21.

14. See the chapter on ethics in Matthew J. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, 2018).

15. Stanford Encyclopedia of Philosophy, "Consequentialism," Stanford University, June 3, 2019.

16. A striking dramatization of weaponized AI was presented by the prominent AI researcher Stuart Russell in his "Slaughterbots" video. The video dramatized the hypothetical use of inexpensive drones, guided by commodity facial recognition software, to assassinate selected individuals. While the technology to do this does not yet exist, Russell was motivated to illustrate "the property of autonomous weapons to turn into weapons of mass destruction automatically because you can launch as many as you want." The video's release was timed to put pressure on diplomats attending a United Nations conventional weapons meeting in Geneva. Eric Ting, "UC Berkeley professor's eerie lethal drone video goes viral," SFGate, November 18, 2017.

17. In machine learning parlance, "labels" are what statisticians call "outcome variables" or "target variables." Labels are known quantities for cases in historical data set, but unknown for the future scenarios in which a predictive algorithm is intended to apply. For example, consider a database containing the heights of a million adult males along with the heights of each man's parents. An algorithm can be created to predict the height of an adult male from the heights of his parents. In this scenario, the height of the child is the "label," and it is an objectively measurable quantity. In other scenarios, the label is provided by humans using judgment. Examples include, labeling a bit of social media content as "offensive," an employee's performance as

"acceptable," or a tumor as "cancerous." The term "human-in-the-loop machine learning" is often used to connote the latter types of scenarios.

18. The hype and semantic confusion surrounding AI, as well as a tendency to overinterpret the "intelligence" of novel forms of AI, results in confusion even among experts. The prominent machine learning researcher Michael Jordan comments that the term "AI" is often used as an "intellectual wildcard, one that makes it difficult to reason about the scope and consequences of emerging technology … This is not the classical case of the public not understanding the scientists—here the scientists are often as befuddled as the public." Michael I. Jordan, "Artificial intelligence—the revolution hasn't happened yet," *Harvard Data Science* Review 1.1, June 23, 2019.

19. Jordan discovered this when his wife's ultrasound image resulted in a faulty diagnosis. Jordan, a trained statistician, investigated the situation and reported the differing ultrasounds to the attending geneticist, who replied, "That explains why we started seeing an uptick in Down syndrome diagnoses a few years ago. That's when the new machine arrived." As a result, Jordan's wife avoided the risky amniocentesis procedure.

20. For further discussion, see: David Danks and Alex John London, "Regulating autonomous systems: Beyond standards," *IEEE Intelligent Systems* 32, no. 1 (2017): pp. 88–91.

21. The *Economist* reports that some Chinese firms are adopting this strategy: "In the absence of driving software which can handle chaotic city streets, some Chinese firms are … turning the streets themselves into something that software can handle. The approach involves installing sensors to guide cars, writing and enforcing rules about how humans move around, designing (or redesigning) urban landscapes to be AV-friendly and, critically, limiting AV firms' legal liability in the event of inevitable accidents." *Economist*, "Chinese firms are taking a different route to driverless cars," October 12, 2019.

22. Prominent examples of chatbots being gamed to spread inflammatory content come to mind, but controls for more mundane scenarios should also be considered. For example, some chatbots incorporate personally identifying information (PII) detectors that can either block the PII content or ask the user to confirm its appropriateness.

23. Robert Hof, "Toyota: 'Guardian angel' cars will beat self-driving cars," *Forbes*, April 8, 2016.

24. For a discussion of human-computer symbiosis, see Guszcza, Lewis, and Evans-Greenwood, *Cognitive collaboration*. A recent exploration of human-computer extended intelligence is Thomas Malone's book *Superminds*. For a summary, see: Jim Guszcza and Jeff Schwartz, "Superminds: How humans and machines can work together," *Deloitte Review* 24, January 28, 2019.

25. The prominent computer scientist Kris Hammond states: "Any program can be considered AI if it does something that we would normally think of as intelligent in humans … How the program does it is not the issue, just that it is able to do it at all. That is, it is AI if it is smart, but it doesn't have to be smart like us." Kris Hammond, "What is artificial intelligence?," *Computerworld*, April 10, 2015.

