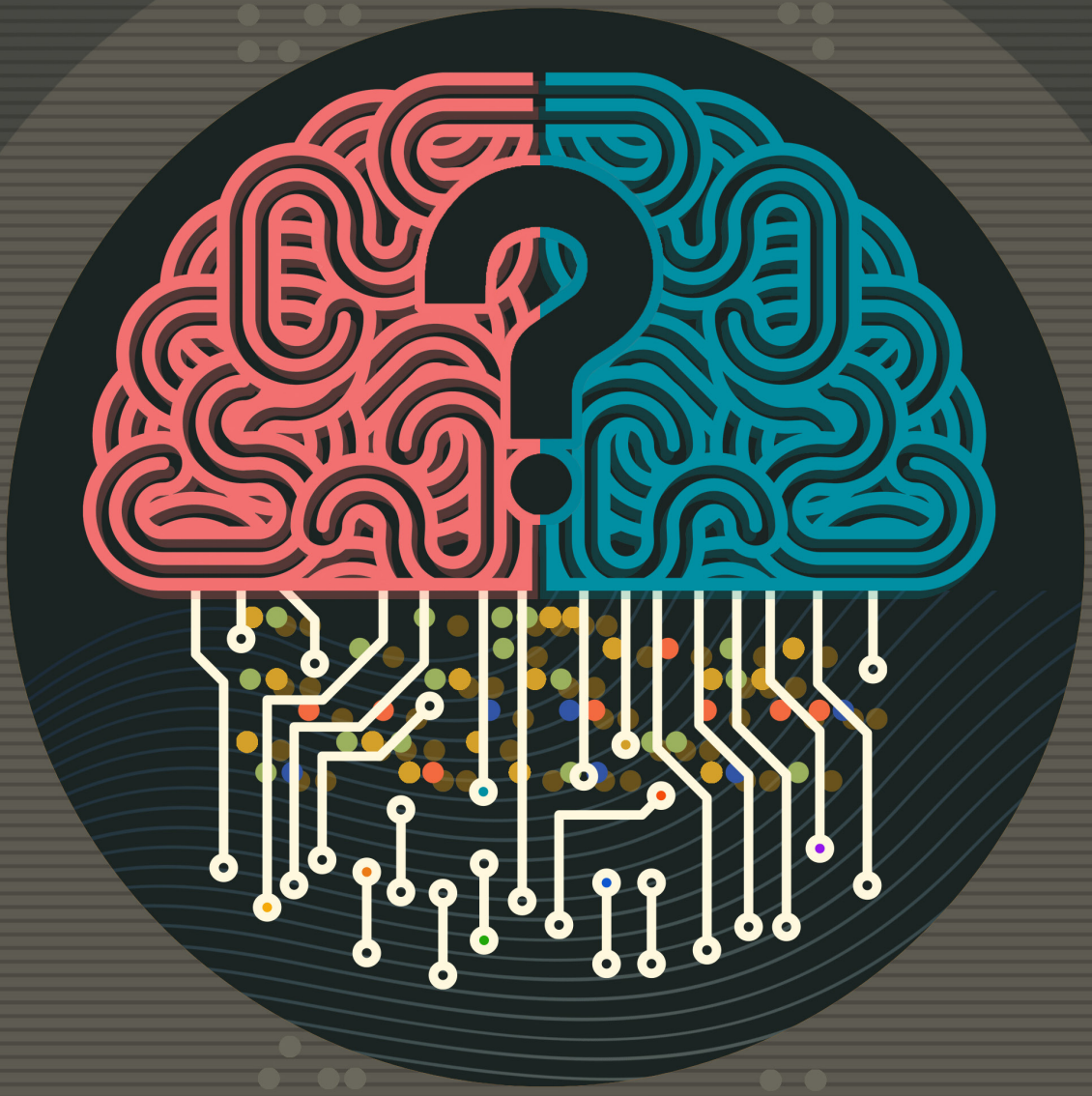
Deloitte. Insights

A report from the Deloitte Australia Centre for the Edge and the Commonwealth Scientific and Industrial Research Organisation

CSIRO DATA 61



A moral license for AI

Ethics as a dialogue between firms and communities

About the Deloitte Australia Centre for the Edge

The Deloitte Australia Centre for the Edge, which is part of our global research network, is designed to help businesses profit from emerging technology opportunities. Technology developments strike at the very heart of a business, and finding strategies for corporate growth is essential. Our mission is to identify and explore opportunities that aren't yet on senior management's agenda but should be. While we're focused on long-term trends and opportunities, we also look at the implications for near-term action.

As a client, you can benefit not only from the Centre's international research, but also from the Australian chapter's research on issues affecting the local operating environment.

About the Commonwealth Scientific and Industrial Research Organisation | Data61

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) is Australia's national science research agency. At CSIRO, we solve the greatest challenges using innovative science and technology. We shape the future. We do this by using science to solve real issues to unlock a better future for our community, our economy, our planet.

Our world is changing, fast, and data is the basic currency of this new world. CSIRO's Data61 is Australia's leading digital research network. We're here to help you create your data-driven future.

Deloitte Australia's Analytics and Cognitive practice

When you harness the power of analytics, automation, and artificial intelligence (AI), you can uncover hidden relationships from vast amounts of data. Implementing the right strategy and technology will balance speed, cost, and quality to deliver measurable business value.

Contents

AI and ethics: A question of social license?	2
Framing the challenge	4
Trust and acceptance	9
The missing parts	15
The need for a moral license for AI	22
Endnotes	23

AI and ethics: A question of social license?

"My company spends US$7 million per year on community programs. We still face work interruptions from the communities we help. Obviously, the money does not buy us the goodwill we need, but I have no idea where we are missing the point."[1]

— ***Managing director of an oil company***

THREE FRIENDS WERE having morning tea on a farm in the Northern Rivers region in New South Wales (NSW), Australia, when they noticed a drilling rig setting up in a neighbor's property on the opposite side of the valley. They had never heard of the coal seam gas (CSG) industry, nor had they previously considered activism. That drilling rig, however, was enough to push them into action. The group soon became instrumental in establishing the anti-CSG movement, a movement whose activism resulted in the NSW government suspending gas exploration licenses in the area in 2014.[2] By 2015, the government had bought back a petroleum exploration license covering 500,000 hectares across the region.[3]

Mining companies, like companies in many industries, have been struggling with the difference between having a *legal* license to operate and a *moral*[4] one. The colloquial version of this is the distinction between what one *could* do and what one *should* do—just because something is technically possible and economically feasible doesn't mean that the people it affects will find it morally acceptable. Without the acceptance of the community, firms find themselves dealing with "never-ending demands" from "local troublemakers" hearing that "the company has done nothing for us"—all resulting in costs, financial and nonfinancial,[5] that weigh projects down. A company can have the best intentions, investing in (what it thought were) all the right things, and still experience opposition from within the community. It may work to understand local mores and invest in the community's social infrastructure—improving access to health care and education, upgrading roads and electricity services, and fostering economic activity in the region resulting in bustling local businesses and a healthy employment market—to no avail.

Without the community's acceptance, without a moral license, the mining companies in NSW found themselves struggling. This moral license is commonly called a *social license*, a phrase coined in the '90s, and represents the ongoing acceptance and approval of a mining development by a local community. Since then, it has become increasingly recognized within the mining industry that firms must work with local communities to obtain, and then maintain, a *social license to operate* (SLO).[6] The concept of a social license to operate has developed over time and been adopted by a range

of industries that affect the physical environment they operate in, such as logging or pulp and paper mills.

What has any of this to do with artificial intelligence (AI)? While AI may seem a long way from mining, logging, and paper production, organizations working with AI (which, these days, seems to be most firms) are finding that the technology's use raises similar challenges around its acceptance by, and impact on, society. No matter how carefully an AI solution is designed, or how extensive user group testing has been, unveiling a solution to the public results in a wide range of reactions. A Bluetooth-enabled tampon can be greeted with both acclaim and condemnation, with some seeing the solution as a boon that will help them avoid embarrassment and health problems while others see privacy and safety concerns or worry about the device being hacked, leaking personal information.[7] Higher-stakes solutions result in more impassioned reactions, as has been the case with COMPAS,[8] a tool for estimating a defendant's risk of recidivism (or reoffending) in a criminal trial,[9] and MiDAS, a solution intended to detect fraud and then automatically charge people with misrepresentation and demand repayment.[10] These solutions are considered biased against less privileged groups, exacerbating structural inequalities in society and institutionalizing this disadvantage. Just as with building an oil rig, the fact that an AI solution is legally and economically feasible doesn't imply that the community will find it morally or ethically acceptable, even if they stand to personally benefit.

AI, like all technology, can benefit as well as harm both individuals and society as a whole. How we use technology—how we transform it from an idea into a solution—determines whether potential benefits outweigh harms. "Technology is neither good nor bad; nor is it neutral."[11] It is how we use technology that matters, for what ends, and by what means is it employed, as both require contemplation. There are choices to be made and compromises to be struck to ensure that the benefits are realized while minimizing, or suitably managing, the problems. Forgoing a technology due to potential problems might not be the most desirable option, though, as a "good enough" solution in an (already) imperfect world might, on balance, be preferable to the imperfect world on its own. The question is, however, what is "good enough"?

The challenge, then, is to discover what we *should* do. How do we identify these opportunities? What processes might be used to make compromises? And how can we ensure that the diverse voices in the community have their concerns listened to and accounted for?

Framing the challenge

AI ENABLES SOLUTIONS as diverse as machine translation, self-driving cars, voice assistants, character and handwriting recognition, ad targeting, product recommendations, music recognition, and facial recognition. AI is being used to instruct, advise, report measurements, provide information and analysis, report on work performed, report on its own state, run simulations, and render virtual environments.[12] Solutions that seemed impossible a few years ago are now embedded in products and services we use every day.

Over this time, our view of AI has also changed. Hopes that AI-powered solutions would counter some of our human weaknesses have given way to fears that AI might be an existential threat. At first, it was thought that regulation could control how AI is used[13]—open letters were sent to government with long lists of signatories attached, asking for regulation to be enacted.[14] This has failed to bear fruit. More recently, the focus has been on developing ethical principles to guide the development of AI-enabled solutions. These principles are useful distillations of what we want from AI (and what we'd like to avoid), but they are not enough,[15] as they fall short of describing how particular solutions should adhere to them.[16] The latest hope is that design (and design methodologies) will enable us to apply these principles, but it's not clear that design will be enough either.

Our efforts to grapple with the challenge of realizing AI's value while minimizing problems have been complicated by three challenges:

- The definitional challenge of understanding what exactly AI is, and therefore, what the problems are
- The challenge of aligning technical (AI) solutions with social norms
- The challenge of bridging different social worlds[17]—the different cultural segments of society that shape how their members understand and think about the world

We'll deal with each of these in turn.

The definitional challenge: What is AI, and what are the problems?

There is no widely agreed-upon and precise definition of what AI is and what it isn't. This is in part because AI is a broad church, home to a range of otherwise unrelated technologies. A useful working definition is:

> **"Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."[18]**
>
> — ***Nils J. Nilsson***

While imprecise, this definition does capture the huge scope and ambition of what we might call the AI project. The lack of a precise definition might also have helped the field grow, as it has enabled AI to be something of a bowerbird,[19] with its practitioners "borrowing" ideas and techniques from other fields in pursuit of their goals.[20] A more cynical approach might be to define AI as "things that don't quite work yet,"[21] as many technologies stop being seen as AI once they are broadly adopted. Roboticist Rodney Brooks[22] once complained: "Every time we figure out a piece of it, it stops being magical; we say, 'Oh, that's just a computation.'"[23] There is a sense that AI is a label for the (currently) impossible.

More pragmatic would be to consider AI as an area of practice, a community working to replicate human cognitive (rather than just physical) achievements. AI technology is simply whatever technology the AI community uses to solve problems that they find interesting. AI can progress by applying old techniques to solve new problems just as much as it can by discovering new techniques to solve old problems. Indeed, a significant driver for the current wave of investment we're seeing in AI is a confluence of cloud services, easy access to data, and low-cost, ubiquitous compute and networks enabling new solutions to be built from old technologies, rather than the development of new disruptive technologies per se.[24] After several decades of steady progress, it seems that discovery of new AI techniques might be stalling.[25]

Regardless of where one draws the line between "intelligent" technologies and others, the growing concern for ethical AI is not due to *new* technology—such as, for instance, the development of CRISPR[26] or genetically modified organisms (GMOs)—that enables us to do new and unprecedented things. The concern is due to dramatic reductions in price-performance that enable existing technologies to be applied in a broad range of new contexts. The ethical challenges presented by AI are not due to some unique capability of the technology, but to the ability to easily and cheaply deploy the technology at scale. It is the scale of this deployment that is disruptive. As technology historian Melvin Kranzberg puts it:

"Many of our technology-related problems arise because of the unforeseen consequences when apparently benign technologies are employed on a massive scale.Hence, many technical applications that seemed a boon to mankind when first introduced became threats when their use became widespread."[27]

Thanks to the growing scale of AI deployment, society seems to be at a tipping point: a transition from a world containing some automated decisions to a world dominated by automated decisions.[28] Society is formalizing decisions in algorithms, cementing them in software to automate them, and then connecting these decisions to each other and the operational solutions surrounding them.[29] Where previously the digital landscape consisted of the isolated islands of enterprise applications and personal computing, the landscape today is one of always online, available, and interconnected cloud solutions and smartphones.

The technology used to automate decisions is less important than the volume of decisions being automated and the impact of connecting these automated decisions so that they affect each other.

We're also integrating these automated decisions with hardware that can affect the real world. And we're doing this at scale, creating a landscape dominated by overlapping *decisioning networks*.[30] It's not that the individual decisions being automated are necessarily problematic on their own (though they may be, and we need guardrails to help ensure that this isn't the case). Rather, problematic behavior often emerges when automated decisions are integrated and affect each other directly, something we might consider *distributed stupidity*[31]—situations where emergent unintended consequences and clashes between automated decisions result in "smart" systems going bad.

A car rental firm, for example, might integrate the end-to-end rental process, from payments through to provisioning, reaching all the way into individual rental cars by using Internet of Things[32] (IoT) sensors and effectors.[33] This could enable the firm to track car location and provide more tailored rental plans and support renters on the road, while also reducing theft by immobilizing (stationary)[34] cars should they be stolen. However, these systems might lead the firm to inadvertently immobilize a long-term rental car while the renters are camping in a remote location with intermittent (at best) mobile phone coverage, believing the car to be stolen due to a temporary fault with a payment gateway that was progressively escalated by a series of automated decisions when the firm was unable to contact the renters via SMS or an outbound call center. The renters in this case would be left without a functioning vehicle in an isolated location and with limited resources, unable to walk out or contact help.

The point is that a bad (automated) decision can now have a cascading series of knock-on effects, triggering further bad decisions that escalate the problem.[35] The unforeseen consequences Kransberg warns of might well, in such instances, be the result of unintended interactions between previously manual decisions that have been automated and then integrated. These interactions could be highly contingent, as with the rental car example. They can also be prosaic, such as mistakenly adding a name to the list of redundancies after a merger, which could force a firm to terminate and then rehire an employee.[36] Integrating payroll with operational and access control systems streamlines internal processes, but it also creates a network of automated decisions that, once started, the firm no longer controls.

This is a difference in degree, not type, with the low (and dropping) cost of technology shifting the question from *can we* to *should we*. We need to consider the four "ares":[37] *Are we doing the right things? Are we doing them the right way? Are we getting them done well?* and *Are we getting the benefits?* The dual edge here is that because the cost to deploy and integrate these automated decisions is low and dropping, governance and oversight are also lowered, while issues concerning privacy, persuasion, and consent come to the fore.

We need to focus on the system, rather than the technology, as it's systems in use that concern us, not technology as imagined.

Aligning technical solutions with social norms

Our second challenge—the problem of aligning technical (AI) solutions with social norms—is one of not seeing the wood for the trees. The technical community, by nature of its analytical approach, focuses on details. The problem of creating an autonomous car becomes the problem of defining how the car should behave in different contexts: what to do when approaching a red light, when a pedestrian stumbles in front of the car, and so on. Designing "correct" car behavior is a question of identifying enough different contexts—different behavioral scenarios—and then crafting appropriate responses for each situation. Similarly, creating an unbiased facial recognition algorithm is seen as a question of ensuring that the set of behavioral scenarios (and responses) used to design the algorithm is suitably unbiased, trained on a demographically balanced set

of images rather than relying on historical (and potentially biased) data sets.

This reductionist approach is rightly seen as problematic, as whether or not a particular response is ethical (or not) is often an "it depends" problem. For autonomous cars, this manifests in the trolley problem, a thought experiment[38] first posed in its modern form by Phillipa Foot in 1967.[39] The trolley problem proposes a dilemma where a human operator must choose whether or not to pull a lever that will change the track that a trolley is running down. The dilemma is that a group of people is standing on the first track, while a separate individual is on the second, so the operator is forced to choose between the group dying due to their inaction, or the individual dying due to their action. The point here is that there is no single "correct" choice; any choice made will be based on subjective values applied to particular circumstances one finds oneself in, nor can one refuse to choose.[40] Many of the scenarios identified for our autonomous car will not have obvious responses, and reasonable individuals may disagree on what the most appropriate response is for a particular scenario. Similarly, attempting to align the training set for a facial recognition system with demographics leads to the question of which group of people will determine the demographic profile to be used.

The diverse and complex real world makes slicing any problem into a sufficient number of scenarios to ensure ethical behavior a Sisyphean task.[41] There will always be another, sometimes unforeseen scenario to consider; newly defined scenarios may well be in conflict with existing ones, largely because these systems are working with human-defined (socially determined) categories and types that are, by their nature, fluid and imprecise. Changing the operating context of a solution can also undo all the hard work put into considering scenarios, as assumptions about demographics or nature of the environment—and therefore, the applicable scenarios—might no longer hold. Autonomous cars designed in Europe, for example, can be confused by Australian wildlife.[42] Or a medical diagnosis solution might succeed in the lab but fail in the real world.[43]

The natural bias of practitioners leads them to think that "fair" or "ethical" can be defined algorithmically. This is not possible[44]—a blind spot, generally, for the technologists.

Bridging social worlds

The third and final challenge is bridging different social worlds. All of us have our own unique lived experience, an individual history that has shaped who we are and how we approach the world and society. The generation that came of age in the Great Depression during the 1930s is a case in point: Failing banks during that time took countless individuals' life savings with them, generating a lifelong distrust of banks among many people.

Disagreements in society are typically framed as differences in values or principles, differences in how we evaluate what we see around us. However, some of society's deepest and most intractable disputes are not primarily about values and principles. Indeed, we can often agree on principles. The differences lie in the social worlds to which we apply these values and principles: the way we interpret what we see around us.[45] We might agree with the principle that "it's wrong to [unjustly] kill people," for example, while disagreeing on what constitutes a person.[46]

Progress on these most intractable disputes is difficult, as it's common to assume that there is a single secular society (a fully normalized social world)[47] against which to measure principles such as fairness. The assumption is that everyone sees the same world as we do ourselves but just approach it with different values, when this is not necessarily the case[48]—a blind spot for many social commentators.

We can see these differences in social worlds come to the fore in some more recent and more controversial AI solutions. COMPAS, the recidivism-predicting tool mentioned earlier, is a good example. The team developing COMPAS took a utilitarian[49] approach, creating a solution for a world where all individuals are treated equally and where harms (roughly, the proportion of incorrect predictions) are minimized for the greatest number of people. If we use a different measure and judge COMPAS according to the norms of a different world, one focused on equity where all individuals experience similar outcomes in life no matter what circumstances they start under,[50] then COMPAS is lacking,[51] as the unintended harms it causes fall disproportionately on disadvantaged groups. This is the "fairness paradox,"[52] as improving COMPAS's performance in one world results in the solution performing worse in others (and vice versa).

While we agree that our AI solutions should be ethical—that they should adhere to principles such as fairness (promoting fair treatment and outcomes) and avoiding harm,[53] we can also disagree on which trade-offs are required to translate these principles into practice—how the principles are enacted. Applying the same clearly defined principle in different social worlds can result in very different outcomes, and so it's quite possible, in our open and diverse society, for different teams working from the same set of principles to create very different solutions. These differences can easily be enough for one group to consider a solution from another to be unethical.

It's common at conferences to pose the (rhetorical) question: Who decides what is ethical? Any design decision is likely to disenfranchise or otherwise affect some demographic group or fail to address existing inequalities or disadvantages, so it's implied that care must be taken to ensure that decisions are made by a suitably sensitive decision-maker. This is likely to be the wrong question, though, as focusing on *who* makes the decision means that we're ignoring *how* this individual's particular social world (which will be used to frame what is or is not ethical) was selected.[54] A better question is: How can one build a bridge between the different social worlds that a particular solution touches? There are trade-offs to be made, but without such a bridge, one cannot begin to determine how to make them.