26. For example, on October 26, 2019, one of this article's authors used a popular app to translate the current headline "Elizabeth Warren is becoming Trump's greatest threat" from English into Burmese, and then from Burmese back into English. The app returned: "Becomes the biggest threat of Elizabeth Warren." When Spanish was substituted for Burmese, however, the translations were flawless. The success of the algorithm is a function of the size and adequacy of the available data, not the sort of understanding characteristic of human speakers.

27. Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Pantheon, 2019).

28. *Economist*, "Driverless cars are stuck in a jam," October 10, 2019; Christopher Mims, "Driverless hype collides with merciless reality," *Wall Street Journal,* September 13, 2018.

29. For an engaging profile of the affective computing pioneers Rosalind Picard and Rana el Kaliouby, see Raffi Khatchadourian, "We know how you feel," *New Yorker*, January 12, 2015. Interestingly, the same affective computing technology is increasingly used for the types of semiautonomous vehicle driver assistance functionality mentioned in the previous section. Khari Johnson, "Affectiva launches emotion tracking AI for drivers in autonomous vehicles," VentureBeat, March 21, 2018.

30. For many peer-reviewed examples, see Data Science for Social Good Summer Fellowship, "Publications from the Data Science for Social Good Fellowship program," accessed November 21, 2019.

31. Simon Hoermann et al., "Application of synchronous text-based dialogue systems in mental health interventions: Systematic review," *Journal of Medical Internet Research* 19, no. 8: (2017).

32. "Growth mindset" interventions—prompting the belief that success comes from effort and perseverance rather than fixed abilities or "gifts"—were pioneered by the respected psychologist Carol Dweck and are commonly used to improve children's educational outcomes. The social robot pioneer Cynthia Breazeal and her collaborators have built peer-like social robots designed to foster growth mindsets in children. Hae Won Park et al., "Growing growth mindset with a social robot peer," *Proc ACM SIGCHI*, (2017): pp. 137–145.

33. For example, colleagues from Deloitte recently teamed with the Penn Medicine Nudge Unit to test such an intervention. Penn Medicine News, "Using a wearable device to exercise more? Add competition to improve your results," press release, September 9, 2019; also see Mitesh S. Patel et al., "Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States," *JAMA Internal Medicine*, September 9, 2019.

34. For example, the prominent behavioral economist Shlomo Benartzi, one of the creators of the celebrated "Save More Tomorrow" behavioral finance program, has worked on such apps. Banartzi coauthored the book *Save More Tomorrow* with his collaborator, Nobel Laureate Richard Thaler. Benartzi's more recent book *The Smarter Screen* is about applying nudge

principles in digital environments to reach larger populations (such as gig workers) and test interventions more rapidly.

35. A false positive is an error resulting from a test or algorithm indicating the presence of a condition (for instance, being a fraudster, having a rare disease, or being a terrorist) that does not in fact exist. If the overall population-level base rate is low, then even the most sophisticated algorithms often yield more false positives than true positives. This is known as the "false positives paradox." To illustrate, suppose that each year a country faces only a small handful of commercial airline terrorists threats, and that the best available algorithm homes in on a few hundred suspects out of millions of passengers. Though tiny relative to the overall population, the great majority of people on this list will be innocent. Furthermore, because no algorithm is perfectly accurate, it is quite possible that this list won't contain all of the actual terrorists, a type of error called a false negative. The tradeoff is that, expanding the list of suspects to reduce the likelihood of false negatives will increase the number of false positives—and therefore the risk of harming or treating unfairly still more innocent people. Analogous scenarios involve selecting algorithmic thresholds for deciding when to treat people at risk of a disease. There is generally a tradeoff between correctly identifying as many people with the disease as possible versus avoiding potentially risky treatments of healthy people. In such cases, applying the principles of beneficence and nonmaleficence requires cost-benefit judgments that might vary across individuals, organizations, and societies. Wikipedia, "Base rate fallacy," accessed November 21, 2019.

36. For Pew Charitable Trusts' discussion of this case study, see The Pew Charitable Trusts, "Behavioral analytics help save unemployment insurance funds: New Mexico uses data to identify misinformation, save money," October 26, 2016. To see a short video describing the case study, see Deloitte US, "Nudging New Mexico: Kindling compliance among unemployment claimants," YouTube video, 4:33, February 11, 2016.

37. For a related discussion, see: Chelsea Barabas et al., "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment," *Proceedings of Machine Learning Research* 81 (2018): pp. 1-15.

38. For an extended discussion, see Jim Guszcza, "The last-mile problem: How data science and behavioral science can work together," *Deloitte Review* 16, January 27, 2015.

39. For an extended discussion of this theme in the context of behavioral data, see Jim Guszcza, David Schweidel, and Shantanu Dutta, "The personalized and the personal: Socially responsible innovation through big data," *Deloitte Review* 14, January 18, 2014. For a discussion of human-centered AI, see: Fei-Fei Li, "How to make AI that's good for people," *New York Times*, March 7, 2018. For further perspectives on human-centered AI, see Jim Guszcza, "Smarter together: Why artificial intelligence needs human-centered design," *Deloitte Review* 22, January 22, 2018.

40. "Justice & Fairness" is the second-most frequently appearing concept in the corpus of declarations studied by the ETH Zurich team, appearing in 68 of the 84 declarations. This is second to "Transparency," which appeared in 73 of the declarations studied, and just ahead

of “Non-maleficence,” which appeared in 60. It is interesting that these top three categories roughly correspond to the three of the four core principles of bioethics. In this article, we discuss transparency as an aspect of the principle of *autonomy*.

41. Stanford Encyclopedia of Philosophy, “Procedural versus substantive justice,” June 26, 2017.

42. Sigal Samuel, “Alexa, are you making me sexist?,” Vox, June 12, 2019.

43. Daniel Cossins, “Discriminating algorithms: 5 times AI showed prejudice,” *New Scientist*, April 12, 2018.

44. Jennifer Langston, “Who’s a CEO? Google image results can shift gender biases,” UW News, University of Washington, April 9, 2015.

45. Tom Simonite, “The best algorithms struggle to recognize black faces equally,” *Wired*, July 22, 2019.

46. Sigal Samuel, “A new study finds a potential risk with self-driving cars: Failure to detect dark-skinned pedestrians,” Vox, March 6, 2019.

47. Jeff Green, Jordyn Holman, and Janet Paskin, “America’s C-suites keep getting whiter (and more male, too),” Bloomberg, September 21, 2018.

48. Julia Angwin et al., “Machine bias,” ProPublica, May 23, 2016.

49. A foundational paper is Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, “Inherent trade-offs in the fair determination of risk scores,” November 17, 2016; for an intuitive discussion, see Sam Corbett-Davies et al., “A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.,” *Washington Post*, October 17, 2016.

50. *Chicago Booth Review*, “Sendhil Mullainathan says AI can counter human biases,” August 7, 2019.

51. Marianne Bertrand and Sendhil Mullainathan, “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination,” National Bureau of Economic Research, accessed November 21, 2019. Further discussion is provided in Jim Guszcza, Josh Bersin, and Jeff Schwartz, “HR for humans: How behavioral economics can reinvent HR,” *Deloitte Review* 18, January 25, 2016.

52. See Richard H. Thaler and Cass R. Sunstein, “Who’s on first,” University of Chicago Law School, September 1, 2003. See also Jim Guszcza, “The importance of misbehaving: A conversation with Richard Thaler,” *Deloitte Review* 18, January 25, 2016. The ability of even simple algorithms to outperform unaided expert judgment gave rise to the decades-long “Actuarial versus clinical judgment” initiative by Daniel Kahneman’s predecessor Paul Meehl. For a discussion of this phenomenon in the context of AI, see Guszcza, Lewis, and Evans-Greenwood, “Cognitive collaboration.”