We might summarize the challenges of developing ethical AI solutions (moral decisioning networks) as being similar to thermodynamics in that you can't win, you can't break even, and you can't leave the game.[55] We can't win, because if we choose to frame "ethical" in terms of a single social world—an assumed secular society—then we must privilege that social world over others. We can't break even, because even if we can find a middle ground, a bridge between social worlds, our technical solution will be rife with exceptions, corner cases, and problems that we might consider unethical. Nor can we leave the game, banning or regulating undesirable technologies, because what we're experiencing is a shift from a world containing isolated automated decisions to one largely defined by the networks of interacting automated decisions it contains.[56]

If we're to move beyond the current stalemate, we need to find a way to address all of these challenges: a method that enables us to address the concerns of all involved social worlds (rather than privileging one over others), that enables us to consider both the (proposed) system and the community it touches (rather than just the technology), and one that also provides us with a mechanism for managing the conflicts and uncertainty, the *ethical lapses*, that are inherent in any automated decisioning system. We need an inclusive dialogue.

Trust and acceptance

A SUCCESSFUL AI SOLUTION— a successful *automated decisioning network*—is one that not only effectively performs its intended function, but one that is accepted, approved, and ultimately trusted by the people it touches.[57] While there will be challenges, managements shouldn't find themselves dealing with "never-ending demands" from "local troublemakers," hearing that "the company has done nothing for us" while incurring costs that weigh the project down. The relationship between management and community[58] should be collaborative rather than adversarial, working together to understand when AI *should* be used. Unfortunately, we're a long way from such a state of affairs.

The concept of a *social license to operate for AI* has the potential to address all three challenges—definitional, aligning a solution with social norms, and bridging social worlds—discussed above. An SLO puts the focus on the overall solution and the social and physical environment into which it is deployed rather than on the technology, avoiding the problem of centering our method on particular AI technologies.[59] It also addresses the challenge of bridging social worlds by acknowledging that the solution can never be considered ethical per se.[60] While a firm might have the legal right to operate, it must also obtain, with the consent of the community, a moral license to operate, and this license must be maintained and renewed as both the solution and community evolve and circumstances change.[61] The ongoing process of developing and maintaining an SLO enables a firm to build a bridge between the social world of the firm and the social world of the affected community—which itself may contain multiple social worlds that also need to be bridged. The SLO process does this by providing a framework within which the firm can work with the community to understand each other, the proposed solution, and each party's goals, norms, and principles. They work together to develop a shared understanding of the proposed solution (focusing on the decisioning network rather than on the particular technologies) and then to determine how shared principles are enacted in real life—addressing the problem of aligning a solution with social norms—by identifying problems and opportunities and finding solutions. Being open to this type of dialogue means being vulnerable, because honesty is required in order for the dialogue to be open and inclusive. Products and services must be fairly represented. Stakeholders need to be willing to trust that technology has not been misrepresented or, in the event of informed consent, that data will be stored and used as promised.

Being open to this type of dialogue means being vulnerable, because honesty is required in order for the dialogue to be open and inclusive.

A case study: An intelligent hospital

Consider a case where a firm is developing a "smart" hospital. This hospital will have all the usual accouterments of a smart building: IoT sensor networks to track how inhabitants use the building—identifying patterns of room use and

individual preferences—and automation to both optimize the building's operation and tailor it to individuals, minimizing maintenance costs, reducing the building's environmental footprint, and improving convenience and comfort for its users. Floor-by-floor and zone-by-zone air quality and staff presence data will enable air conditioning and heating to be optimized, reducing power and water use while improving comfort. Data on ambient light levels and staff activity can be used to minimize lighting. Plant equipment, such as backup generators and oxygen supply lines, can be instrumented to enable just-in-time maintenance. Smartphone apps will enable inhabitants to interact with these systems and personalize their experience. And so on.

AI will be used to string these systems together, transforming our smart hospital into an "intelligent" one. Voice assistants will be ubiquitous—installed in registration (including for the emergency room), patient and treatment rooms, surgery, and so on—providing staff, patients, and their guests with a more convenient way of interacting with hospital processes, calling for help, and bridging any language barriers. Staff, patients, and visitors are tracked from when they first approach the building and associated with records maintained in operational systems—patients should never go missing again, visitors will be directed to whomever they're visiting via wayfinding, and staff can always find the nearest specialist in an emergency. Decision support tools speed diagnosis, highlighting potential problems on medical images and suggesting what a patient's particular collection of symptoms might imply. All this information is fed into AI-powered situational awareness and planning systems that identify problems (possibly before they crystallize into emergencies) and present decision-makers with both potential problems and possible solutions. A patient's mutterings, for instance, are correlated with unusual readings from bedside monitors and interpreted as advanced heart disease,[62] resulting in the situational awareness and planning system dispatching a drone crash cart while alerting support staff and the nearest specialist, and suggesting a change to the operating room schedule to accommodate a potential emergency.

While a boon, our intelligent hospital will likely suffer from many of the problems associated with a large-scale AI deployment. Voice assistants, for example, must support a range of languages, but which dialects within each language should be supported to avoid biasing the solution,[63] and how should the hospital support those who (for whatever reason) can't talk? A tool that "reads" X-ray images and highlights lung damage or other signs of pneumonia, a tool that worked well in the hospital where it was developed, might be biased against one of the demographic groups that our intelligent hospital serves, providing an undesirably high level of false negatives or positives. What should the situational awareness and planning system prioritize when confronted with conflicting needs for a scarce resource, such as a particular specialist or machine: Which patient gets priority, and should the system be empowered to make these decisions on its own?[64] There is also the possibility of unexpected interactions between these systems causing problems via emergent distributed stupidity: A voice assistant in a patient room might consistently misrecognize a patient with an uncommon dialect[65] and, exacerbated by biases in diagnosis recommendation solutions,[66] cause situation analysis to create many erroneous low-level requests that the staff soon dismiss, leading the staff to turn off decision support and so miss the patient's underlying problem before it becomes critical.[67]

Our intelligent hospital can also amplify existing discrimination, disadvantage, and privacy concerns. Flawed AI behavioral profiling derived from social media and smartphone data could, for example, influence medical risk profiles determining which treatments are offered. Data from medical devices pieced together by situation analysis—blood

oxygen, heart rate, and so on—might provide accurate prognoses that are implicitly treated by staff as do not resuscitate (DNR) decisions, decisions that might not be in a patient's best interest but represent the most efficient use of hospital resources.[68]

AI enables the hospital to take data generated by the sensor network (security cameras, for example), identify individual people, profile them, and then to discriminate[69] between them, either individually or as groups, and treat them differently. This discrimination can be a boon—allowing the firm to adjust the building or medical treatment to their needs and preferences while smoothing their journey through the day. The discrimination could also be harmful—creating undue stress by enabling the firm to track toilet breaks, generating maps of who is talking to whom and use them to identify groups unrelated to work for union-busting purposes,[70] determining what treatments are offered to a patient, or even determining which patient is treated when resources are scarce. This discrimination relies on a wealth of personal data (both captured and inferred) stored in operational systems, elevating the risks and consequences of our intelligent hospital's operational systems being hacked or leaking personal data.

From acceptance through approval to trust

To understand how a firm might go about gaining a social license, it's important to consider the major role that trust plays in this effort. The benefits of a social license to operate are the result of the community's *acceptance*[71] and *approval*[72] of a solution—the intelligent hospital in our example—and this acceptance and approval stems from the community's *trust* in the firm. If the firm is to realize the anticipated benefits of the intelligent hospital, it needs to ensure the acceptance and approval of the community that will be using it. Failure to do this is likely to result in disruptions that drive up cost and prevent the benefits from being realized. These can range from the minor (small disobediences such as sabotaging the sensors on a floor or using patterned clothing to hinder AI profiling and location tracking)[73] to the major (attempts to hack the system and render it inoperable, or protests). Unanticipated bias in voice assistants, for example, could lead to protests by affected community groups unable to engage with hospital systems. Prioritization decisions by the planning system that are not aligned with community norms, or simply surprising to many in the community, could result in the entire project being questioned.

What is important is what decisions are made, which of these decisions are automated and which are not, how these decisions affect the quality of the working and private lives of the people using the building, the effect of the decisions on the human dignity of the people they touch, and how the decisions align with community expectations.

The firm has a great deal of freedom in how AI is used to realize the intelligent hospital. While voice assistants will require some form of voice recognition technology, a range of audio and video techniques can be used to track inhabitants to similar effect. A number of different

approaches—many configurations of sensors, decisions (potentially made by AI technologies), and (consequential) actions—are possible, though only some of them will be acceptable, and even fewer may be desirable, to both the people working in and using the hospital and the firm commissioning it.

What is important is what decisions are made, which of these decisions are automated and which are not, how these decisions affect the quality of the working[74] and private lives of the people using the building, the effect of the decisions on the human dignity of the people they touch, and how the decisions align with community expectations. The firm needs a social license for the intelligent hospital. The community needs to trust the firm if it is to grant the license: trust that the firm will do (and is doing) what it says it will, and trust the firm's ability to execute and deliver on its commitments.

Ultimately, trust is a relationship of reliance. It's the belief that a counterpart will behave in certain ways, as well as the belief that the counterpart is dependable and competent, that they can be relied on. A firm that works collaboratively with the community, demonstrating integrity and competence in how it shapes the solution and manages operational risk, will likely be seen in a positive light. A firm that takes advantage of a community's vulnerabilities, is seen as cynical or incompetent, or shows poor stewardship of its own vulnerabilities, will be viewed poorly.

Trust-building enables members of the groups associated with the initiative to accept being vulnerable to one another (something many businesses may need to learn), and it also helps deescalate conflicts. Failure by the firm to meet community expectations, either for reasons beyond the firm's control or because the results of the firm's labors don't align with community expectations, erodes trust. When trust breaks down, it is often replaced by suspicion—suspicion that results in "never-ending demands" from "local troublemakers."

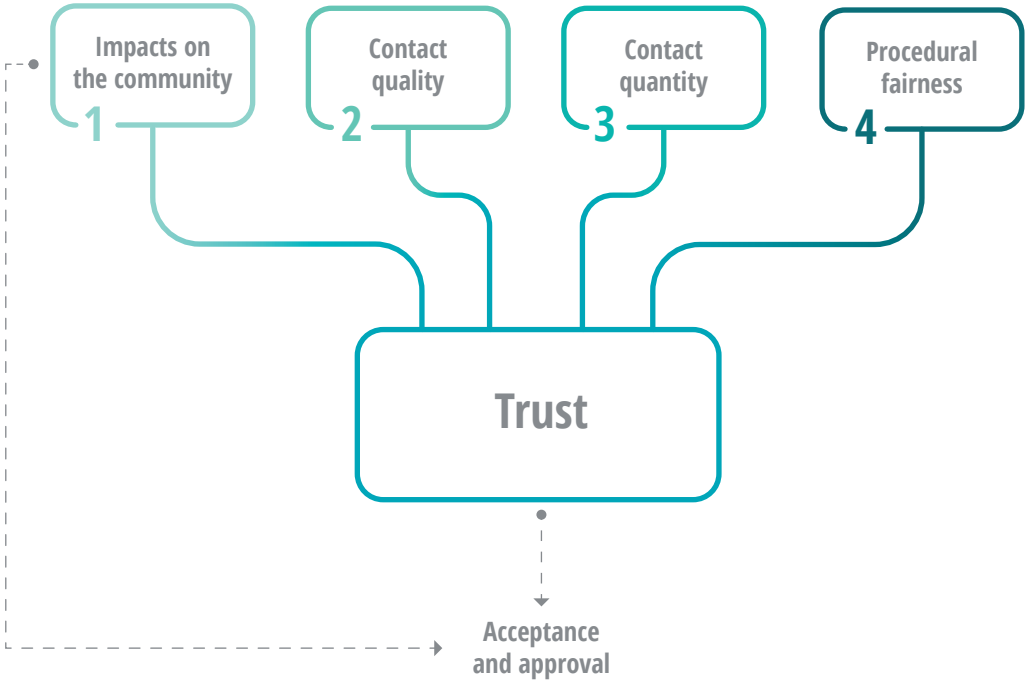
Within the context of social license to operate, trust relies on four factors: a firm's (or its solution's) *impact* on the community; the *quantity* of contact between the firm and the community; the *quality* of that contact; and the procedural *fairness* of decisions made regarding the solution (figure 1).[75] A firm can take action in all four of these areas to build trust with the community and so increase the community's acceptance and approval of its actions.

Understanding a solution's impact on the community entails recognizing that all solutions bring with them problems as well as benefits. Our intelligent hospital potentially has a smaller environmental footprint though more efficient energy use. It may facilitate more inclusive operations by enabling staff to support a broader range of languages. And diagnoses might be more accurate and swifter. However, the building also has the potential to increase work stress; introduce the privacy risks of sensitive personal data being leaked or otherwise misused; or institutionalize undesirable biases, inequalities, or disadvantage; as well as being subject to emergent distributed stupidity. But many of these benefits and problems can be anticipated by firms, enabling them to bolster benefits while mitigating problems.

It is also important to consider how the community experiences a solution, and how individuals experience it personally. For instance, integrating Bluetooth-enabled medical devices directly into the intelligent hospital's IoT network might be met with a similar response to the Bluetooth-enabled tampon discussed earlier. Or a desire to streamline operations by simplifying how staff can collaborate around an image recognition solution might not adequately address concerns about privacy and human dignity.[76] It's quite possible for different stakeholders within the community to have different expectations for a solution's benefits and problems. Similarly, an

FIGURE 1

Trust in the context of social license to operate depends on four factors



Source: Adapted from Kieren Moffat and Airong Zhang, "The paths to social licence to operate: An integrative model explaining community acceptance of mining," *Resources Policy* 39 (March 2014): pp. 61–70.

unanticipated dialect could result in frustration or even exclusion of an individual unless the speech recognition failure is dealt with gracefully. This mismatch between the firm's intention and community expectations of a solution's impact and benefits can be a significant source of the unanticipated consequences for the firm.

The distinction between the impact of a smart hospital and an intelligent one, between a hospital without and with AI, is one of degree rather than kind. AI increases the potential benefits, but it also elevates the risks.

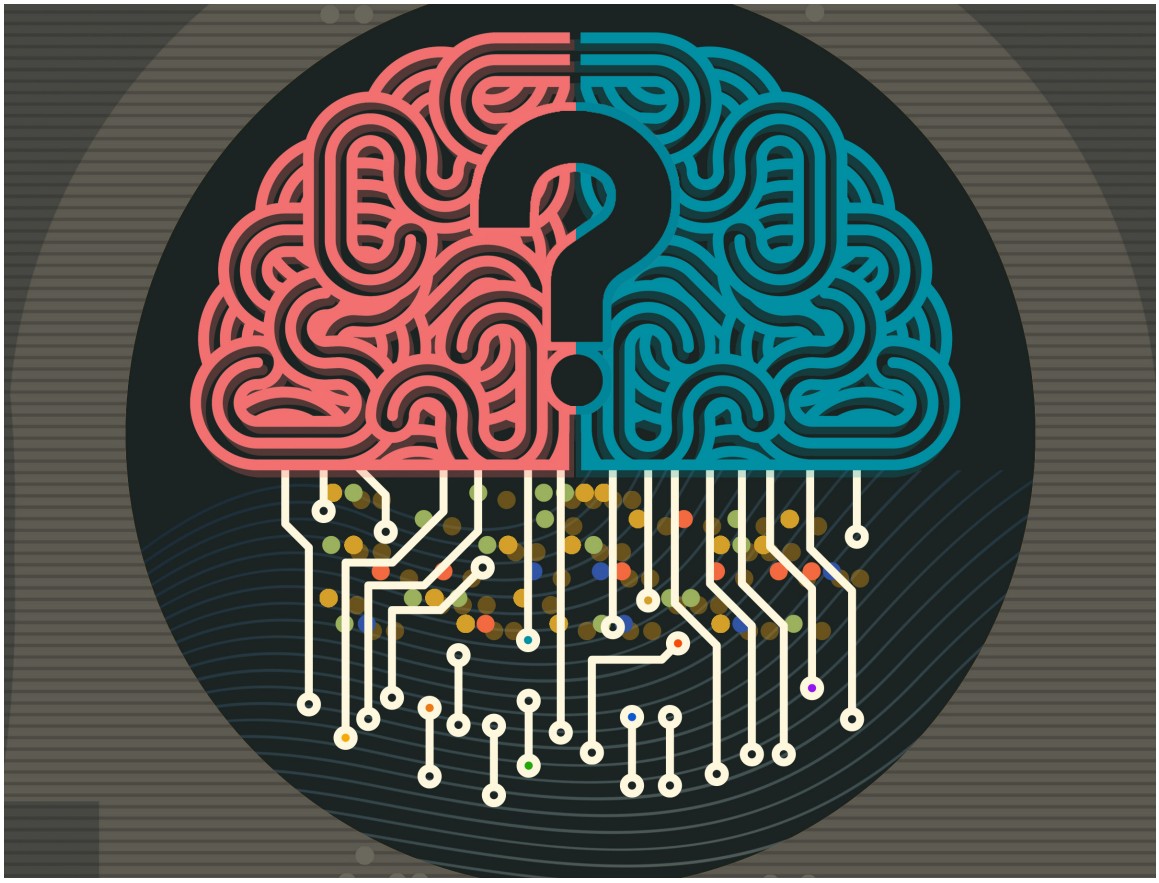
This brings us to the next two factors supporting trust: the quantity and quality of the firm's contact with the community. Trust is the result of frequent positive contact between the firm and the community. The firm that builds our hospital needs to present a human face to the community (the firm, after all, is also a community), a face that the community can learn to trust and work with.

Contact should be frequent (quantity) and meaningful (quality). Practically, contact can range from formal impact studies attempting to gauge how a solution will affect a community and their disposition toward it, to day-to-day contact in the field via community groups or between individuals and representatives of the firm,[77] as well as contact with stakeholders who are not directly affected by the solution but who have an interest in influencing the outcome.[78] Some of this contact might also be mandated via regulations such as the General Data Protection Regulation (GDPR) or those associated with the industry in which the firm is operating. Frequent, meaningful contact enables the community and the firm to learn about each other,

reducing the unknowns (and the unexpected) by minimizing misinterpretation and avoiding the projection of one's own belief systems onto the other.

TThe fourth factor influencing trust, procedural fairness, is the decision-making and dispute resolution processes that govern a solution's development and operation. Individuals must perceive that they have a reasonable voice in the decision-making process, that the decision-makers have treated them respectfully, and the procedure is one they regard as fair. They must also feel that there is equal power between parties—community and firm—so that the solution is truthful.

For the community to accept our intelligent hospital and trust the firm behind it, they need to feel that their opinions are valued, that their point of view has been accounted for, that they are being treated respectfully and with dignity, and that their view is being integrated into the solution. End-of-life care or intensive care treatment augmented by AI, for instance, needs to support patients and treat them with dignity and respect, rather than be based on an economic calculus. It should be practical for individuals and groups, for example, to respond to the proposal to use voice assistants throughout the hospital, pointing out problems and suggesting alternatives. Both decision-making and dispute processes need to be understandable and navigable by individuals so that they can see their views being accommodated and weighed against not only those of others in the community, but with technical and financial constraints and the firm's own interests.



The missing parts

THE CONCEPT OF social license to operate can provide us with a solid foundation for a moral license for AI, but work needs to be done to adapt it to the needs of firms developing AI solutions. There are three questions that we've been skirting in this article so far that we need to address if we're to move forward. These questions are:

- How do we describe the (proposed) solution without unnecessary (and confusing, for many stakeholders) technical details or reverting to overly abstract concepts?
- What constitutes "community" for our solution—that is, how do we identify our stakeholders?
- How do we evolve the solution, working from a proposed solution to one that the stakeholders consider ethical, identifying where the trade-offs are to be made and how to make them?

We can deal with these in order.

Describing the solution

The first hurdle to overcome is to find a way to describe our solution, such as the smart hospital in our example. While our familiarity with voice assistants makes them easy to understand, it is more challenging to understand a situational awareness and planning solution due to its more nebulous nature, as it requires data to be sourced from around the hospital to drive a network of interconnected decisions that provides recommendations and triggers actions for a diverse range of (potential) patient problems. We need a language that the community and the people proposing it can use to discuss the shape the AI solution will take—how inhabitant location will be tracked and what the tracking data will be used for, how it will interact with situational awareness, what actions and processes situational awareness can drive, and so on—as well as the relative problems and benefits of alternative approaches to realizing this functionality. It's AI's ability to integrate this broad range of sensors and effectors, to transform our smart hospital containing isolated automated decisions into an intelligent hospital that contains an integrated automated decisioning network, that highlights this need.