53. For a discussion of the need for professionalizing a data science subprofession of algorithm auditors, see James Guszcza et al., “Why we need to audit algorithms,” *Harvard Business Review*, November 28, 2018.

54. Stanford Encyclopedia of Philosophy, “Autonomy in moral and political philosophy,” January 9, 2015.

55. Regarding privacy in particular, Frederike Kaltheuner from the civil rights organization Privacy International states: “People want to negotiate who they are and how they want to interact with the world around them. Privacy is about enabling all of this and empowering individuals to do this all. Framed like this, privacy isn’t the opposite of *connecting* and *sharing*—it’s fundamentally about human dignity and autonomy.” Privacy International, “It’s about human dignity and autonomy,” July 12, 2018.

56. Floridi and Cowls, “A unified framework of five principles for AI in society.” Though fundamental, autonomy appears in only four of the six declarations studied by the AI4People team, and 34 of the 84 declarations studied by the ETH Zurich team. Echoing Salganik’s point made above, this illustrates the benefit of starting with fundamental principles.

57. See, for example Vili Podgorelec, et al., “Decision trees: An overview and their use in medicine,” *Journal of Medical Systems* 26, no. 5 (2002): pp. 445–63.

58. Dr. Matt Turek, “Explainable artificial intelligence (XAI),” Defense Advanced Research Projects Agency, accessed November 21, 2019.

59. The philosopher David Danks is especially helpful on the concept of trust. For Danks, a starting point for trust involves the user having “a reasonable belief that the system (whether human or machine) will behave approximately as intended.” David Danks, “The value of trustworthy AI,” proceedings of the 2019 AAAI/ACM conference, January 27–28, 2019. In Deloitte’s 2020 *Tech Trends*, Danks states: “To me, trust is a willingness to make yourself vulnerable because you expect the broader system to act in ways that support your values and interests. That doesn’t mean that you expect the company will never make a mistake or experience an unintended outcome. Instead, what’s important is that if something goes wrong, you’re confident that the company will take care of it.” Deloitte, *Tech Trends 2020*, forthcoming. The ethical philosopher Onora O’Neill discusses the importance of linking trust to the inherent trustworthiness of the agent. She comments, “To place and refuse trust intelligently, we must link trust to trustworthiness, and must focus on evidence of honesty, competence, and reliability.” Onora O'Neill, “Linking trust to trustworthiness,” *International Journal of Philosophical Studies* 26, no. 2 (2018): pp. 293–300.

60. An Insurance Institute of Highway Safety article states that one such crash “demonstrates the operational limits of advanced driver assistance systems and the perils of trusting them to do all of the driving, even though they can’t.” Insurance Institute for Highway Safety (IIHS), “Fatal Tesla crash highlights risk of partial automation,” August 7, 2018.

61. Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2013).

62. Sunstein and Thaler state that when third parties are not at risk and the welfare of choosers is all that’s involved, the central objective of nudging is to “influence choices in a way that will make choosers better off, as *judged by themselves*.” Cass Sunstein, “The ethics of nudging,” *Yale Journal on Regulation* 32, no. 2 (2015), pp. 413–50.

63. Dirk Helbing et al., “Will democracy survive big data and artificial intelligence?,” *Scientific American*, February 25, 2017.

64. Acorns, “Home page,” accessed November 18, 2019; Greg Szwartz, “STEP UP: When paired with gamification and a science-based engagement program, wearables can boost (and sustain) activity levels,” Deloitte Life Sciences & Health Care blog, September 9, 2019; Katie Burke, “Read my lips: How AI enables safe driving,” *Forbes*, June 27, 2018; Humu, “Make work better,” accessed November 18, 2019.

65. Cass R. Sunstein, “Choosing not to choose,” *Duke Law Journal* 65, no. 1 (2014).

Deloitte.
Insights

Follow @DeloitteInsight #DeloitteReview

www.linkedin.com/company/deloitte-insights

GET IT ON Google Play Download on the App Store Go straight to smart.
Download the Deloitte Insights app

Subscribe to receive email updates at www.deloitte.com/insights