Describing our solution involves solving what we might call the brewing problem.

Describing our solution involves solving what we might call the brewing problem. Brewing required the development of microbiology—a language integrating biology and chemistry—before it could transition from craft to engineering. This made it possible to fine-tune the brewing process and obtain more consistent results. Similarly, if we're to fine-tune our AI solution, then we need to be able to describe and discuss it in a language that is accessible to both the community and the people proposing it, a language that encompasses both ethics and implementation, but without including too many technical details.[79] To be both comprehensible and useful, this language needs to

be more specific than our high-level ethical principles, but more general than implementation details. It should also avoid technical jargon, using straightforward and accessible terms to support a common understanding that contributes to building trust. We need to be able to describe the interconnected and aggregated set of decisions (the *decisions* and their *relationships*) in our proposed solution; which actor (*human* or *machine*) enacts each decision; what *information* drives the decision; the *consequences* (and information) resulting from a decision; and the *impact* of these actions (and changing information) on humans.[80]

It can be important to distinguish between decisions made by a human and those made by AI,[81] as humans and machines think (and decide) differently.[82] As humans, we use our senses and lived experience when we make a decision, even if we're making it unconsciously. We notice the unusual and unexpected and factor it into our deliberations. Machines, on the other hand, only consider that data that they're designed to consider. If a decision is consequential—such as the decision to fire a missile, withdraw an individual's social benefits, or to move a lifesaving machine to a different patient—then it is common to prefer that the decision is made by a human,[83] as only a human will consider an unusual factor, something unexpected but important enough to sway a decision. In some cases, regulation might require particular decisions to be made by a human (or even by a group) rather than algorithmically.[84] However, while we want to distinguish between human and machine decisions, we might be less interested in how the machine decision is implemented.

Our intelligent hospital might be described in terms of what information is captured, the decisions that are informed by this information, the entities that make the decisions (human or machine), and the information and actions that spring from each decision. For instance, the description may specify that a temporary identification badge issued to a visitor (*information*) will be associated with video images and a voice print (*information*) to identify the visitor (via a *machine decision*) so that the hospital can track them as they move through the building. (The technology used to associate the two is less important than the fact that the association is made.) If the building determines that the visitor wanders into a prohibited area, then it notifies (a *machine decision*) security staff on the floor who will determine what to do (a *human decision*). A complete description of a solution could contain many of these information-decision-action threads covering our intelligent hospital's operations ("Man is an animal suspended in webs of significance he himself has spun"),[85] which will be evolved and refined in collaboration with the community.

Defining the community

Before we begin any work, we need to delineate the social boundary of our system. We must establish who the stakeholders are, understand their dispositions, discover the social worlds at play, and identify our "experts," gatekeepers, and informants.[86]

"Community" may well be too narrow a term to capture the diverse set of stakeholders that a complex solution such as our intelligent hospital touches and whose lives it affects. It's easy to assume a social license to be a single license granted by a well-defined community. This is not true in complex environments, where the community is composed of a diverse collection of subgroups drawn from other geographic areas and communities. In these cases, it's more productive to think of a social license to operate as a continuum of multiple licenses across these subgroups, across multiple overlapping and interrelated communities.[87]

An anthropologist might start by listing the different behaviors, thoughts, and attitudes that should be considered, along with demographic attributes such as employment status, income, gender, primary language, and so on—factors that describe differences in the community. These factors are mapped to a set of *community factors*,[88] with each factor capturing a tension or difference in preference that might exist in the community. Obvious examples from our intelligent hospital are a worker's attitude to gender (whether gender is considered strictly binary or if a broader definition is accommodated), the nature of their work (analytical and bureaucratic or manual), their educational attainment, their religious or belief system, socioeconomic (dis)advantage, or whether they work in the hospital regularly or only visit occasionally. A complete set of factors provides us with a mud map[89] of the landscape our community might cover.

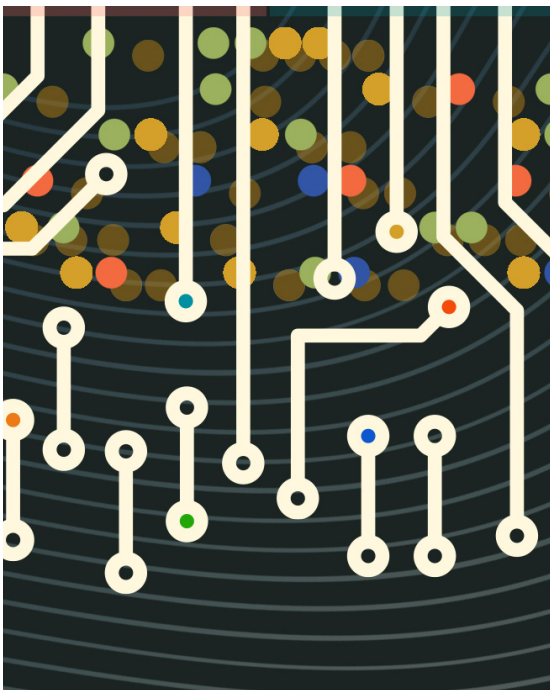
Firms can use a range of formal and informal methods to investigate community members' behaviors, thoughts, and attitudes, such as observation, structured and semi-structured interviews, group discussions, diary studies, or workshops with members of the group being studied. The important thing is to establish an open dialogue where information flows back and forth between researchers and subjects. Participants can be selected from the community to ensure that all known factors are covered, with particular attention given to edge cases. The goal is to learn as much as possible about the community's history and the individuals within it to develop a full understanding of the social worlds the community contains and how it functions.

What the firm learns can be captured in an *actor network*[90]—a web of human and nonhuman "actants," their relationships, conflicts and alliances, and the processes that bind them together—which can be used to identify a set of representative community member profiles (and representative community members) and how they might relate to the proposed solution.

Refining the solution

Our last challenge is to work with our community to refine our solution. In an approach inspired by the technique of general morphological analysis[91] (GMA), we can break this into four phases.

First, we take an idea, such as our intelligent hospital, and create a description of it. The building might use *this* data to drive *these* decisions, with *this* decision resulting in *these* actions. This is the language discussed earlier in the article, the information-decision-action threads that describe how the building will monitor visitors while in the building, support diagnosis, identify and help manage emergencies, and so on. The description can be kept general at this point by, for example, not concerning ourselves with whether a decision is made by a machine or a human.



Next, we refine our solution over two phases: eliminating the impossible, and then discovering what is allowable (and acceptable) to the community (as regulation lags behind ever-evolving social norms).

Eliminating the impossible involves enumerating all possible solution configurations—combinations of which information might feed which decision to trigger which action—and then eliminating the ones that are clearly impossible, such as those configurations that are either technically impossible or that are prevented by regulation.

Eliminating the impossible involves enumerating all possible solution configurations[92]—combinations of *which* information might feed *which* decision to trigger *which* action—and then eliminating the ones that are clearly impossible, such as those configurations that are either technically impossible or that are prevented by regulation. Regulation might require that a particular decision must be made, or supervised, by a human, leading us to add "*this* decision is performed by a human" to our solution description. Our intelligent hospital, for example, might require that any decision to transfer a lifesaving machine to a higher-priority patient is made by a human. In cases where we want the benefits of both human and machine decision-making, we might split the decision in two: a machine suggestion that can be considered as part of a human decision. The situation analysis and planning solution could be restricted to providing recommendations to a human manager who is responsible for determining the course of action. Or a decision might be required to be made by a suitably qualified person, or one with a particular level of seniority, such as a medical specialist—a requirement that is noted in the description of the decision. We might also require that a machine decision is also *understandable* by a human, noting in the machine decision's specifications that whatever technique used must provide a rationale for the decisions it makes. Our planning engine, for example, might be better implemented via rule-based constraint satisfaction[93] rather than machine learning, as this may simplify users interacting with and tweaking the solution's reasoning.

This first phase of analysis will also determine when a piece of data represents personal data (such as gender) that can only be used as an input to a few specific decisions. Based on this analysis, our description of the solution can be evolved, either by changing elements—information, decisions, and actions and their relationships—or by annotating them to restrict how each element might be used or implemented.

The next step, removing the unacceptable, is a similar process, but must be done in consultation with the community. Working with community representatives (aligned with the representative community member profiles identified earlier), a firm can identify what outcomes and processes are more or less acceptable to the community.

This phase can also explore the solution's benefit (to the community) and maturity, using a tool such as a Wardley map[94] to expose assumptions, permit challenges, and create consensus. For example, if a

particular decision is required to be "fair"—such as the choice in COMPAS between equality and equity, or the prioritization of patient needs in an emergency—then how fairness is to be enacted could be determined in collaboration with the community representatives and noted in the decision's description. Groups of related components—such as an automated registration process that integrates voice and touch interfaces with image recognition—can be reviewed to ensure the ensemble as a whole will not disadvantage or otherwise negatively affect individuals even though particular AI components are not perfect. The community's attitude to (potentially) controversial technologies can also be considered: The community may be uncomfortable with ubiquitous video surveillance, prompting our intelligent hospital's owners to find a more acceptable way to track inhabitants as they move through the building.[95] The role of situation analysis and planning might be questioned due to concerns (mentioned earlier) that accurate prognoses will be treated as implicit DNR recommendations that are not in a patient's best interests. With challenging questions such as this, the firm may need to consult with many diverse groups in the community to develop a coherent approach that is acceptable to the community as a whole.

At the conclusion of eliminating the impossible and discovering what is allowable, we have a detailed outline of our solution—though not a complete solution, as it won't have details that the firm and community do not consider pertinent. The algorithm used to maintain the temperature in a building zone will likely, for example, remain unspecified. Other details, on the other hand, might be quite tightly specified, such as the allowable uses for the video streams emanating from security cameras, how consequential recommendations from AI solutions (such as accurate prognoses) should be treated, the extent to which behavioral profiles can influence decision-making, which machine decisions are required to be understandable by a human, or how "fair" should be interpreted when dealing with conflicting patient priorities.

The processes of eliminating the impossible and discovering what is allowable enable the firm, in collaboration with the community, to determine how ethical principles (such as fairness or preventing harm) are enacted, documenting this in a shared description of the solution. We have what might be called an "ethical requirements architecture."

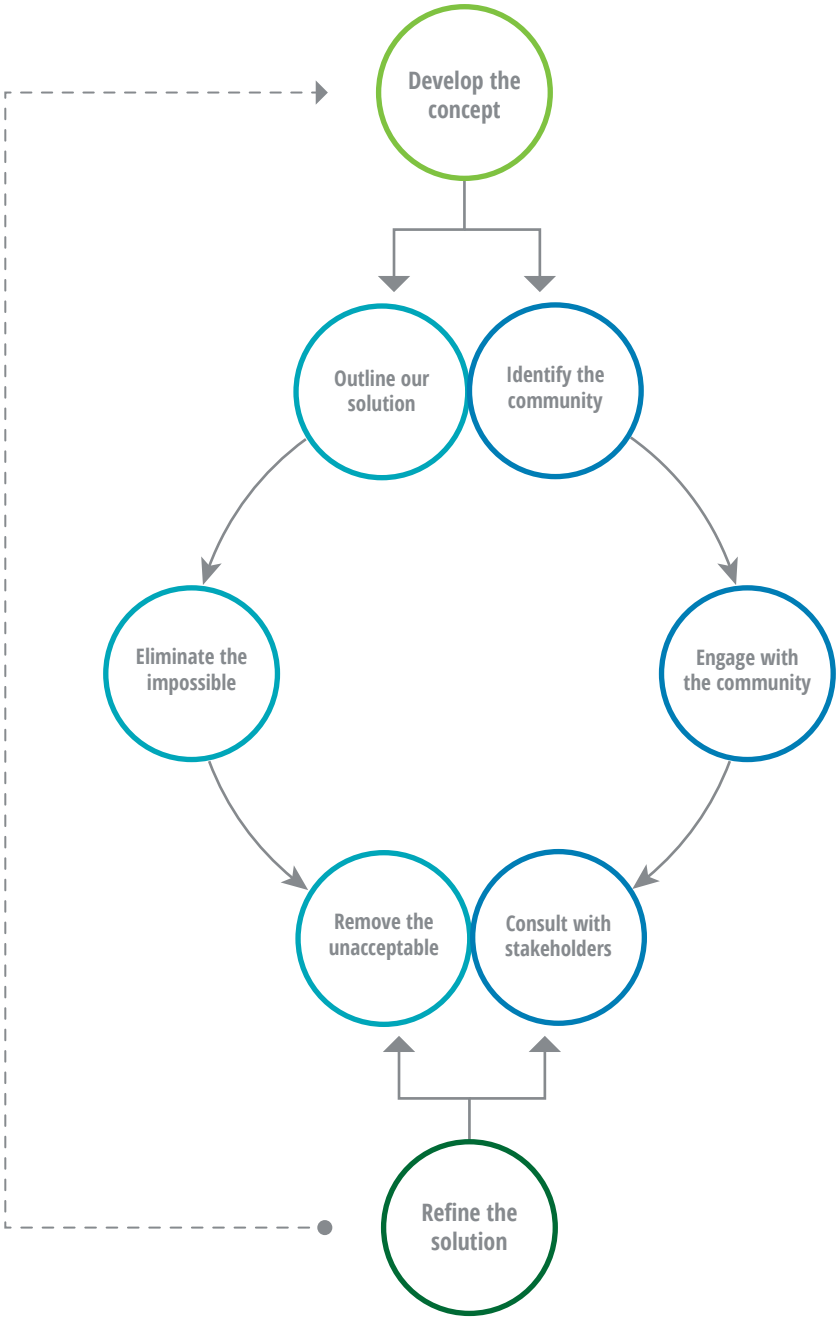
The final, fourth phase is the technical challenge of taking the refined solution description and determining how it should be realized.[96] It's in this phase that the wealth of work on methodologies and techniques to create unbiased and ethical algorithms—"trustworthy AI"[97]—is leveraged.

Developing an ethical requirements architecture

If we integrate these frameworks—social license to operate, a language for our ethical requirements, understanding the community via social science approaches, and developing and refining the solution via GMA—then we might have something similar to what is shown in figure 2. (Though we would like to note that this article has only been sufficient to develop an outline or description for this process, and not the process itself.)

FIGURE 2

A framework for vetting and refining artificial intelligence solutions with stakeholders



Source: Deloitte analysis.

The left half of the circle in figure 2 depicts the GMA process, from initial solution proposal through refinement, to create an ethical solution architecture and an eventual system. The circle's right half depicts the social sciences flow, where the firm develops a mud map of the community by identifying community factors before mapping the community terrain via an impact analysis and then engaging with the community to refine the solution.

There are two touch points between these streams. At the first, the firm develops the initial solution outline and its understanding of community factors side by side; the second comes later, when the firm works with community stakeholders to shape what an acceptable solution might look like. Cocreating the ethical solution architecture in this way addresses the possibility of the community not wanting the solution in the first place—such as our intelligent hospital's users resisting the very concept due to fears of invasive monitoring—and enables a firm to establish trust in the community by being transparent about the firm's motivations and goals. The entire process also loops back since, as we pointed out much earlier in this article, an SLO must be revisited and maintained as both the solution and the community evolve over time.

What we deem to be "good" evolves in a way that enables humans to live together in groups and ultimately create harmonious civilizations. Some morals ("thou shalt not kill," "thou shalt not steal," and so on) have been with us for millennia, while others (inclusive voting and marriage rights are good examples) have emerged relatively recently. Some morals are unique to particular communities. It is this quality of morals, their emergence as communities come together and interact, that we need to be sympathetic to. For AI solutions to be accepted and trusted by a community, one must engage with the communities they are being applied to and their emergent morals. The engagement must be meaningful and ongoing, and it must ensure that what the community believes is moral and "right" is incorporated into the solution.

It's important to note that this journey should be undertaken with an open mindset. The process we've outlined could be followed with a closed and unempathetic, or even merely disinterested, mindset, which would provide quite a different outcome. Without the right skills and mindset, the process may not yield the desired result. Interactions between firm and community members are touchpoints to establish and build trusted, respectful relationships to work toward a mutually acceptable outcome. As discussed earlier, the challenge is to bridge social worlds, rather than privileging one over the other.

We also need to allow for the possibility that firms may choose not to engage with a community in good faith, or for situations when there is a power asymmetry between the firm developing the AI and the community it will affect—or, at worst, situations where the firm's intentions are nefarious. In this case, a framework such as a moral license for AI could provide regulators with the leverage required to enact reporting requirements that ensure that nefarious firms have at least gone through the motions, using known methods while documenting their interactions with the community and the outcomes. Developing an "ethical requirements architecture" could well become the regulatory equivalent of an environmental impact study for AI.

The need for a moral license for AI

WORK ON ETHICAL AI has focused on developing the principles, requirements, technical standards, and best practices needed to realize ethical AI. However, while there is a clear consensus that AI should be ethical, and a global convergence around principles for ethical AI, there remain substantive differences on how these principles should be realized, on what "ethical AI" means in practice.[98]

While this article is notionally about "ethical AI," it never addresses the question of ethics and AI directly, taking a different tack. Rather than attempt to define what AI uses are and aren't ethical, it proposes that firms need to work with the communities they touch, and obtain and maintain a *moral license* for the AI-enabled solutions they want to operate. Moreover, firms should consider doing this for any solution that automates decisions and integrates them with other operational systems to create decisioning networks—not just solutions that contain what is currently considered AI technology.

This difference in approach is due to three observations:

- That AI solutions cannot be made ethical though the development of "fair" or "ethical" algorithms or development methodologies
- That there is no single secular society (a fully normalized social world, an objective standard) against which we can determine if a solution is ethical or good
- That the importance of ethical AI is not due to the development of disruptive AI technology or an existential threat from isolated, self-aware, AI solutions, but rather due to the widespread emergence of automated decisioning networks

Ethical AI—the development of regulation, techniques, and methodologies to manage the bias and failings of particular technologies and solutions—isn't enough on its own. Ethics are the rules, actions, or behaviors that we'll use to get there. Our goal should be moral AI. We must keep a clear view of our ends as well as our means. In a diverse, open society, the only way to determine if we should do something is to work openly *with* the community that will be affected by our actions to gain their trust and then acceptance for our proposal.

Endnotes

1. Luc Zandvliet and Mary Anderson, "Introduction," *Getting it Right: Making Corporate-Community Relations Work* (Sheffield, UK: Greenleaf Publishing Limited, 2009), p. 5.
2. ABC News, "Bentley gas protest makes history," May 20, 2014.
3. ABC News, "NSW government buys back Lismore CSG licence for $1 million," October 19, 2015.
4. We'll use "moral" as being the outcome we're after, while "ethical" is the process we use to get there. "Moral" is what good "ethics" is.
5. While some costs land directly on the P&L statement for a project, others (such as reputational damage) are much harder to quantify.
6. For a conceptual overview of social license to operate, see Joel Gehman, Lianne M. Lefsrud, and Stewart Fast, "Social license to operate: Legitimacy by another name?," *Canadian Public Administration* 60, no. 2 (June 2017): pp. 293–317.
7. my.Flow, accessed July 1, 2020; Jordan White, "Bluetooth tampons—YES!," SmartFem, 2016; Ashley Carman, "A Bluetooth-connected tampon. Hoo boy.," *Verge*, May 18, 2016; Gemma Mullin, "Controversial bluetooth tampon lets you know when it needs changing—but has 12ins string," *Sun*, January 8, 2020.
8. Julia Angwin et al., "Machine bias," *ProPublica*, May 23, 2016.
9. Tim Brennan and William Dieterich, "Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)," in J.P. Singh et al. (eds.), *Handbook of Recidivism Risk/Needs Assessment Tools* (Chichester, UK: John Wiley & Sons, Ltd, 2018), pp. 49–75.
10. Stephanie Wykstra, "Government's use of algorithm serves up false fraud charges," Undark, June 1, 2020.
11. Known as Kranzberg's First Law. From Melvin Kranzberg, "Technology and history: 'Kranzberg's Laws,'" *Technology and Culture* 27, no. 3 (July 1986): pp. 544–60.
12. B. J. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do* (Burlington, MA: Morgan Kaufmann Publishers, 2002), p. 126.
13. Ben Loewenstein, "Regulation of AI: Not if but when and how," *RSA*, November 21, 2017.
14. Industry requests for regulation can be considered something of a red flag, as firms typically don't ask to be regulated unless they see significant risks. The inability to regulate is likely one contributor to recent decisions by a number of firms to suspend work on face recognition technology. See Bobby Allyn, "IBM abandons facial recognition products, condemns racially biased surveillance," NPR, June 9, 2020.
15. Brent Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence,* November 2019.
16. We note that it can also be challenging to implement the principles in any complex regulatory framework, such as data privacy (ISO29100:2011).
17. "Social world" is a term frequently used in sociology that refers to "universes of discourse" through which common symbols, organizations, and activities emerge. They involve cultural areas that need not be physically bounded. Typical examples might be the "social worlds" of surfing, nursing, politics, or science.
18. Nils J. Nilsson, "Preface," T*he Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge, UK: Cambridge University Press, 2009), p. 13.

19. A bowerbird is an Australasian bird noted for the male's habit of constructing an elaborate structure, or bower. The bower is adorned with brightly colored ornaments such as feathers and shells to attract females for courtship.

20. The line between data science and AI, for example, is unclear, with the two fields borrowing many techniques from the other.

21. Paraphrasing Marvin Minsky, among others.

22. Rodney Brooks is an Australian roboticist, Fellow of the Australian Academy of Science, author, and robotics entrepreneur. He is most known for popularizing the actionist approach to robotics.

23. Jennifer Kahn, "It's alive!," *Wired*, March 1, 2002.

24. We can see this in how it is now fairly straightforward to use existing technology to develop solutions that seemed to be science fiction only a few years ago—such as autonomous cars—but continues to be very difficult to advance beyond the current state of the art. Rather than exponential improvements in these solutions, we're seeing exponential increases in cost and effort for only modest, sublinear gains in performance. See Stefan Seltz-Axmacher, "The end of Starsky Robotics," Starsky Robotics 10–4 Labs, March 19, 2020.

25. Matthew Hutson, "Core progress in AI has stalled in some fields," Sciencemag.org, May 29, 2020; *Economist*, "An understanding of AI's limitations is starting to sink in," June 11, 2020.

26. CRISPR is a gene-editing technology used for a range of agricultural and public health purposes, from developing disease- and pest-resistant crops to (more recently) enabling a diagnostic test for the virus that causes COVID-19.

27. Kranzberg, "Technology and history: 'Kranzberg's Laws.'"

28. The authors have previously discussed this shift and how it's changing our relationship with technology in Peter Evans-Greenwood, Robert Hillard, and Alan Marshall, *The new division of labor: On our evolving relationship with technology*, Deloitte Insights, April 9, 2019.

29. Cities, for example, are integrating automated decisions into the systems that manage infrastructure and transport networks and service delivery. Everything from bus schedules and maintenance schedules to policing patterns are being made "smarter." See Beryl Lipton, "Smarter government or data-driven disaster: The algorithms helping control local communities," *MuckRock*, February 6, 2020.

30. We use "decisioning networks" rather than the more common term "cyber-physical systems," as our focus is on how amoral and immoral automated decisions affect networks of integrated decisions and the moral hazard associated with them.

31. Which we might also call *algorithmic moral hazard*.

32. "Internet of things" (IoT) is a term coined around 1994 to refer to a network of small, smart internet-connected devices, the prototypical example being a vending machine modified so that one could check that one's preferred beverage was in stock before taking the long walk from desk to machine. See Bennet Yee, "bsy's list of internet accessible coke machines," August 29, 2003.

33. *Effector* is a term from biology used to refer to an organ or cell that acts in response to a stimulus. It has been adopted by robot developers to refer to an appendage—a wheel or arm—that can be used to interact with, to affect, the world. An alternative term is *influencers*, which is used in some military doctrine and complements a "sensors" and "shooters" AI taxonomy, though this doesn't align with the term as it is commonly used.

34. We don't want to immobilize a moving car due to the potential to cause an accident.

35. For examples, see Greg Jennett, "Robodebt removed humans from Human Services, and the government is facing the consequences," ABC News, May 29, 2020; Wykstra, "Government's use of algorithm serves up false fraud charges."

36. Ibrahim Diallo, "The machine fired me," iD, June 17, 2018.

37. John Thorpe, *The Information Paradox: Realizing the Benefits of Information Technology* (McGraw-Hill Higher Education, 2003).

38. Significant effort has been devoted to developing thought experiments for a wide range of philosophical conundrums, exploring increasingly narrowly defined issues. The application of this thinking to the real world is not clear, however. We might imagine a problem with an autonomous car that requires it to discriminate between two individuals via demographic or physical features, applying an ethical principle to determine which one to hit. This is likely a moot point, though, as no existing or imagined technology can sufficiently discriminate in this way; it's not possible to account for every subtle variation of the problem, nor can we anticipate (and resolve) every possible conflict between principles. It's more likely that a small set of simple principles will be used—brake, and only swerve if there are no obstructions of any kind—with other issues dealt with via insurance and liability. See James Wilson, "The trolley problem," Aeon, May 28, 2020.

39. Philippa Foot, "The problem of abortion and the doctrine of the double effect" in Philippa Foot, *Virtues and Vices and Other Essays in Moral Philosophy* (New York: Clarendon Press and Oxford, UK: Oxford University Press, 2002).

40. One issue with applying the trolley problem to autonomous vehicles is that the thought experiment is designed around agency, and the finding that proximity impacts agency. (The subject is more likely to kill an individual over the group if they are removed or abstracted from the act, but when they have to touch the victim they hesitate or refuse.) This proximity (to the impact) factor is also important in the pluralistic approach taken in this article, and the converse moral hazard problem.

41. Self-driving cars are a case in point. See Kelsey Piper, "It's 2020. Where are our self-driving cars?," *Vox*, February 28, 2020.

42. Naaman Zhou, "Volvo admits its self-driving cars are confused by kangaroos," *Guardian*, June 30, 2017.

43. Will Douglas Heaven, "Google's medical AI was super accurate in a lab. Real life was a different story.," *MIT Technology Review*, April 27, 2020.

44. Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, "On the (im)possibility of fairness," arXiv:1609.07236v1 [cs.CY], September 23, 2016.

45. See p. 162 in Gerald Gaus, *The Tyranny of the Ideal: Justice in a Diverse Society* (Princeton, NJ: Princeton University Press, 2016). "It is important to stress that disagreements about the nature of the social worlds in which we live are neither peripheral nor can they be redescribed as value or preferential disputes (i.e., pushed into the evaluative standards element of a perspective). Some of our deepest and most intractable disputes are not about values or principles of justice, but about the world to which these principles apply. The most obvious instance is the long-standing and persistent struggle concerning abortion rights. Advocates of such rights see the case as decisively about fundamental rights of personal autonomy; opponents of abortion rights are depicted as having little sensitivity to a woman's claim to control her own body. But this by no means follows, and often is simply not the case; opponents of abortion can be deeply devoted to such autonomy, but not in cases where it entails overriding another's right to life. And, of course, in the abstract, most advocates of abortion rights would also draw back in such situations. The dispute is centrally about the social world to which the principles of autonomy and the right to life apply: The two social worlds do not have the same set of persons, and so even perfect agreement about abstract principles of justice would not resolve the dispute.

It is only because so many moral philosophers agree with Sen that there is only a single, fully normalized, secular social world that the dispute has to be misdescribed as one simply about values or abstract principles of justice."

46. Steve Jacobs, "How views on 'when life begins' drive Americans' abortion attitudes," *Heterodox Academy*, July 9, 2020.

47. That is, the "society" that we all inhabit, the normalized social world that is the foundation of our institutions and which all regulation and norms are measured against—a shared, common moral existence that is accepted to all perspectives. There is no single normalized perspective from which we can reason about ethics and justice. "At the heart of the particular problem of a unique impartial resolution of the perfectly just society is the possible sustainability of plural and competing reasons of justice, all of which have claims to impartiality and which nevertheless differ from—and rival—each other." See Amartya Sen, "Introduction" in *The Idea of Justice* (Cambridge, MA: Belknap Press of Harvard University Press, 2011), p. 12.

48. A good, and challenging, example of how our assumptions of "justice" and so on are particular to our own lived experience can be found in Sam Dubal, "Against humanity: What the Lord's Resistance Army can teach us about flaws in the ideal of human rights and the fight for justice," *Aeon*, March 18, 2020.

49. Utilitarianism promotes actions that maximize happiness and well-being for the affected individuals.

50. That is to say, a social world prioritizing equity rather than equality.

51. Angwin et al., "Machine bias."

52. Tafari Mbadiwe, "Algorithmic injustice," *The New Atlantis* 54 (Winter 2018): pp. 3–28.

53. The authors use the following article as a sound set of principles for ethical AI: Jim Guszcza et al., *Human values in the loop: Design principles for ethical AI*, Deloitte Insights, January 28, 2020.

54. This is a good adjunct to the "four ares" by Thorpe, or the Zen practice of asking "Who is asking the question?" If we're to answer the question of "Who decides what is ethical?," then we need to understand the relationship between the group answering the question and those affected by the question's answer, and the intentions of the group answering the question in terms of the "four ares." Who is included (and who is excluded) when answering the question posed, and what factors have been considered?

55. C. P. Snow is reputed to have had an excellent way of remembering the laws of thermodynamics that goes along the following lines:
    0. You must play the game (as the physical world is inescapable).
    1. You cannot win (that is, you cannot get something for nothing, because matter and energy are conserved).
    2. You cannot break even (you cannot return to the same energy state because there is always an increase in disorder; entropy always increases, and all systems have inescapable losses that make perpetual motion impossible).
    3. You cannot get out of the game (because absolute zero is unattainable).
    Wikiquote, "Thermodynamics," accessed July 1, 2020.

56. Indeed, it's likely that we're already surrounded by these automated decisioning networks; it's just that their misbehavior hasn't yet been recognized as a significant problem.

57. Developing a social license to operate passes through three thresholds: *rejection* (via social norms), *acceptance*, and finally *approval* (biases employed for trust). See Robert G. Boutilier and Ian Thomson, "Modelling and measuring the social license to operate: Fruits of a dialogue between theory and practice," Shinglespit Consultants Inc., 2011.

58. We note that "community" is often too narrow a term to describe the web of overlapping and interrelated social groups that any complex solution touches. This is explored in more detail later.

59. As a "method for ethical rule networks" will necessarily be different to a "method for ethical neural network classifiers."

60. This also implies that the question of whether or not a particular solution is ethical is one for the community affected by the solution, not some third party or other external authority. This does *not* imply complete freedom for the firm and community, though, as national norms and regulations still apply.

61. The key difference between a legal license and social license is the "ongoing" nature of a social license, making social license harder to obtain (as there are no explicit rules that govern how they are granted) and easier, and faster, to withdraw.

62. Emily Mullin, "Voice analysis tech could diagnose disease," *MIT Technology Review*, January 19, 2017.

63. This is a problem for both humans and machines, which we can see in subtitles for TV shows set in Scotland and broadcast into England. See Jennifer Hale, "Scots in stitches after STV use subtitles to help viewers understand Glasgow accent in Ross Kemp: Behind Bars as he looks at life in HMP Barlinnie," *Scottish Sun*, November 2, 2017.

64. A recent and topical example is the challenge of allocating ventilators in a hospital overwhelmed by COVID-19. See Robert D. Truog, Christine Mitchell, and George Q. Daley, "The toughest triage—allocating ventilators in a pandemic," *New England Journal of Medicine* 382, no. 21 (2020): pp. 1973–75.

65. One of the Melbourne-born authors found it impossible, when living in San Francisco, to book a restaurant over the phone for 8 p.m. No matter how they phrased "8 p.m.," the restaurant staff answering the phone could not understand what they were saying. The solution was to request tables for 7:30 p.m. or 9 p.m., and accept 8 p.m. if the restaurant offered it as an alternative.

66. Medical research is biased toward males, and particularly manifests during patient-physician gender discordance. AI can amplify this bias and, for example, misdiagnose a female patient's heart attack. For an example, see Brad N. Greenwood, Seth Carnahan, and Laura Huang, "Patient–physician gender concordance and increased mortality among female heart attack patients," *Proceedings of the National Academy of Sciences* 115, no. 34 (August 6, 2018): pp. 8569–74.

67. The problem of being seen to be "crying wolf."

68. This links back to the trolley problem, mentioned earlier, and the impact of proximity, moral hazard, and the point that there is no correct choice.

69. We use "discriminate" here as we have to accept that, for the system to provide benefits, it will often need to identify people and treat them differently, to *discriminate* between individuals. The problem emerges when this discrimination produces undesirable outcomes for some individuals.

70. Harmon Leon, "Whole Foods secretly upgrades tech to target and squash unionizing efforts," *Observer*, April 24, 2020.

71. A disposition to tolerate, agree, or consent to.

72. Have favorable regard, agree to, or be pleased with.

73. Alex Hern, "Anti-surveillance clothing aims to hide wearers from facial recognition," *Guardian*, January 4, 2017.

74. By "quality of working" we imply "*good* jobs—jobs that, without undue intensity or stress, make the most of workers' natural attributes and abilities; where the work provides the worker with motivation, novelty, diversity, autonomy, and work/life balance; and where workers are duly compensated and consider the employment contract fair. Crucially, good jobs support workers in learning by doing—and, in so doing, deliver benefits on three levels: to the worker, who gains in personal development and job satisfaction; to the organization, which innovates as staff find new problems to solve and opportunities to pursue; and to the community as a whole, which reaps the economic benefits of hosting thriving organizations and workers. This is what makes good

jobs productive and sustainable for the organization, as well as engaging and fulfilling for the worker. It is also what aligns good jobs with the larger community's values and norms, since a community can hardly argue with having happier citizens and a higher standard of living." See Peter Evans-Greenwood, Alan Marshall, and Matthew Ambrose, *Reconstructing jobs: Creating good jobs in the age of artificial intelligence*, Deloitte Insights, July 18, 2018

75. This is complicated by the fuzzy agency relationship between a firm, the platform(s) it operates, third parties, and the consumers inhabiting the platform. Often, the boundaries between these parties are unclear.

76. Sydney Bauer, "Trans travellers face 'invasive' airport security at Thanksgiving," *Thomson Reuters Foundation News*, LGBT+, November 28, 2019.

77. Operational community engagement.

78. Strategic community engagement.

79. By "technical details," we mean the details from both the technical workings of the AI solution and ethical methodology and framework technicalities.

80. This could be seen as a simplification of the sort of requirements modelling done in methods such as Tropos and i*, though without the requirement for a connected graph. It would be quite possible to have a set of disconnected information-decision-action chains, as we're only offering up the idea of a solution rather than something that can be translated into implementation.

81. This distinction should really be between human decisions and abstract or formalized decisions, as algorithms are just the latest formalism. See James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven, CT: Yale University Press, 1999).

82. In *Seeing Like a State*, Scott gives an excellent and sustained discussion on how adopting an algorithmic approach to understanding the world has also changed the world.

83. Indeed, with robodebt, it's likely that a human was required, by law, to make the decision to send a claim letter. Dana McCauley and Rob Harris, "Lawyers warned federal government robodebt scheme was 'unlawful,'" *Sydney Morning Herald*, February 6, 2020.

84. Ibid.

85. Clifford Geertz, "Chapter 1. Thick description: Toward an interpretive theory of culture," in *The Interpretation of Cultures*, third edition (New York: Basic Books, 2017).

86. Anthropologists often use the word "informant" in field work to refer to the community members they work with. There is also the idea of the gatekeeper who can broker the introduction of the anthropologist into the community and host the anthropologist in the community. Who introduces, hosts, and befriends the anthropologist informs the anthropologist's insight into a community, and to whom the anthropologist is aligned also influences the type of information they will receive. Hence, that initial introduction into a community and who it is brokered through is something anthropologists often spend a lot of time searching and planning for, as it can influence so much of what comes afterward. As time goes on, a network of informants is created in a community, and between the people (informants) they speak with and the experiences anthropologists have themselves (as active members in the community), they can discover the community factors and gain an understanding of the community.

87. Melanie (Lain) Dare, Jacki Schirmer, and Frank Vanclay "Community engagement and social licence to operate," *Impact Assessment and Project Appraisal* 32, no. 3 (2014), pp. 188–97.

88. The authors chose to call these things “community factors,” as it was a challenge to find an existing term that would understandable in the article. It’s usual to choose a name that makes sense to the community being researched, and typical choices include scenario, preference, use case, dimension, or even use case dimensions. Unfortunately, none of these worked in the context of this article. Our point here is that that our word choice is not typical for anthropology as a discipline.

89. An informal Australian term for a map drawn on the ground with a stick, or any other roughly drawn map.

90. For a more comprehensive definition of actor networks and actor-network theory (ANT), see M. Callon, “Actor network theory,” in Neil J. Smelser and Paul B. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences* (Oxford, UK: Pergamon, 2001), pp.62–66.

91. General morphological analysis was developed by Fritz Zwicky—a Swiss astrophysicist and aerospace scientist based at the California Institute of Technology—as a method for structuring and investigating the total set of relationships contained in multidimensional, non-quantifiable problem complexes. For an overview of the method, see Tom Ritchey, “General morphological analysis: A general method for non-quantified modeling,” 2002.

92. “Matrixing,” in GMA terminology.

93. Constraint satisfaction is an AI and operations research technique that finds solutions to a set of constraints (rules) that represent conditions imposed on the parameters used to model or define a problem.

94. Simon Wardley, “An introduction to Wardley (value chain) mapping,” Bits or pieces?, February 2, 2015.

95. We note that the tracking itself may be the problem, and consequently, no amount of reconfiguring of the tracking solution will lead to community acceptance.

96. This is the opposite of the common approach of taking a predefined technological solution to a problem to the community.

97. Irfan Saif and Beena Ammanath, “‘Trustworthy AI’ is a framework to help manage unique risk,” *MIT Technology Review*, March 25, 2020.

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

About the authors

Peter Evans-Greenwood | pevansgreenwood@deloitte.com.au

Peter Evans-Greenwood is a fellow at the Deloitte Australia Centre for the Edge, helping organizations embrace the digital revolution through understanding and applying what is happening on the edge of business and society. He has spent 20 years working at the intersection between business and technology. These days, he works as a consultant and strategic adviser on both the business and technology sides of the fence.

Rob Hanson | rob.hanson@csiro.au

Rob Hanson is a senior research consultant at the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia's national science agency. Hanson is a transdisciplinary researcher who works at the intersection of emerging technology and their policy implications. He has a professional background in technology, security, risk management, and strategic foresight. Hanson's current research focus includes data privacy, consumer data, and trust. As a PhD candidate, his thesis is titled "Fables for future technology."

Sophie Goodman

Sophie Goodman is an applied anthropologist who previously worked in customer-led strategy and service design projects as part of Deloitte Digital. She uses her experience and training in ethnographic research to help organizations better understand and act on the needs and experiences of their customers. Goodman has worked in research roles for an Australian university and a global workplace culture consulting firm as well as on customer and user experience projects.

Dennis Gentilin | degentilin@deloitte.com.au

Dennis Gentilin helps organizations build the infrastructure required to promote ethical behavior and drive sustainable performance. A unique career experience catalyzed his strong interest in ethics. His book, *The Origin of Ethical Failures*, won the textbook prize in the 2017 UK Chartered Management Institute Book of the Year awards. He is an adjunct fellow at Macquarie University and an honorary fellow at the Centre for Ethical Leadership.

Contact us

Our insights can help you take advantage of change. If you're looking for fresh ideas to address your challenges, we should talk.

Industry leadership

Alan Marshall
Partner | Deloitte Touche Tohmatsu
+61 893 658 139 | almarshall@deloitte.com.au

Alan Marshall is the national lead of Deloitte Australia's Analytics and Cognitive practice. He specializes in developing strategies and solutions that integrate human and machine decision-making to improve yield and productivity.

Kellie Nuttall
Partner | Deloitte Touche Tohmatsu
+61 733 087 075 | knuttall@deloitte.com.au

Kellie Nuttall specializes in consumer psychology, using her expertise to develop digital twins to support decision-making in asset-intensive industries as part of Deloitte Australia's Analytics and Cognitive practice.

The Deloitte Australia Centre for the Edge

Peter Evans-Greenwood
Fellow | Deloitte Australia Centre for the Edge
+61 439 327 793 | pevansgreenwood@deloitte.com.au

Peter Evans-Greenwood is a fellow at the Deloitte Australia Centre for the Edge.

Deloitte.
Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights.

 Follow @DeloitteInsight

About Deloitte Insights

Deloitte Insights publishes original articles, reports, and periodicals that provide insights for businesses, the public sector, and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific, and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses, and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.

In Australia, the member firm is the Australian partnership of Deloitte Touche Tohmatsu. As one of Australia's leading professional services firms, Deloitte Touche Tohmatsu and its affiliates provide audit, tax, consulting, and financial advisory services through approximately 6,000 people across the country. Focused on the creation of value and growth, and known as an employer of choice for innovative human resources programs, we are dedicated to helping our clients and our people excel.

For more information, please visit Deloitte's web site at www.deloitte.com.au.

Liability limited by a scheme approved under Professional Standards Legislation.